

**Item Response Theory,  
Adaptive Assessment, and  
Health Services Research**

Thomas E. Love, Ph. D.  
November 30, 2004  
thomaslove@case.edu  
[www.chrp.org/love](http://www.chrp.org/love)

1

**Summary: Three Breakthroughs**

- Modern psychometric models
  - Dichotomous or Multi-category items?
  - Parametric and non-parametric models
- Computerized dynamic assessments
  - What must we learn about item banks?
- Availability on the Web

2

**Dichotomous Items Measure  
One Scale Level**

Did you feel downhearted and depressed  
some of the time during the past month?

-5 -4 -3 -2 -1 0 1 2 3 4 5

3

**Did You Feel Downhearted and  
Depressed During the Past Month?**

YES, Some  
of the Time

Worst Health -5 -4 -3 -2 -1 0 1 2 3 4 5 Best Health

NO, More Often      NO, Less Often

4

**Did You Feel Downhearted  
and Depressed During the Past  
Month?**

All of  
the time

None of  
the time

Worst Health -5 -4 -3 -2 -1 0 1 2 3 4 5 Best Health

5

**Categorical Rating Scale**

**How Much of the Time Have You  
Felt Downhearted and Depressed  
During the Past Month?**

All      Most      Some      A little      None

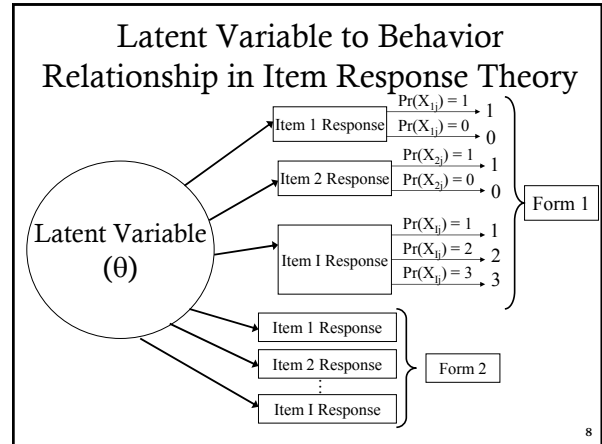
Worst Health -5 -4 -3 -2 -1 0 1 2 3 4 5 Best Health

6

### Item Response Theory [IRT]

- Item response models describe the confrontation of a particular subject with a particular item in terms of parameters for that subject and item.
- IRT models individual responses to items, can then specify performance of the subjects as well as the items.

7



### The Latent Trait

- Early notions suggested that items all ought to measure the same thing.
- IRT formalizes this by explicitly positing a single underlying trait on which all of the items rely.
- So each item may be placed on this single dimension,  $\theta$ .

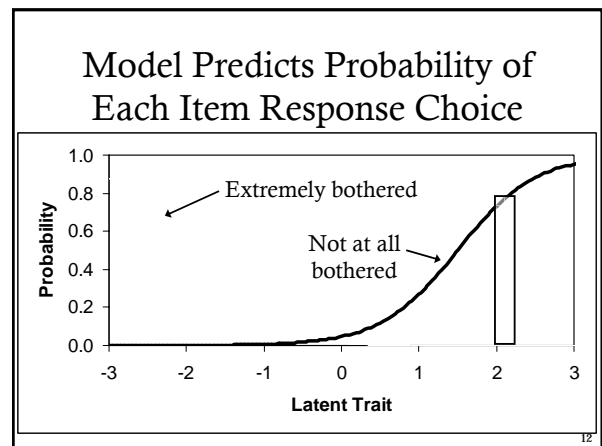
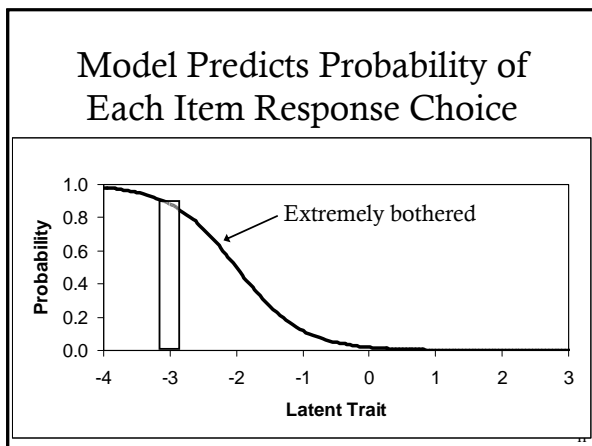
9

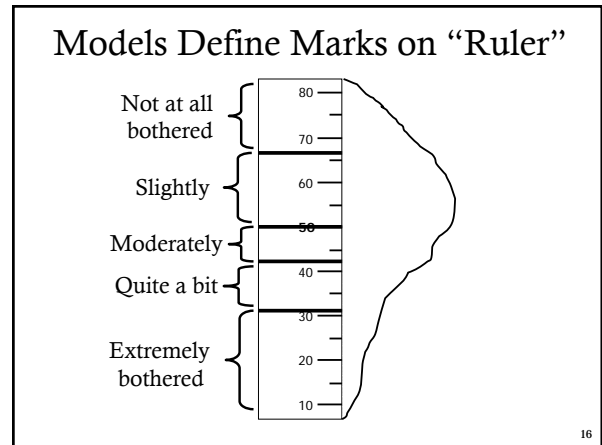
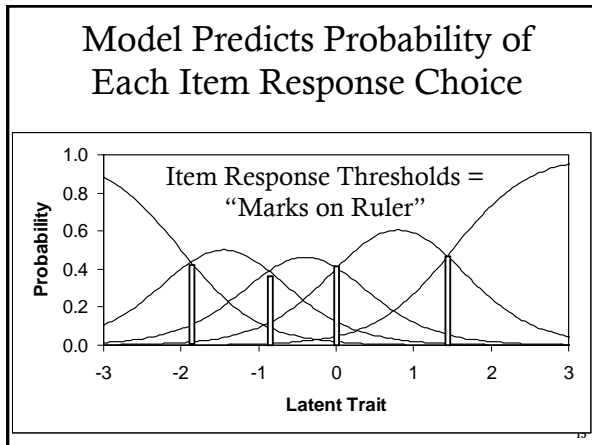
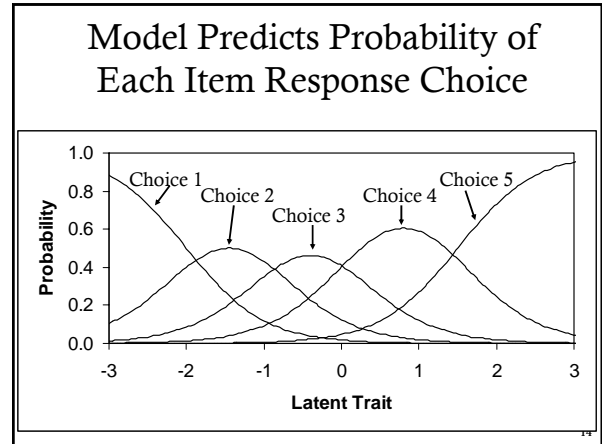
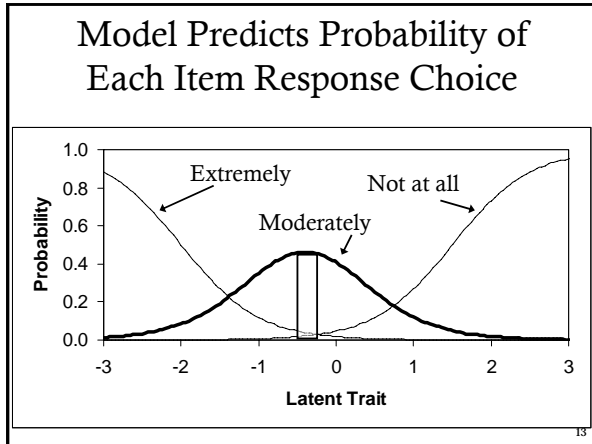
### New Psychometric Models Improve Scoring of Questionnaire Items

- During the past 4 weeks, how much have you been bothered by emotional problems (such as feeling anxious, depressed or irritable)?

- Not at all
- Slightly
- Moderately
- Quite a bit
- Extremely

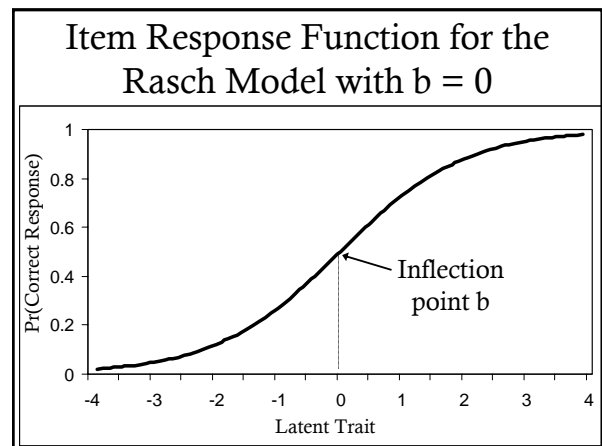
10

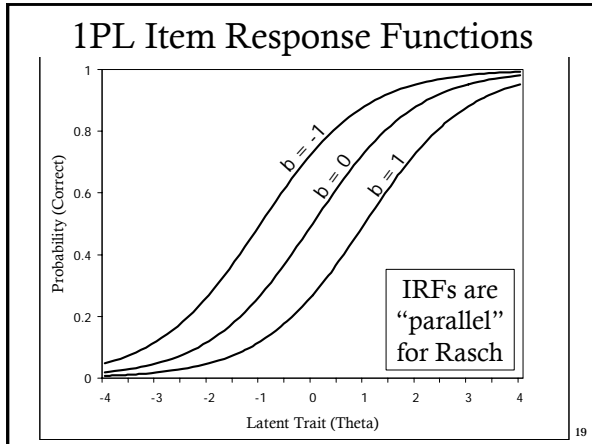




### The Rasch Model: A Simple IRT Model for Binary Items

- Probability of someone with ability  $\theta$  endorsing (i.e. choosing "Yes") for an item with location parameter  $b$
- One-parameter Logistic (1PL):

$$P(\theta) = \frac{1}{1 + e^{-(\theta - b)}}$$




### No Distributional Assumptions about $\theta$ !

- Scaling of the latent trait is arbitrary.
- Convenient to use a scale with mean of zero and a standard deviation of one.
- Regardless of the  $\theta$  distribution, it is common to find nearly all scores within three standard deviations of the mean.

### Key IRT Assumption: Local (Conditional) Independence

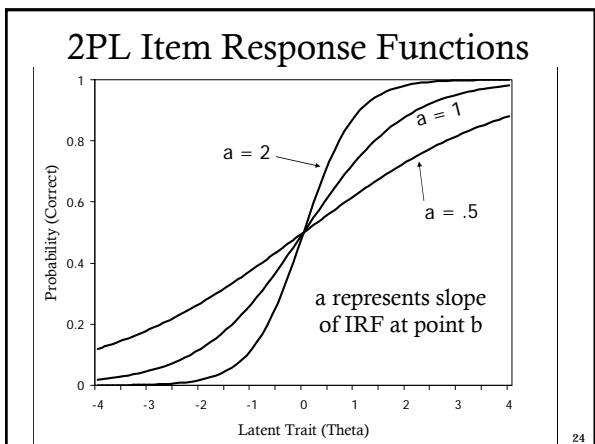
- We model response probability using only subject's latent trait  $\theta$  and item's location  $b$ .
  - No parameters for other items, or for the order of the items, etc.
  - Imposes requirements on the patterns we expect to see in data.

### Key Modeling Assumptions

- Monotonicity
  - People with higher levels of the latent trait will have a higher probability of endorsing items that are scaled higher
- Unidimensionality of Latent Trait
  - A single (dominant) trait influences the probability of item endorsement

### Generalizing the Model

- Rasch forces parallel IRFs ("slopes" must be the same for all items)
  - Can allow the slopes to diverge
- Item's discrimination, denoted  $a$ , characterizes the slope of the IRF.
- 2PL Model:
 
$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}}$$



### Personality Modeling

- 2PL is the most common approach for modeling personality data
  - We can extend the 2PL model to help deal with multi-categorical responses (graded response model)
- 2PL is rarely used in educational measurement because of “guessing”

25

### The 3PL Model

- $c$  = a binomial “floor” on probability of endorsing an item
- Appropriate when “guessing” prevents IRF from flooring at zero...
- 3PL:
 
$$P(\theta) = c + \frac{(1-c)e^{a(\theta-b)}}{1+e^{a(\theta-b)}}$$

26

### The 3PL Model

- Slope
- Location
- “Guessing” / Floor

27

### Is The Third Parameter Useful in Personality Modeling?

- Why do some personality items have a nonzero lower asymptote (“c”)?
- Lower asymptote may reflect an item’s social desirability value.
- Could also be an indication of hidden multidimensionality...
- Ideally, we want items so that low-trait people almost never endorse and high-trait people almost always endorse

28

See Reise and Waller (2003) Psychological Methods 8(2), 164-184

### A Four-Parameter Logistic Model

- Allows us to estimate lower bounds greater than 0 and upper bounds less than 1 for endorsement probability

$$P(\theta) = c + \frac{(d-c)e^{a(\theta-b)}}{1+e^{a(\theta-b)}}$$

29

### Non-Parametric IRT Models: Mokken’s MMH Model

- Model of Monotone Homogeneity
  - Unidimensional measurement
  - Increasing IRF as latent trait  $\theta$  increases
  - IRF need not be logistic – can be used in situations where 2PL and 3PL don’t fit
- Model yields subject ordering by  $\theta$  on the basis of total scores
- Provides diagnostics as to whether ordering by total score is justified

30

See Meijer and Baneke (2004) Psychological Methods 9(3), 354-368

### Nonparametric IRT through Nonparametric Regression Models

- Estimate the item response function without assuming a logistic form
  - Isotonic regression estimation, or
  - (More Common) Kernel smoothing
- TESTGRAF does this work for you
  - Estimate IRFs for binary, multiple choice or scaled items using kernel smoothing, and without enforcing monotonicity in  $\theta$

<http://www.math.mcgill.ca/~cao/Research/ramsay/testgraf/> 31

### Motivation for using Nonparametric IRT Models

- Models do not impose a specific form on item response functions
- Can be used with relatively small samples ( $n = 300-400$ , for instance)
- These models and graphs alert you that things don't work uniformly well across all elements of the development sample

32

### What happens with Likert Scales?

- Lots of outcomes are not binary.
- Models for ordered polytomous data
  - Adjacent-categories logit models (partial credit model)
  - Cumulative logit models (graded response model) – let's look at this one a little more closely

33

### Graded Response Model

- Appropriate for Likert-style scales
- Ordered categorical responses
- 5-category items described by
  - one slope parameter  $a$
  - 4 “threshold” parameters  $b_1, b_2, b_3, b_4$  represent trait level required to get to higher response with probability .50
- Item treated as 4 dichotomies...

34

### Graded Response Model for An Item with Five Response Choices

- Item has categories 0, 1, 2, 3, and 4.
- $\Pr(\text{response } x \text{ is category } j \text{ or higher given latent trait level } \theta) =$

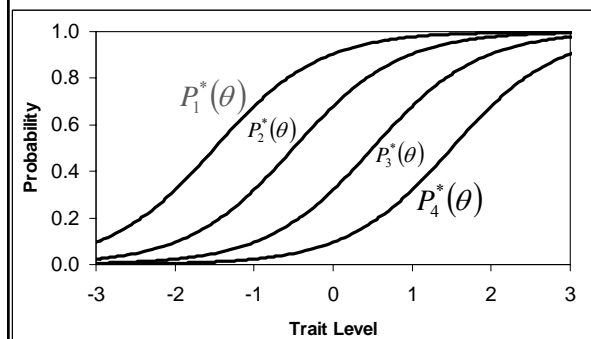
$$P_{ix}^*(\theta) = \frac{\exp\{a_i(\theta - b_{ij})\}}{1 + \exp\{a_i(\theta - b_{ij})\}}, x = j = 1, \dots, 4$$

- Category Response Curves:

$$P_{ix}(\theta) = P_{ix}^*(\theta) - P_{i(x+1)}^*(\theta)$$

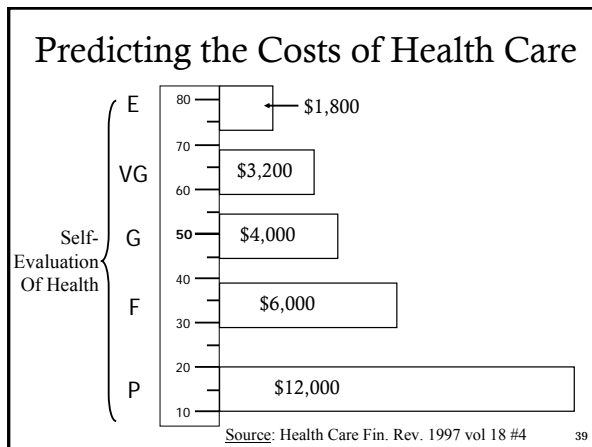
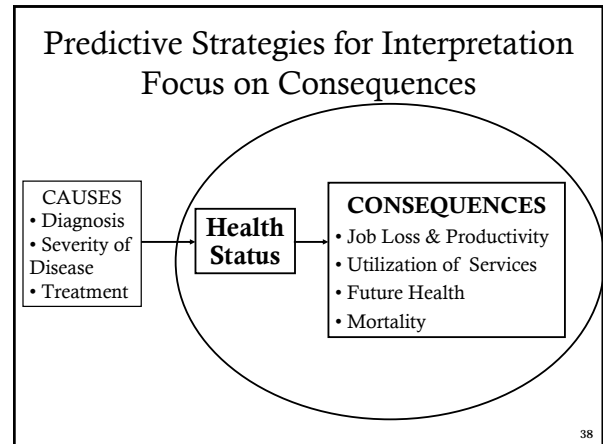
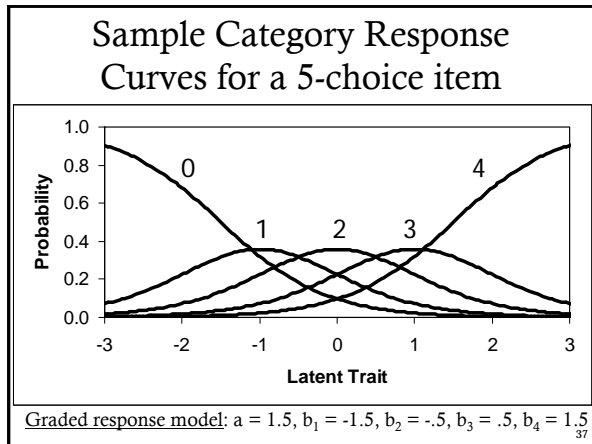
35

### Sample OCCs for a 5-choice item



Graded response model:  $a = 1.5, b_1 = -1.5, b_2 = -.5, b_3 = .5, b_4 = 1.5$

36



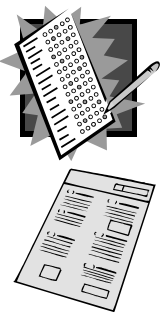
- ### Example: Items from Multiple Forms Define Physical Functioning "Ruler"
- 11 - Vigorous activities
  - 10 - Vigorous activities with limitations
  - 9 - Moderate activities
  - 8 - Moderate activities with limitations
  - 7 - Walk slowly - trouble bending, stooping
  - 6 - Need help to bathe
  - 5 - Cannot maintain balance
  - 4 - Move about with help
  - 3 - Stand up with help
  - 2 - Staying in bed / partly undressed
  - 1 - Lying down most of time
  - 0 - Confined to room, bed
- Source: Health Assessment Lab (HAL)

- ### Interpreting a 5 Point Improvement in Physical Health (PCS)
- **Content-Based:** Decrease from 28% to 13% limited in climbing stairs.
  - **Normative:** A 5 point change in PCS is about half a standard deviation.
  - **Psychometric:** Very unlikely to be due to chance.
  - **Disease Burden:** Larger than the impact of Type II diabetes patients relative to the norm.
  - **Consequences:**
    - 1/3 reduction in Pr(hospitalized in next 6 months)
    - Predicted reduction in estimated health care cost from \$1500 to \$1100.

### Measurement Strategies

- Suppose the runners ran on different days, or different courses, or both?
- Establish standards for all courses. [Control]
- Statistically correct for differences. [Adjust]

### Standardization of Tests



- Adjustment strategy for large-scale assessments.
- Make tests and testing situations as identical as possible.
- Different forms with different items taken at different times are still comparable.
- Scaling and Equating


43

### “Static” Testing

- Items are selected to be suitable for some hypothetical “average” person.
- Each subject receives the same batch of items, in the same order, and is scored in the same way.
- Pool of items is homogenous – often designed to cover entire population, regardless of efficiency.

44

### “Tailoring” and High Jumping



- Starting height that competitors choose depends on their ability.
- Very good jumpers “pass” lower heights.

Which jumper is superior?

	6 feet	7 feet
Jumper A	“Pass”	Clears
Jumper B	Fails to Clear	

45

### Adaptive Testing

- Each subject receives an individually tailored sequence of items.
- Items in the pool vary markedly in terms of information to the scorer.
- High Jump: if you miss, you’re out.
- Adaptive Testing: if you miss, you get an easier question.

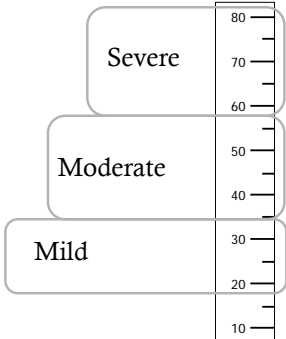
46

### Adaptive Selection of Items

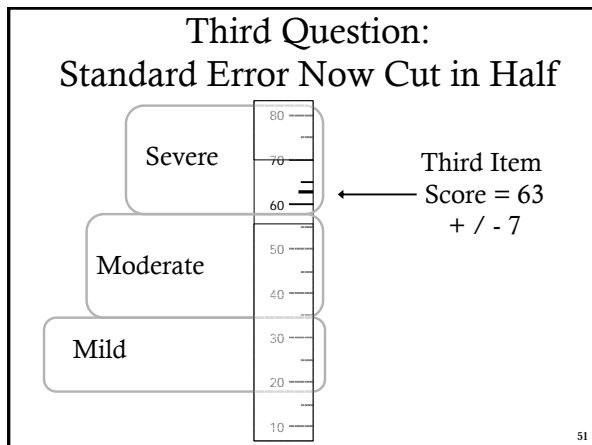
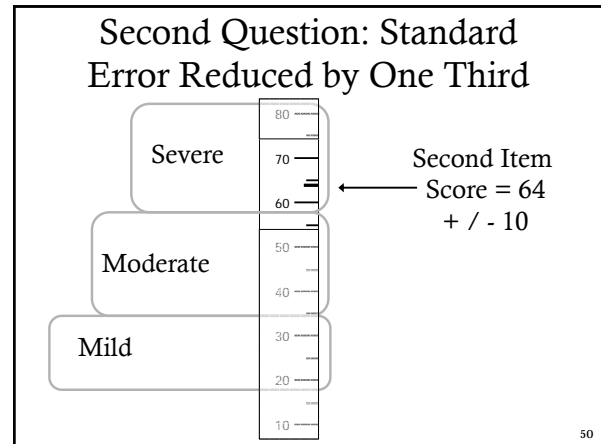
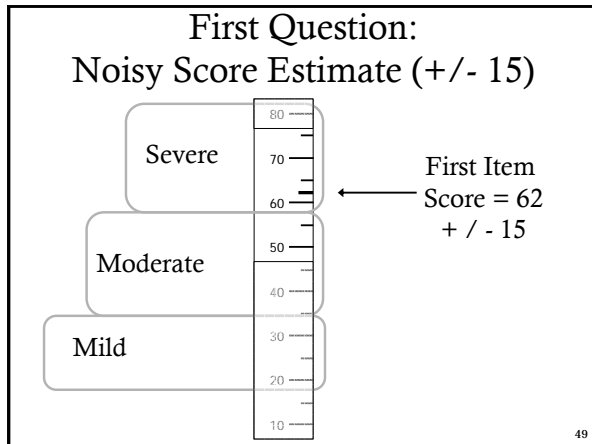
- Present each subject with items that are particularly informative.
- Key Issue: how can we maximize the precision of your measurement for a given amount of assessment time?
- Key Step Forward: you keep giving us some new information.

47

### Dynamic Assessments Match Questions to Each Patient’s Level



48



- Challenges of Adaptive Assessment**
1. Characterize the variation among items usefully.
  2. Determine efficient rules for selecting items to administer to a subject.
  3. Arrive at scores on a common scale, even though subjects see different sets of items.
- 52

- Item Selection**
- No point in asking questions about whether you can walk several hundred yards if you've already said you can't walk 100 yards.
  - No point in asking if you have trouble climbing stairs if you've said you have no physical problems.
- 53

- Which Item To Select Next?**
- Want to estimate your trait value efficiently.
  - Want to pick an item that will help us distinguish you from people who have given similar previous responses.
  - Want to "hone in" on your "correct" score for the trait.
- 54

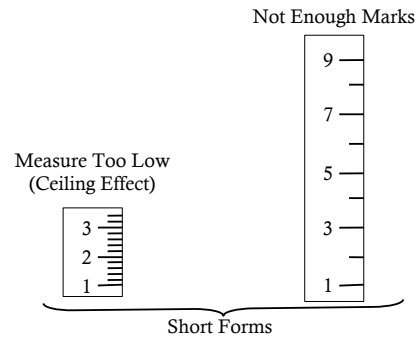
### “Maximum Information” Strategy for Adaptive Assessment

1. Evaluate all items not yet administered to determine which will be best to administer next given currently estimated trait level
2. “Best” next item administered, and subject responds
3. New estimate of trait computed using responses to all administered items
4. Repeat 1 through 3 until stopping criterion.

Rudner, LM (1998). *An On-line, Interactive, Computer Adaptive Testing Tutorial* <http://ericae.net/scripts/cat>

55

### Short-Form Surveys Alone Do Not Provide Practical Solutions



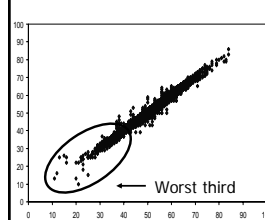
56

### What Are The Advantages of Dynamic Assessments?

- More accurate risk screening
- Reliable enough to monitor individual patient outcomes
- Brevity of a short form – 90% reduction in respondent burden
- Markedly reduced data collection costs
- Eliminate “ceiling” and “floor” effects

57

### Do Dynamic Assessments Reduce Respondent Burden?



(N = 2753, overall r = .985)

- Among patients in poorer mental health (worst third), 92% met clinical standard of precision with five or fewer questions

Source: John Ware at Academy Annual Meeting 2001 Short Course

58

### Dynamic Internet Assessments Greatly Lower Costs

- Face-to face interview \$200-\$300
- Telephone interview \$50-\$75
- Self-administered \$25-\$50
- Internet administered \$1

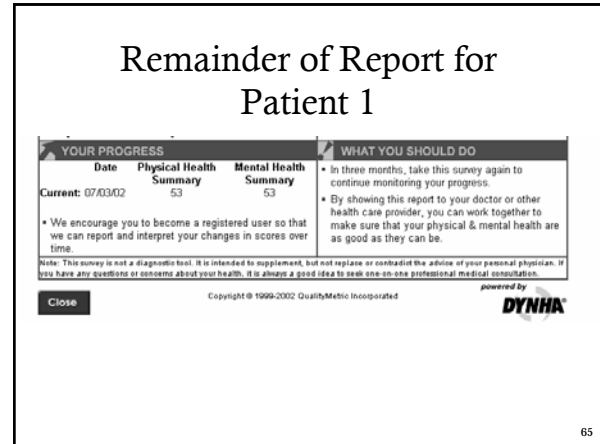
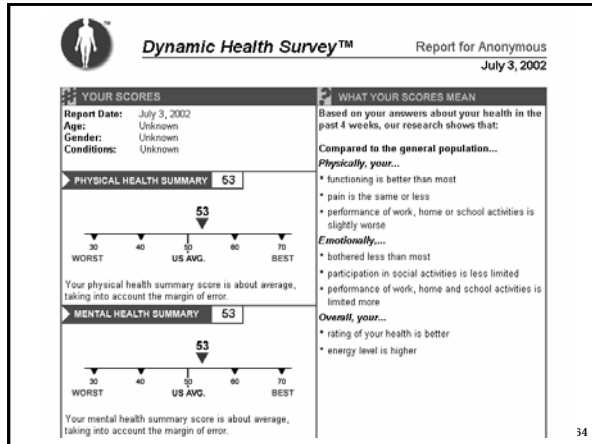
Source: John Ware at Academy Annual Meeting 2001 Short Course

59

### Question 1



61



**Summary: What is IRT?**

- Item Response Theory is a family of mathematical descriptions of what happens when a subject meets an item.
- The most fundamental aspect of IRT is its depiction of the interaction between item and subject.

**Summary: Advantages of Dynamic Assessments**

- More accurate risk screening
- Reliable enough to monitor individual patient outcomes
- Brevity of a short form – up to 90% reduction in respondent burden
- Markedly reduced data collection costs

**For More Information...**

- The Basics of Item Response Theory (Frank Baker) <http://ericae.net/irt>
- SF-36, SF-12, and SF-8 Surveys [www.sf-36.com](http://www.sf-36.com)
- Dynamic Health Assessments [www.qmetric.com](http://www.qmetric.com) [www.amlhealthy.com](http://www.amlhealthy.com) [www.headachetest.com](http://www.headachetest.com)

**More Sources for Information**

- Understanding Health Outcomes Videos Educational Series [www.healthstatprod.com](http://www.healthstatprod.com)
- Latent Class Analysis <http://ourworld.compuserve.com/homepages/jsuebersax/> <http://www2.chass.ncsu.edu/garson/pa765/latclass.htm>
- General Resources <http://lib.stat.cmu.edu> <http://www.math.yorku.ca/SCS/StatResource.html> <http://www.education.umd.edu/EDMS/tutorials/frontpage.html>