Methods Seminar Five: Item Response Models for Health Services Research

Thomas E. Love, Ph. D.

July 12, 2002

tel3@po.cwru.edu www.chrp.org/love

What's a Questionnaire Item?

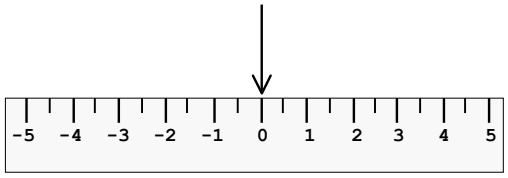
- Question about health?
  - ❑ Response Choice 1
  - ❑ Response Choice 2
  - ❑ Response Choice 3
  - ❑ Response Choice 4
  - ❑ Response Choice 5

Sample Dichotomous Item

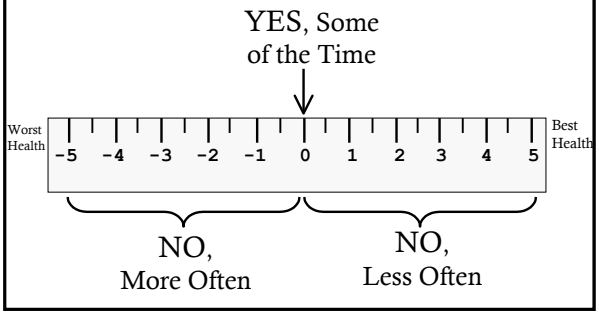
- Did you feel downhearted and depressed some of the time during the past month?
  - ❑ Yes
  - ❑ No

Dichotomous Items Measure One Scale Level

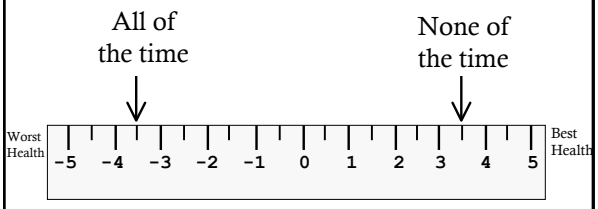
Did you feel downhearted and depressed <u>some of the time</u> during the past month?



Did You Feel Downhearted and Depressed During the Past Month?

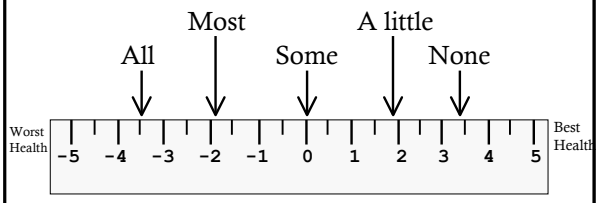


Did You Feel Downhearted and Depressed During the Past Month?



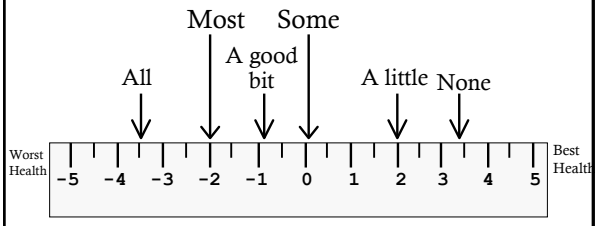
Categorical Rating Scale

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



Sixth Response Category Doesn't Add Much

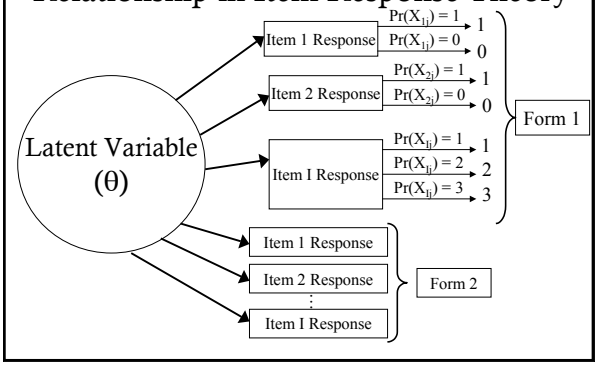
How Much of the Time Have You Felt Downhearted and Depressed During the Past Month?



Item Response Theory [IRT]

- IRT is a mathematical model for 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.

Latent Variable to Behavior Relationship in Item Response Theory



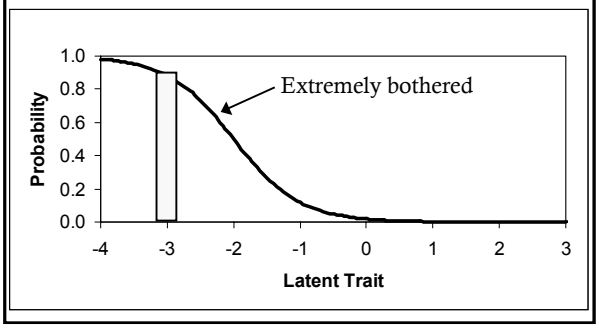
The Latent Trait

- 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, θ.

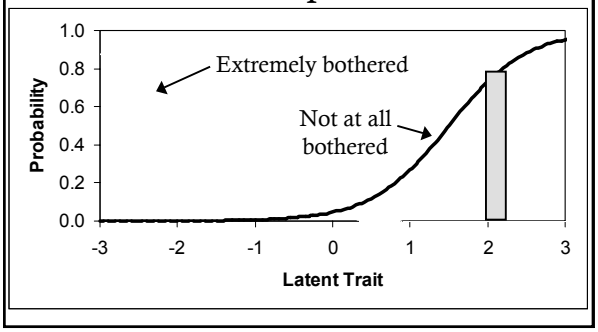
New Psychometric Models Improve Scoring of Questionnaire Items

- 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

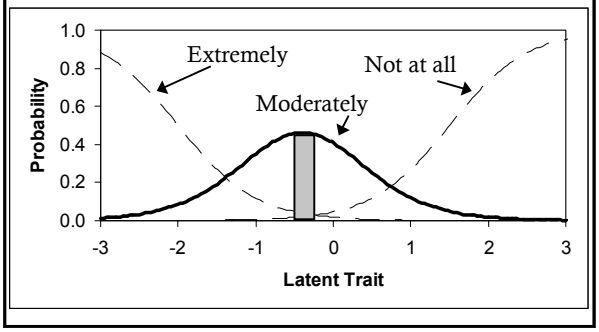
Model Predicts Probability of Each Item Response Choice



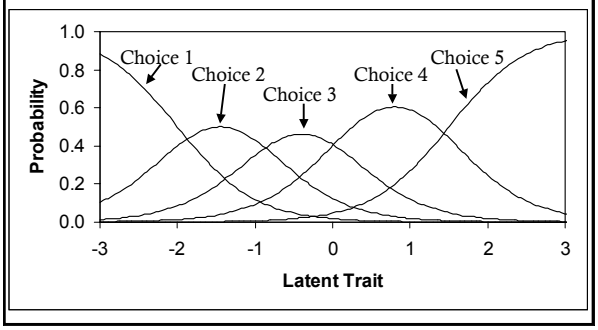
Model Predicts Probability of Each Item Response Choice



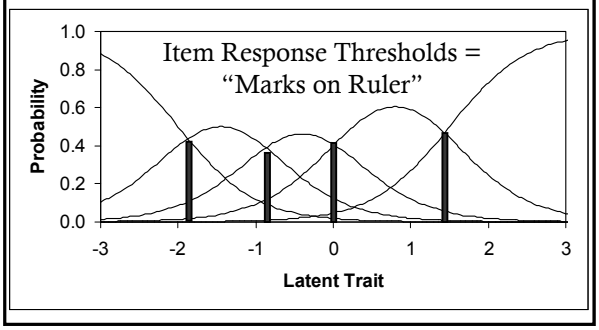
Model Predicts Probability of Each Item Response Choice



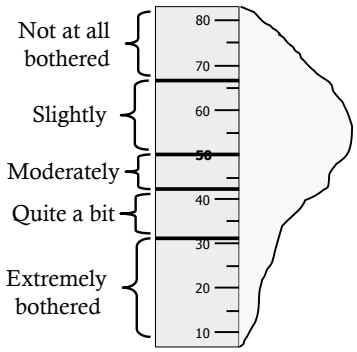
Model Predicts Probability of Each Item Response Choice



Model Predicts Probability of Each Item Response Choice



Models Define Marks on "Ruler"



Educational Testing and Health Care Research

- Why am I talking about tests?
  - Easy to think about psychological or achievement tests – foundations.
  - Nothing forces this, though. We can talk about any sort of item.
- Will I cut it out?
  - I'll try to focus on health measures.

Item "Difficulty" and Subject "Proficiency"

- Items: Items are characterized as more or less difficult. "Difficulty" = b
- Subjects: Subjects are more or less proficient. Proficiency = θ
- IRT gives Pr(response x for item i) based on the values of b and θ.

The Rasch Model: Simple IRT

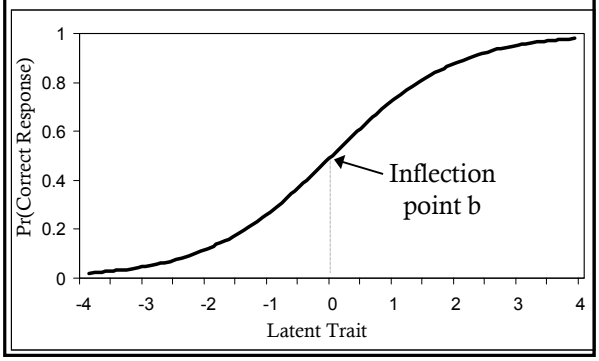
- Probability of someone with proficiency θ responding "correctly" to an item of difficulty b, described by a logistic model
- One-parameter Logistic (1-PL):

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

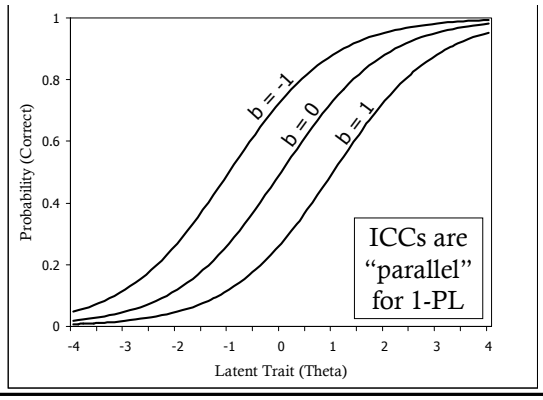
Interpreting the Rasch (1-PL) Model

- Lots of interesting applications when you have dichotomous items.
  - Rasch (1960, 1980)
  - Wright & Stone (1979)
- To see the effect of parameters, we look at item characteristic curves.

Item Characteristic Curve for the 1-PL with b = 0



1-PL Item Characteristic Curves



No Distributional Assumptions about θ!

- 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 θ distribution, it is common to find nearly all scores within three standard deviations of the mean.

Conditional Independence

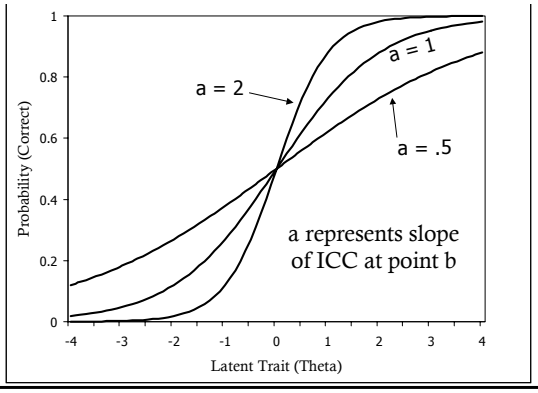
- We model response probability using only subject's θ and item's b.
- No parameters for other items, or for the order of the items, etc.
- Imposes requirements on the patterns we would expect to see in data.
- Basis of estimating both θ and item parameter(s).

Generalizing the Model

- 1-PL forces parallel ICCs ("slopes" must be the same for all items)
- Can allow the slopes to be divergent
- Item's discrimination, denoted a, characterizes the slope of the ICC.
- 2-PL Model:

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

2-PL Item Characteristic Curves

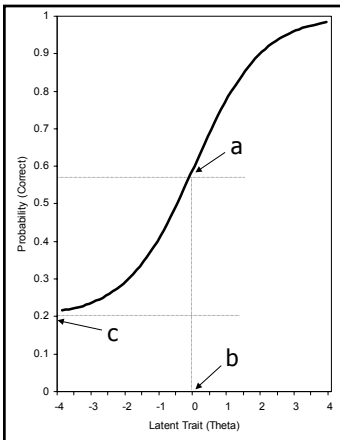


The 3-PL Model

- c = a binomial floor on the probability of correctly responding to an item
- Appropriate when "guessing" prevents ICC from flooring at zero…
- 3-PL:

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

The 3-PL Model



a. Slope
b. Difficulty / Location
c. "Guessing" / Floor

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…

Graded Response Model for An Item with Five Response Choices

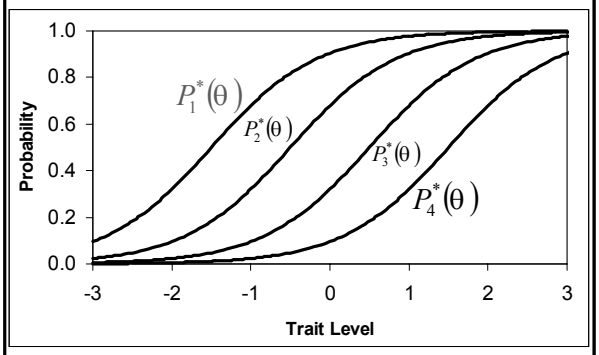
- Item has categories 0, 1, 2, 3, and 4.
- Pr(response x is category j or higher given latent trait level θ) =

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

- Category Response Curves:

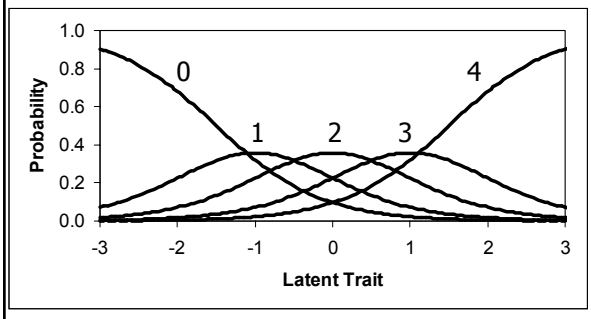
$$P_{ix}(\theta) = P^*_{ix}(\theta) - P^*_{i(x+1)}(\theta)$$

Sample OCCs for a 5-choice item



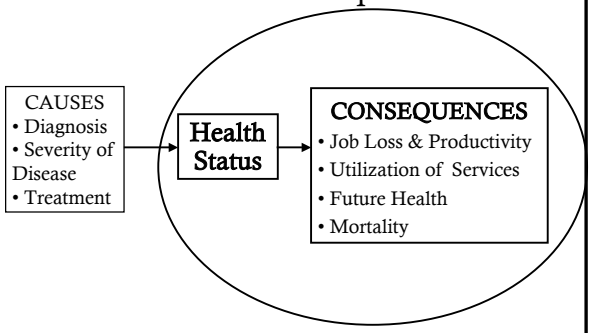
Graded response model: a = 1.5, b_1 = -1.5, b_2 = -.5, b_3 = .5, b_4 = 1.5

Sample Category Response Curves for a 5-choice item

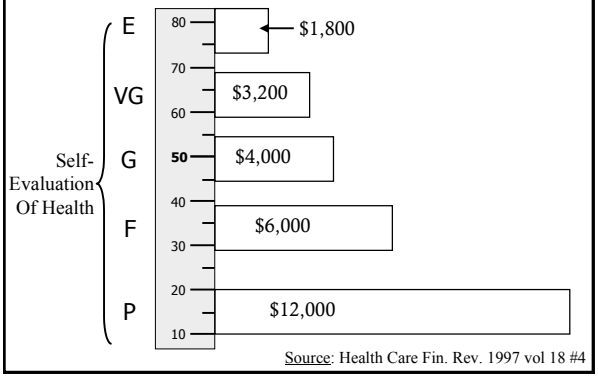


Graded response model: a = 1.5, b_1 = -1.5, b_2 = -.5, b_3 = .5, b_4 = 1.5

Predictive Strategies for Interpretation Focus on Consequences

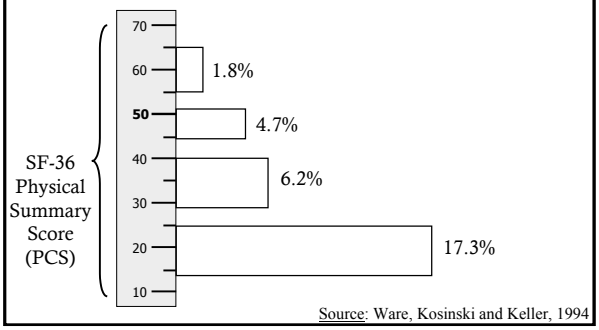


Predicting the Costs of Health Care



Source: Health Care Fin. Rev. 1997 vol 18 #4

Predicting 5-Year Mortality



Source: Ware, Kosinski and Keller, 1994

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)

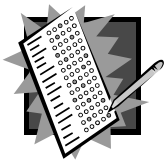
Interpreting a 5 Point Improvement in Physical Health (PCS)

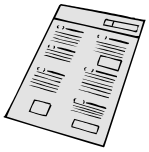
- Content-Based: Decrease from 28% to 13% limited in climbing stairs.
- Normative: A 5 point change in PCS is about half a standard deviation.
- Psychometric: Very unlikely to be due to chance.
- Disease Burden: Larger than the impact of Type II diabetes patients relative to the norm.
- Consequences:
  - 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

"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.

"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	

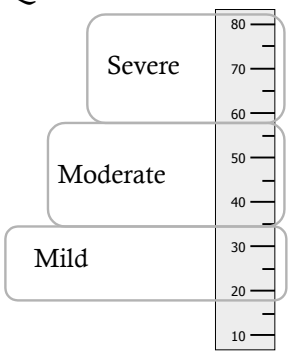
Adaptive Testing

- 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.

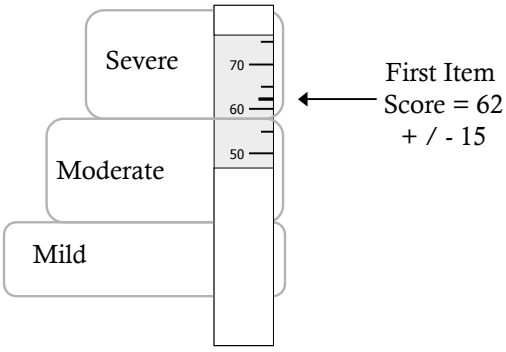
Adaptive Selection of Items

- 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.

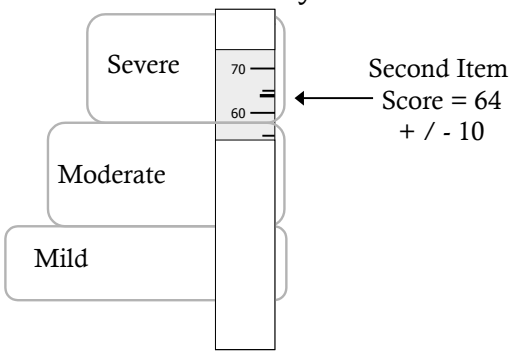
Dynamic Assessments Match Questions to Each Patient's Level



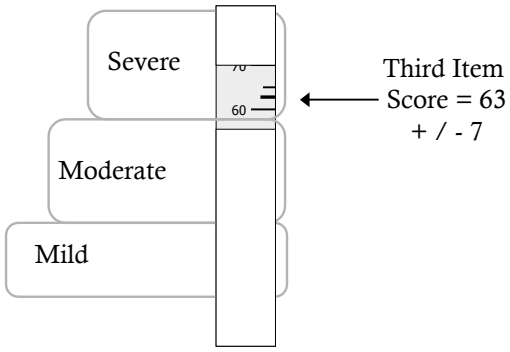
First Question: Noisy Score Estimate (+/- 15)



Second Question: Standard Error Reduced by One Third



Third Question: Standard Error Now Cut in Half



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.

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.

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.

"Maximum Information" Strategy for Adaptive Assessment

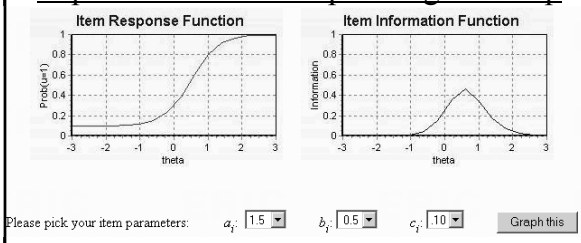
1. Evaluate all items not yet administered to determine which will be best to administer next given currently estimated trait level
2. "Best" next item administered, and subject responds
3. New estimate of trait computed using responses to all administered items
4. Repeat 1 through 3 until stopping criterion.

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

Item Information Function

- Estimates the information at various trait levels provided by an item.
- At what level is this item effective?
- How well can this item discriminate?
- Information Function is the reciprocal of the variance of the likelihood function…

Item Response Theory Tutorial at http://128.8.182.4/scripts/cat/genicc.asp



- An item provides its maximum information at ability level b.
- Amount of information provided is a function of the item's a and c parameters.

National Center for Fair and Open Testing

- Test-makers claims that the scores of computerized and pencil-and-paper tests are equivalent are inadequately supported.
- Computerized tests constrain test-takers compared to paper-and-pencil tests (only one item on the screen at a time, etc.)
- The additional cost of computerized tests is certain to have a large effect on who chooses to take them.

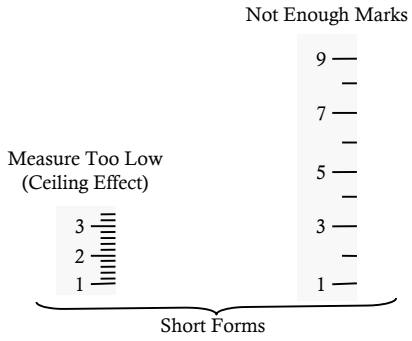
http://www.fairtest.org/facts/computer.htm

What's Wrong with Today's Outcome Measures?

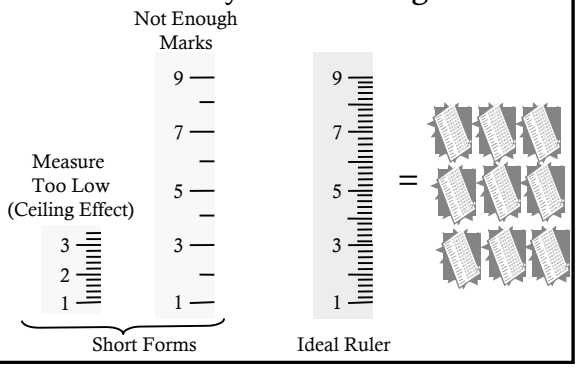
Health Status Questionnaires Are Not Practical

		Short Forms
• Face-to face interview	~~$300~~	$200
• Telephone interview	~~$75~~	$38
• Self-administered	~~$50~~	$15

Short-Form Surveys Alone Do Not Provide a Solution



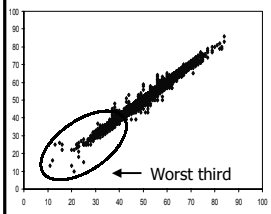
The Ideal Ruler is Not Practical Using the Old Way of Collecting Data



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
- Elimination of "ceiling" and "floor" effects

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

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

amIhealthy.com Dynamic SF-36

SF-36 Scales

- Physical Functioning [PF]
- Role Functioning – Physical [RP]
- Bodily Pain [BP]
- General Health Perceptions [GH]
- Vitality [VT]
- Social Functioning [SF]
- Role Functioning – Emotional [RE]
- Mental Health [MH]

Blurb for Patients

- The **Dynamic SF-36® Health Survey** is a state of the art "smart" survey designed to measure your overall health. After you answer the first few questions, the survey can customize the number and types of additional questions you get.

Question 1

Questions 2-7 for Patient 1

amIhealthy.com™
measuring your health...

Dynamic SF-36® Health Survey

2. During the PAST 4 WEEKS, how much did physical health problems limit your usual physical activities (such as walking or climbing stairs)?

- (●) Not at all
- ( ) Very little
- ( ) Somewhat
- ( ) Quite a lot
- ( ) Could not do physical activities

Next Exit

Copyright© 1999-2002 QualityMetric Incorporated

Questions 8-13 for Patient 1

amIhealthy.com™
measuring your health...

Dynamic SF-36® Health Survey

13. During the PAST 4 WEEKS, how much did personal or emotional problems keep you from doing your usual work, school or other daily activities?

- (●) Not at all
- ( ) Very little
- ( ) Somewhat
- ( ) Quite a lot
- ( ) Could not do daily activities

Next Exit

Copyright© 1999-2002 QualityMetric Incorporated

Dynamic Health Survey™

Report for Anonymous
July 3, 2002

YOUR SCORES

Report Date: July 3, 2002
Age: Unknown
Gender: Unknown
Conditions: Unknown

PHYSICAL HEALTH SUMMARY 53

53

30 WORST — 40 — 50 US AVG. — 60 — 70 BEST

Your physical health summary score is about average, taking into account the margin of error.

MENTAL HEALTH SUMMARY 53

53

30 WORST — 40 — 50 US AVG. — 60 — 70 BEST

Your mental health summary score is about average, taking into account the margin of error.

WHAT YOUR SCORES MEAN

Based on your answers about your health in the past 4 weeks, our research shows that:

Compared to the general population...

Physically, your...

- functioning is better than most
- pain is the same or less
- performance of work, home or school activities is slightly worse

Emotionally,...

- bothered less than most
- participation in social activities is less limited
- performance of work, home and school activities is limited more

Overall, your...

- rating of your health is better
- energy level is higher

Remainder of Report for Patient 1

YOUR PROGRESS

	Date	Physical Health Summary	Mental Health Summary
Current:	07/03/02	53	53

- We encourage you to become a registered user so that we can report and interpret your changes in scores over time.

WHAT YOU SHOULD DO

- In three months, take this survey again to continue monitoring your progress.
- By showing this report to your doctor or other health care provider, you can work together to make sure that your physical & mental health are as good as they can be.

Note: This survey is not a diagnostic tool. It is intended to supplement, but not replace or contradict the advice of your personal physician. If you have any questions or concerns about your health, it is always a good idea to seek one-on-one professional medical consultation.

Close

Copyright © 1999-2002 QualityMetric Incorporated

powered by DYNHA®

First 7 Items for Patient 2

amIhealthy.com™
measuring your health...

Dynamic SF-36® Health Survey

7. How TRUE or FALSE is the following statement for you? I am as healthy as anybody I know.

- ( ) Definitely true
- ( ) Mostly true
- ( ) Don't know
- ( ) Mostly false
- (●) Definitely false

Next Exit

Copyright© 1999-2002 QualityMetric Incorporated

Report for Patient 2 (saw 33 items)

Dynamic Health Survey™

Report for Anonymous
July 3, 2002

YOUR SCORES

Report Date: July 3, 2002
Age: Unknown
Gender: Unknown
Conditions: Unknown

PHYSICAL HEALTH SUMMARY 36

36

30 WORST — 40 — 50 US AVG. — 60 — 70 BEST

Your physical health summary score is very much below average, even taking into account the margin of error.

MENTAL HEALTH SUMMARY 54

54

30 WORST — 40 — 50 US AVG. — 60 — 70 BEST

Your mental health summary score is about average, taking into account the margin of error.

WHAT YOUR SCORES MEAN

Based on your answers about your health in the past 4 weeks, our research shows that:

Compared to the general population...

Physically, your...

- functioning is worse
- pain is much worse
- performance of work, home or school activities is worse

Emotionally,...

- bothered less than most
- participation in social activities is less limited
- performance of work, home and school activities is limited more

Overall, your...

- rating of your health is much worse
- energy level is lower

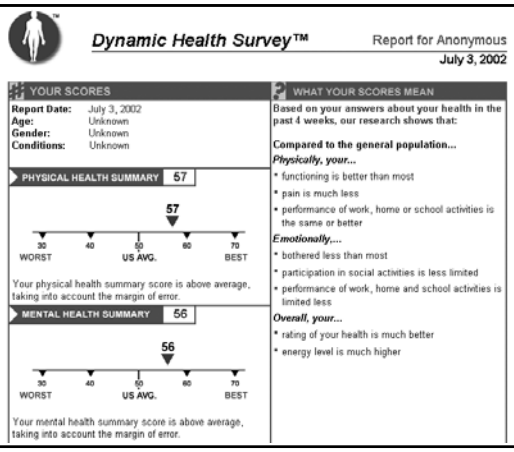
WHAT YOU SHOULD DO

On your next visit to the doctor be sure to discuss the impact health problems are having on your life.

In one month, take this survey again to continue monitoring your progress.

By showing this report to your doctor or other health care provider, you can work together to make sure that your physical & mental health are as good as they can be.

Question Set for Patient 3



Summary: Three Breakthroughs

- Modern psychometric models
- Computerized dynamic assessments
- Availability on the world-wide web

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

Internet Sources

- The Basics of Item Response Theory (Frank Baker) http://ericae.net/irt
- SF-36, SF-12, and SF-8 Surveys www.sf-36.com
- Dynamic Health Assessments www.qmetric.com www.amIhealthy.com www.headachetest.com
- Understanding Health Outcomes Videos Educational Series www.healthstatprod.com
- Latent Class Analysis http://ourworld.compuserve.com/homepages/jsuebersax/index.htm 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

Some Books

- Embretson SE, Reise SP (2000) *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Wainer H, ed. (2000) *Computerized Adaptive Testing: A Primer* (2nd Edition). Mahwah, NJ: Lawrence Erlbaum.
- Aday LA (1996) *Designing & Conducting Health Surveys* (2nd Edition). San Francisco: Jossey-Bass.
- Hambleton RK, Swaminathan H, Rogers HJ (1991) *Fundamentals of Item Response Theory*, volume 2. Newbury Park, CA: Sage.
- Lord FM (1980) *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.
- van der Linden WJ, Hambleton RK, eds. (1997) *Handbook of Modern Item Response Theory*. New York: Springer.
- van der Linden WJ, Glas CAW, eds. (2000) *Computerized Adaptive Testing: Theory and Practice*. Boston: Kluwerr-Nijhoff.

A Few Articles

- Bjorner JB, Kreiner S, Ware JE, Damsgaard MT, Bech P (1998) Differential item functioning in the Danish translation of the SF-36. *J Clin Epi*, 51(11): 1189-1202.
- Bjorner JB, Ware JE (1998) Using psychometric methods to measure health outcomes. *Medical Outcomes Trust Monitor*, 3(2): 11-16.
- Douglas JA (1999) Item response models for longitudinal quality of life data in clinical trials. *Statistics in Medicine*, 18: 2917-2931.
- Groenvold M, Bjorner JB, Klee MC, Kreiner S (1995) Test for item bias in a quality of life questionnaire. *J Clin Epi*, 48(6): 805-816.
- Juniper EF, Guyatt GH, Streiner DL, King DR (1997) Clinical impact versus factor analysis for quality of life questionnaire construction. *J Clin Epi*, 50(3): 233-238.
- Kantz ME, Harris WJ, Levitsky K, Ware JE, Davies AR (1992) Methods for assessing condition-specific and generic functional status outcomes after total knee replacement. *Medical Care*, 30(suppl. 5): MS240-MS252.
- Marx RG, Bombardier C, Hogg-Johnson S, Wright JH (1999) Clinimetric and psychometric strategies for development of a health measurement scale. *J Clin Epi*, 52: 105-111.
- Skevington SM, Bradshaw J, Saxena S (1999) Selecting national items for the WHOQOL: Conceptual and psychometric considerations. *Social Science & Medicine*, 48: 473-487.
- Wainer H (1989) The future of item analysis. *J of Educ Measurement*, 26: 191-208.
- Ware JE, Bjorner JB, Kosinski M (2000) Practical implications of item response theory and computerized adaptive testing. *Medical Care* 38(9 suppl. II): II-73 – II-82.
- Ware JE (2000) SF-36 health survey update. *SPINE* 25(24): 3130-3139.
- Wilson IB, Clearly PD (1995) Linking clinical variables with health related quality of life: a conceptual model of patient outcomes. *JAMA*, 273(1): 59-65.