NATIONAL RESEARCH
UNIVERSITY

6th International Scientific Practical Conference
“Health and Personality Development: Integrative Approach”

Use of Rasch Modeling in Research

Elena Kardanova
National Research University
Higher School of Economics, Moscow

Riga, 2020

Overview of the Session

- Introduction
- Classical Test Theory: advantages and limitations
- Modern Test Theory: IRT vs.Rasch Modeling
- Principles of Rasch measurement
- Item and scale analysis with Rasch modeling
- Detecting and measuring rater effects using Many-Facet Rasch Measurement
- Example of the research: IRT analysis of the Russian version of the Grit scale

Higher School of Economics / Institute of Education

- National Research University
- Established in 1995
- Young, but rapidly developing university with a focus on international partnership
- 4 campuses (Moscow, S.-Petersburg, Perm, Nizhny Novgorod)
- Over 30,000 students, 7,000 instructors and researchers
- Institute of Education at HSE: established in 2012 to carry out a research agenda in education at HSE University
- Master Program “Measurement in Psychology and Education” & PhD program “Educational Measurement and Evaluation”
- Annual International Summer school “Applied Psychometrics in Psychology and Education”

Classical test theory

Advantages

- based on relatively weak assumptions (i.e., they are easy to meet in real test data)
- well-known
- easy to apply

Main limitations

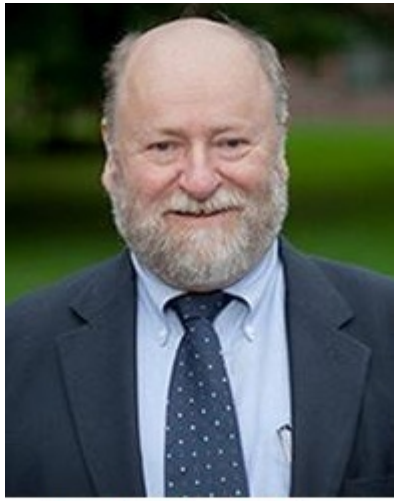
- Person parameters (i.e., true scores) are dependent on the test
- Item parameters (i.e., item difficulty and item discrimination) are sample dependent
- Scale of test scores is ordinal. No scale transformation raises the level of the scale
- Abilities levels and items difficulties are on different scales
- Doesn’t allow to solve satisfactorily many specific problems of testing: equating, item banking, computer adaptive testing, multidimensional modeling, analysis of additional factors (e.g.raters’ behaviour), etc.
- Poor opportunities for item analysis and scale construction, especially in psychology

Two lines of modern test theory: IRT line

This is the world as it was given to us and we have done our best to faithfully reproduce the observed item responses

Educational Testing Service (Princeton, NJ, USA)

- Lord & Novick “Statistical Theories of Mental Test Scores” (1968):
  - ✓ Statistical justification of the Classical Test Theory
  - ✓ 4 chapters devoted to IRT
- Scientific seminar: F. Lord, M. Novick, A. Birnbaum, S. Messick, F. Samejima, R. McDonald, W. Meredith



Ronald Hambleton

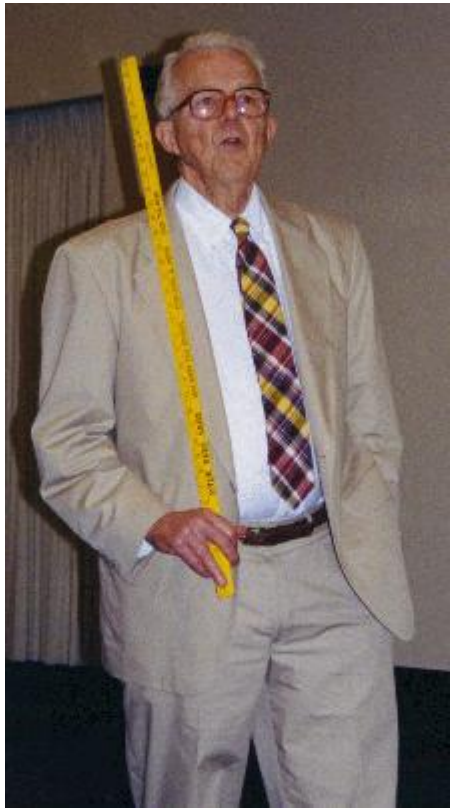
✓*Ron Hambleton and H. Swaminathan.* Item response theory: Principles and applications *(Kluwer, 1985)*

✓***Ron Hambleton, H. Swaminathan, and H. Jane Rogers*. Fundamentals of Item Response Theory (Sage, 1991)**

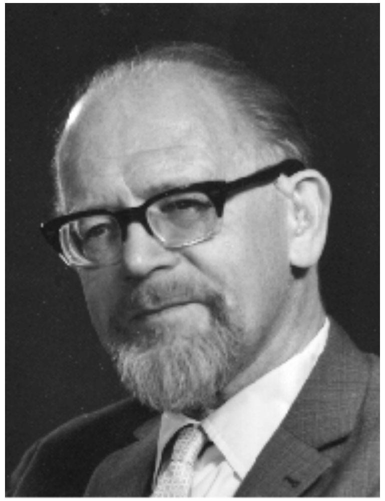
✓*Wim van der Linden and Ron Hambleton (edited).* Handbook of modern item response theory (Springer, 1997)

Two lines of the modern test theory: Rasch modeling

The Rasch model ... is not designed to fit any data, but instead is derived to define measurement



- **G.Rasch** “Probabilistic models for some intelligence and attainment tests” (1960)
- **B. Wright** (University of Chicago)
  - ✓ “Best Test Design” (1979, co-authored with Stone M.N)
  - ✓ “Rating Scale Analysis” (1982, co-authored with Masters G.N.).
- His students and followers: David Andrich, Geoffrey Masters, Mark Wilson, Mikle Linacre, Larry Ludlow, Carol Myford, and many others.



Georg Rasch (1901-1980)

Benjamin Wright (1926-2015)

Principles of Rasch Measurement

The measurement principles of Georg Rasch (1960) are:

(a) the items should be of the "same sort" or possess "uniformity of content"

(b) they should vary from "very easy to very difficult"

(c) "the difficult and the easy parts are spread evenly"

(d) their easy-to-difficult spread should follow a progression

(e) the items should be of equal discrimination

(f) the items should be independent in the sense that an answer to one is not dependent upon the answer to another

(g) "weeding" should be conducted so that "they on the whole fit well"

In current terminology these requirements may be stated as:

(a) Unidimensionality

(b) variation

(c) uniform spread of items along a continuum

(d) hierarchical in the nature of their progression along the continuum

(e) equally discriminating

(f) Independent

(g) well fitting as a match between theory and data.

These principles define how an instrument should be constructed, not simply whether a set of items may be described in these terms *post hoc.*

Rasch Measurement Development

Conceptual idea: how might we think about a "variable"?

- **Variable definition**
  - ✓What are we trying to measure? How should a variable "vary"? What does it mean to be characterized as "high" or "low" or "moderate" on a variable?
- **Unidimensionality**
  - ✓How many "things" should the variable measure? The world is complex. To understand behavior well, it often means we have to first break down the task into simple constructs
- **Continuum**
  - ✓What does movement along a variable imply? Every variable stretches across a continuum of simple-to-complex tasks, levels of knowledge, affective characteristics, or cognitive abilities

Family of Rasch models

> The family of Rasch measurement models is a way to make sense of the world
>
> Benjamin D. Wright

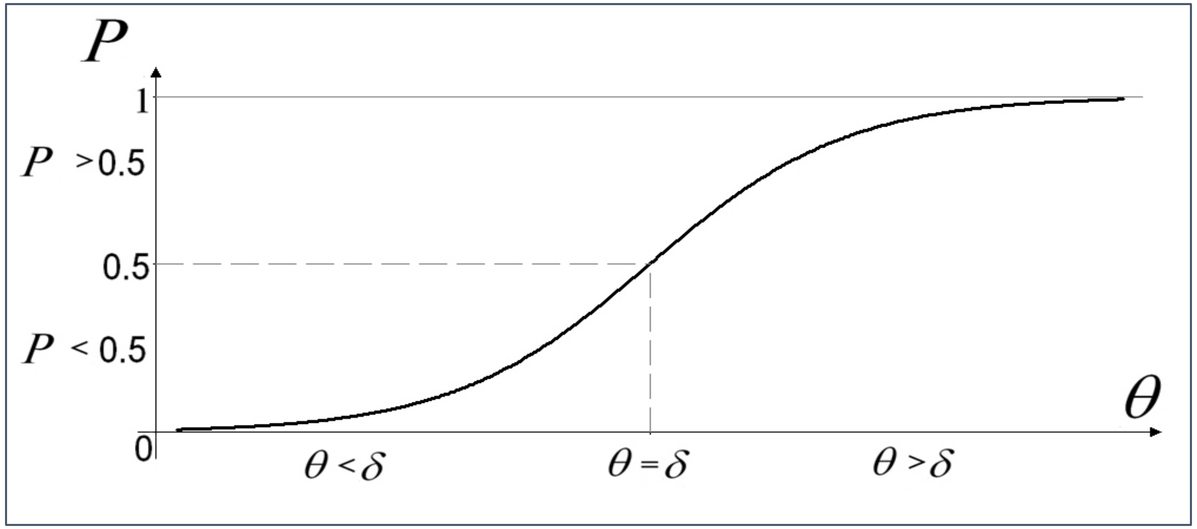
- Dichotomous Rasch Model
- Partial Credit Model
- Rating Scale Model
- Many-Facet Rasch Model
- Multidimensional Rasch Model

- All models are probabilistic
- The model sets the relationship between the probability of completing an item (or its part) / endorsing a certain category and the parameters of items and persons (and additional facets if necessary)

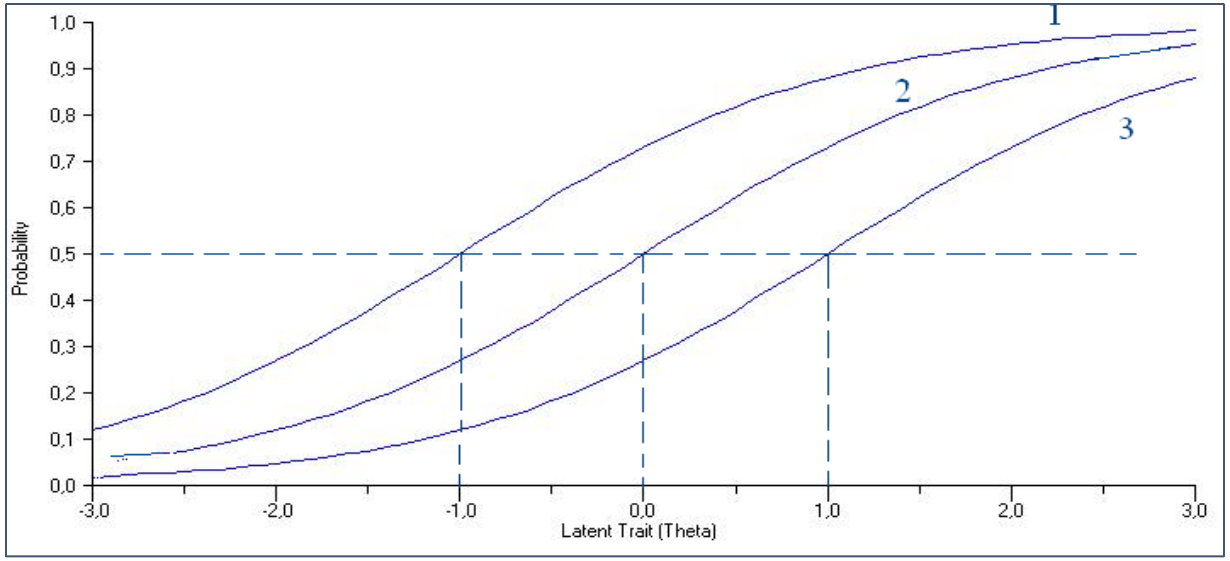
Rasch dichotomous model

$$P_{ni} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}$$

- P_{ni} is the probability of person *n* completing the item *i* (scoring 1 to an item)
- θ_n is the ability of person *n*
- δ_i is the difficulty of item *i.*



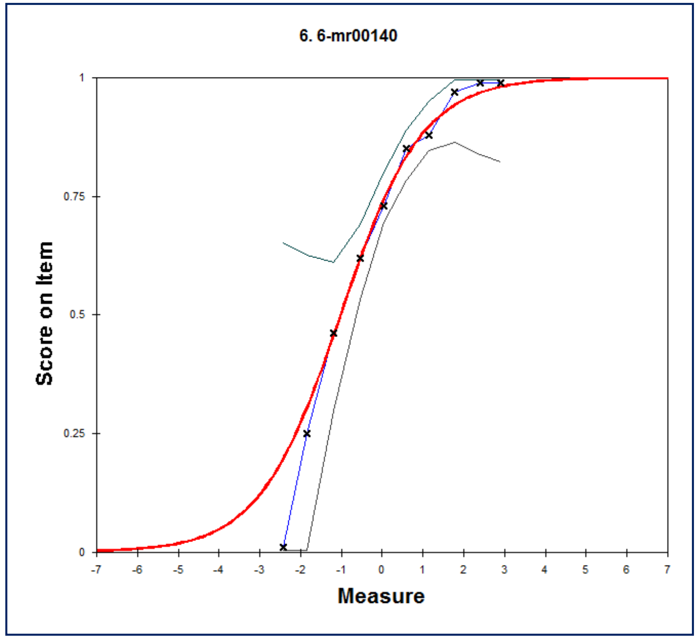
Item Characteristic Curve (ICC)



Advantages of Rasch modeling

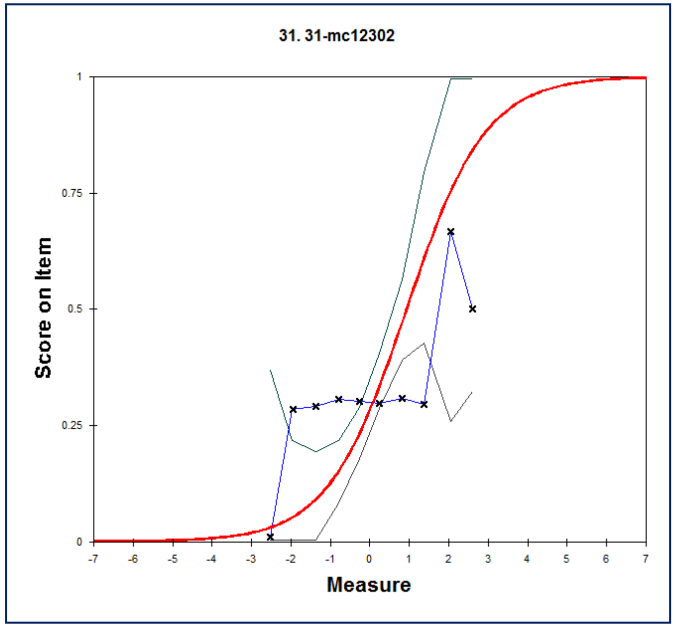
- Rasch models define the probability of response options as a function of latent trait levels, allowing the consideration of one or more traits simultaneously, allowing to implement both unidimensional and multidimensional models, the comparison of different models, allowing to include into consideration additional factors that can influence the probability of response (Embretson &Reise, 2000)
- Additionally, from a practical viewpoint, Rasch modeling offers the substantial advantage of providing a way to investigate the quality of test items, response categories for Likert-type scales, to study dimensionality, etc. (Linacre, 2002; Smith, 2002)

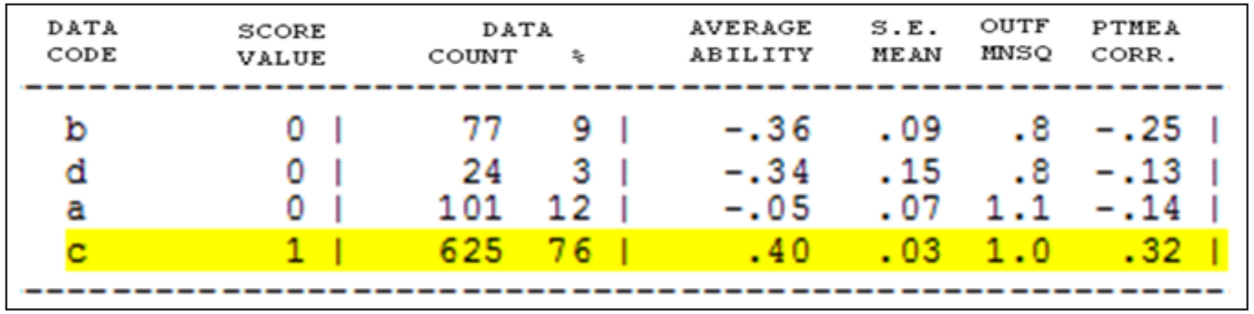
Item analysis with Rasch model: diagnostics of problems



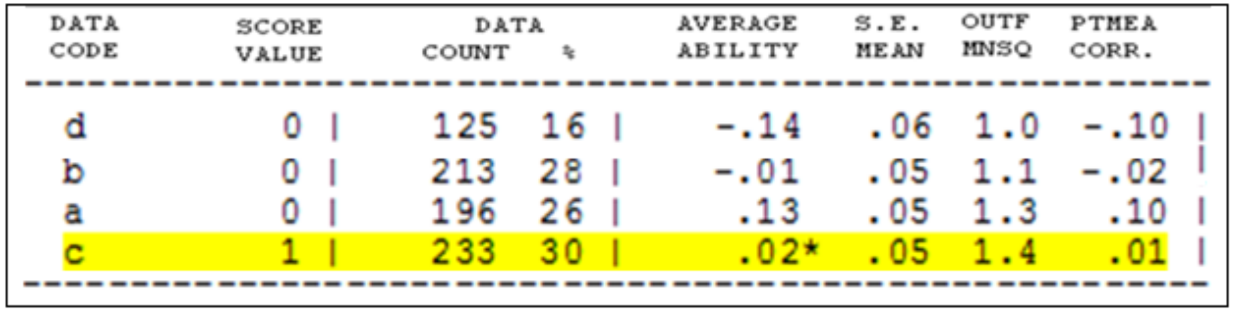
Good item example

Bad item example: random guessing





```
 DATA      SCORE       DATA       AVERAGE  S.E.  OUTF  PTMEA
 CODE      VALUE     COUNT   %    ABILITY  MEAN  MNSQ  CORR.
--------------------------------------------------------------
  b          0  |     77    9  |    -.36   .09    .8   -.25 |
  d          0  |     24    3  |    -.34   .15    .8   -.13 |
  a          0  |    101   12  |    -.05   .07   1.1   -.14 |
  c          1  |    625   76  |     .40   .03   1.0    .32 |
--------------------------------------------------------------
```

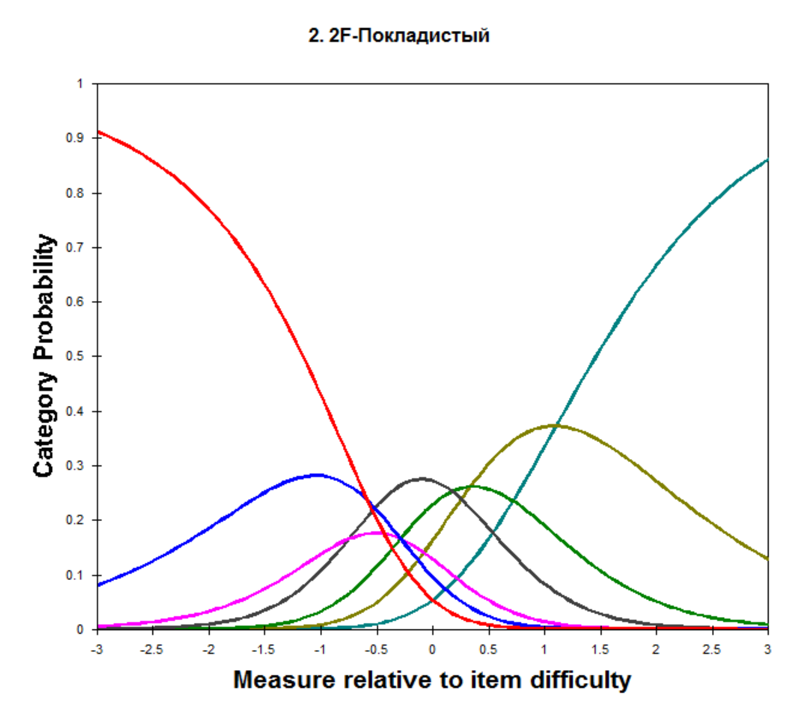


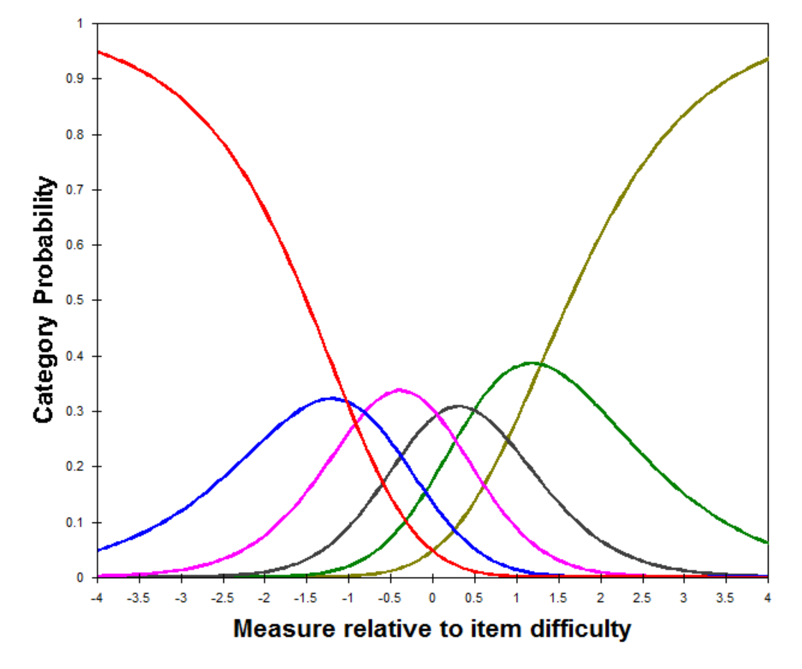
```
 DATA      SCORE       DATA       AVERAGE  S.E.  OUTF  PTMEA
 CODE      VALUE     COUNT   %    ABILITY  MEAN  MNSQ  CORR.
--------------------------------------------------------------
  d          0  |    125   16  |    -.14   .06   1.0   -.10 |
  b          0  |    213   28  |    -.01   .05   1.1   -.02 |
  a          0  |    196   26  |     .13   .05   1.3    .10 |
  c          1  |    233   30  |     .02*  .05   1.4    .01 |
--------------------------------------------------------------
```

76% of students selected the correct answer (c), and the students with the right answer were of higher ability on average (as measured by performance on other questions)

Many responses are at 25%. With 4 choices of answers, this suggests random guessing

Scale construction: optimization of category number





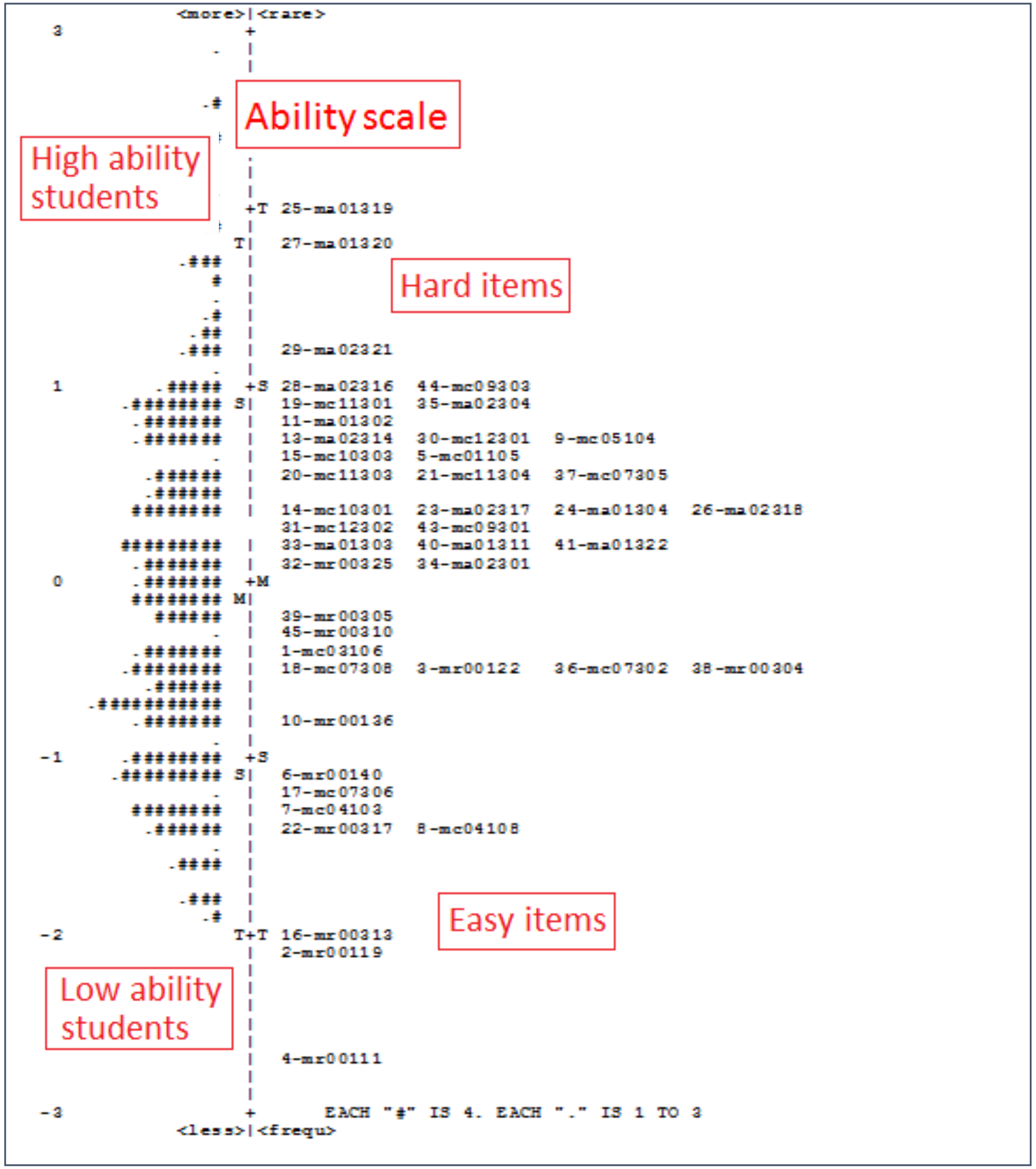
Using Rasch models, one can evaluate the following assumptions underlying the use of rating scales:

- ✓ Each consecutive number of the scale is to represent more of the latent construct under investigation.
- ✓ As one moves up the construct continuum, each point on the rating scale should, in turn, become the most probable response (i.e., respondents can reliably discriminate among all levels provided by the rating scale).

Rasch modeling allows to investigate the quality of the response categories.

- ✓ In the RSM context, a set of criteria has been posed for verifying useful functioning and diagnosing malfunctioning of a rating scale

Linacre, J.M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3, 85-106.



```
                 <more>|<rare>
   3                   +
                     . |
                       |
                    .# |
                       |
                       |
                       |
                       +T  25-ma01319
                       |
                      T|   27-ma01320
                  .### |
                     # |
                     . |
                    .# |
                   .## |
                  .### |   29-ma02321
                     . |
   1            .##### +S  28-ma02316  44-mc09303
           .######### S|   19-mc11301  35-ma02304
             .######## |   11-ma01302
             .######## |   13-ma02314  30-mc12301  9-mc05104
                     . |   15-mc10303  5-mc01105
              .####### |   20-mc11303  21-mc11304  37-mc07305
              .####### |
             ######### |   14-mc10301  23-ma02317  24-ma01304  26-ma02318
                           31-mc12302  43-mc09301
           ########### |   33-ma01303  40-ma01311  41-ma01322
             .######## |   32-mr00325  34-ma02301
   0         .######## +M
             ######### M|
                ###### |   39-mr00305
                     . |   45-mr00310
             .######## |   1-mc03106
           .########## |   18-mc07308  3-mr00122   36-mc07302  38-mr00304
              .####### |
       .############## |
             .######## |   10-mr00136
                     . |
  -1       .########## +S
         .########### S|   6-mr00140
                     . |   17-mc07306
             ######### |   7-mc04103
              .####### |   22-mr00317  8-mc04108
                     . |
                 .#### |
                       |
                  .### |
                    .# |
  -2                  T+T  16-mr00313
                       |   2-mr00119
                       |
                       |
                       |
                       |
                       |
                       |   4-mr00111
                       |
                       |
  -3                   +      EACH "#" IS 4. EACH "." IS 1 TO 3
                 <less>|<frequ>
```

Variable map

- Items (on the right) and students (on the left) are placed on the same scale
- The scale is interval
- Knowing the model, we can find for any person the probability of completing any item, regardless of whether it was offered to him or not.

Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement (MFRM)

- An extended version of the basic Rasch measurement models
- Makes possible the analysis of data that has more than the “traditional” two facets of examinees and items
- Considers other “facets” of interest in assessments that involve human judgment (raters, rater background variables, situational variables (e.g., time of day)

- *Myford Carol M., Wolfe Ed.W*. Detecting and Measuring Rater Effects Using Many-Facets Rasch Measurement: Part I // Journal of Applied Measurement, 2003, Vol. 4, № 4.
- *Myford Carol M., Wolfe Ed.W*. Detecting and Measuring Rater Effects Using Many-Facets Rasch Measurement: Part II // Journal of Applied Measurement, 2004, Vol. 5, № 2.

What kinds of questions can a MFRM analysis answer?

- Do raters differ in the levels of severity they exercise when rating? Just how interchangeable are they?
  - ✓ If raters differ in the levels of severity they exercise, which raters are rating more severely than others, and which are rating more leniently?
  - ✓ Is rater severity invariant across student subgroups? Does each rater behave in a similar way when rating all students? Are there certain raters who exercise differential severity, depending upon the race, ethnicity, or gender of the student?
  - ✓ Is rater severity invariant across time? Are there certain raters whose individual level of severity fluctuates depending upon the time of day, or the day of the week the rater is rating?
- Do raters differ in the consistency with which they apply rating scales?
  - ✓ If raters differ in the consistency with which they apply rating scales, which raters appear unable to use the scales in a consistent fashion?

A Catalog of Rater Effects

“Classic”

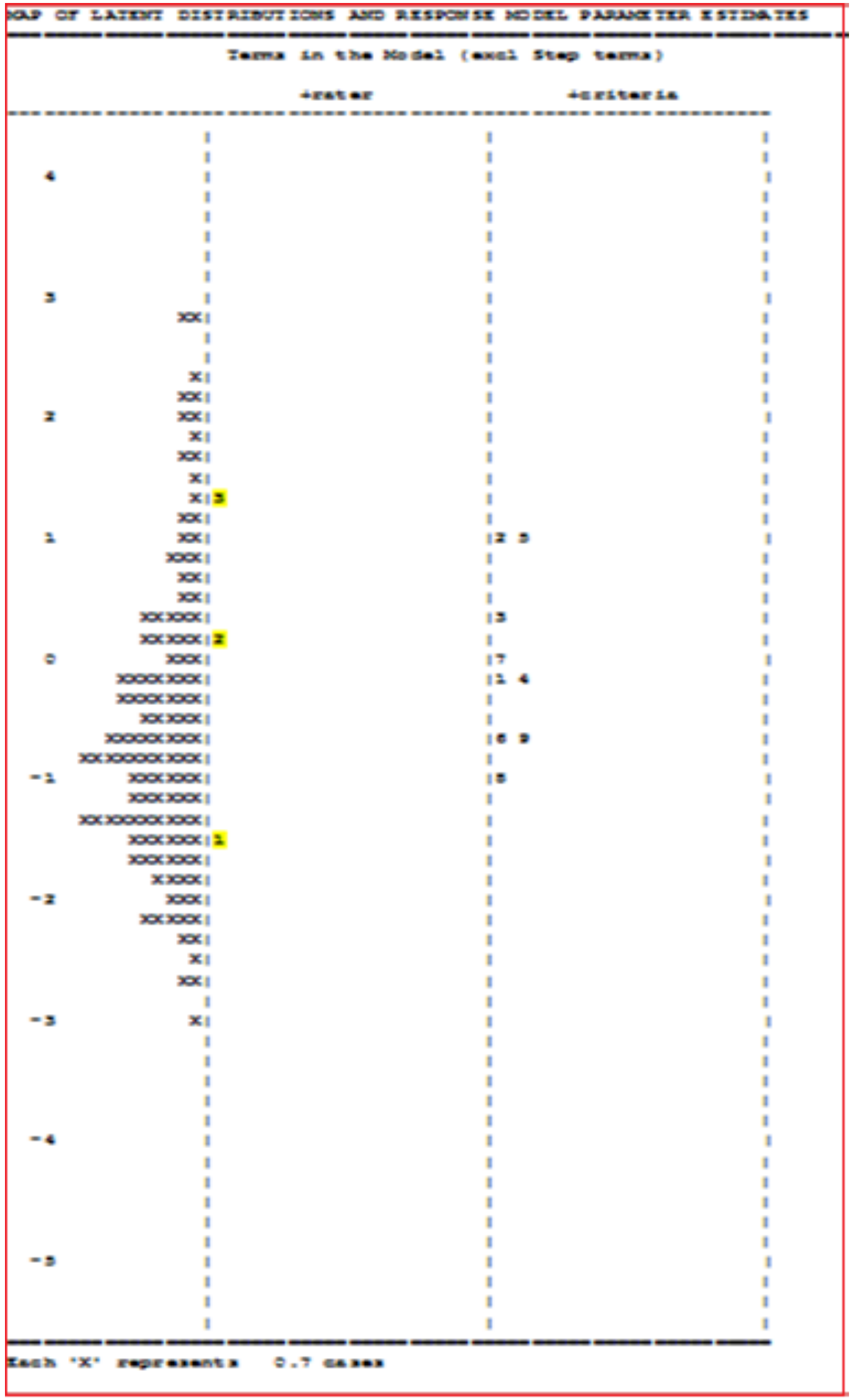
- leniency/severity
- halo
- central tendency
- restriction of range
- randomness

Other effects (less frequently mentioned, more difficult to detect):

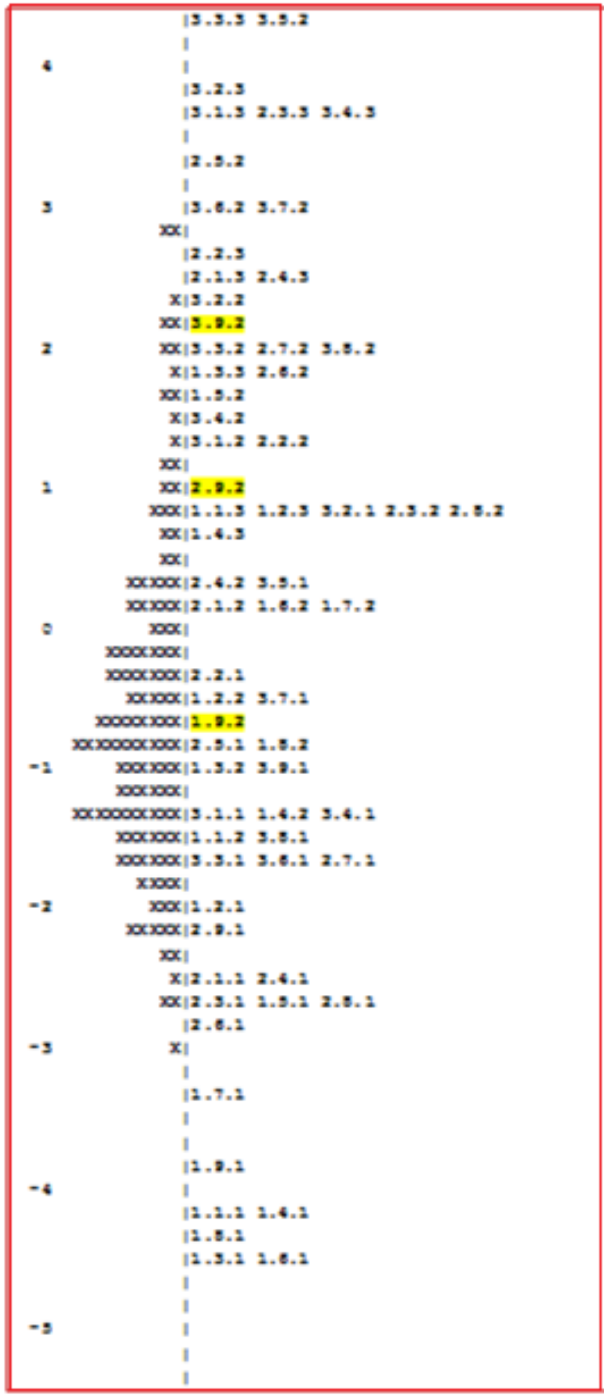
inaccuracy; influences of rater beliefs, attitudes, and personality characteristics; influences of rater background characteristics; etc.

“a rating involves an evaluative summary of past or present experiences in which the ‘internal computer’ of the rater processes the input data in complex and unspecified ways to arrive at the final judgment”

Thorndike and Hagen (1977)



```
MAP OF LATENT DISTRIBUTIONS AND RESPONSE MODEL PARAMETER ESTIMATES
==================================================================================
                       Terms in the Model (excl Step terms)
                   +rater                            +criteria
--------------------------------------------------------------------------
                |                                |                        |
                |                                |                        |
  4             |                                |                        |
                |                                |                        |
                |                                |                        |
                |                                |                        |
                |                                |                        |
                |                                |                        |
  3             |                                |                        |
              XX|                                |                        |
                |                                |                        |
                |                                |                        |
               X|                                |                        |
              XX|                                |                        |
  2           XX|                                |                        |
               X|                                |                        |
              XX|                                |                        |
               X|                                |                        |
               X|3                               |                        |
              XX|                                |                        |
  1           XX|                                |2 5                     |
             XXX|                                |                        |
              XX|                                |                        |
              XX|                                |                        |
           XXXXX|                                |3                       |
           XXXXX|2                               |                        |
  0          XXX|                                |7                       |
         XXXXXXX|                                |1 4                     |
         XXXXXXX|                                |                        |
           XXXXX|                                |                        |
        XXXXXXXX|                                |6 9                     |
      XXXXXXXXXX|                                |                        |
 -1       XXXXXX|                                |8                       |
          XXXXXX|                                |                        |
      XXXXXXXXXX|                                |                        |
          XXXXXX|1                               |                        |
          XXXXXX|                                |                        |
            XXXX|                                |                        |
 -2          XXX|                                |                        |
           XXXXX|                                |                        |
              XX|                                |                        |
               X|                                |                        |
              XX|                                |                        |
                |                                |                        |
 -3            X|                                |                        |
                |                                |                        |
                |                                |                        |
                |                                |                        |
                |                                |                        |
                |                                |                        |
 -4             |                                |                        |
                |                                |                        |
                |                                |                        |
                |                                |                        |
                |                                |                        |
                |                                |                        |
 -5             |                                |                        |
                |                                |                        |
                |                                |                        |
                |                                |                        |
==========================================================================
Each 'X' represents    0.7 cases
```



```
                |3.3.3 3.5.2
                |
  4             |
                |3.2.3
                |3.1.3 2.3.3 3.4.3
                |
                |2.5.2
                |
  3             |3.6.2 3.7.2
              XX|
                |2.2.3
                |2.1.3 2.4.3
               X|3.2.2
              XX|3.9.2
  2           XX|3.3.2 2.7.2 3.8.2
               X|1.3.3 2.6.2
              XX|1.5.2
               X|3.4.2
               X|3.1.2 2.2.2
              XX|
  1           XX|2.9.2
             XXX|1.1.3 1.2.3 3.2.1 2.3.2 2.8.2
              XX|1.4.3
              XX|
           XXXXX|2.4.2 3.5.1
           XXXXX|2.1.2 1.6.2 1.7.2
  0          XXX|
         XXXXXXX|
         XXXXXXX|2.2.1
           XXXXX|1.2.2 3.7.1
        XXXXXXXX|1.9.2
      XXXXXXXXXX|2.5.1 1.8.2
 -1       XXXXXX|1.3.2 3.9.1
          XXXXXX|
      XXXXXXXXXX|3.1.1 1.4.2 3.4.1
          XXXXXX|1.1.2 3.8.1
          XXXXXX|3.3.1 3.6.1 2.7.1
            XXXX|
 -2          XXX|1.2.1
           XXXXX|2.9.1
              XX|
               X|2.1.1 2.4.1
              XX|2.3.1 1.5.1 2.8.1
                |2.6.1
 -3            X|
                |
                |1.7.1
                |
                |
                |1.9.1
 -4             |
                |1.1.1 1.4.1
                |1.8.1
                |1.3.1 1.6.1
                |
                |
 -5             |
                |
                |
```

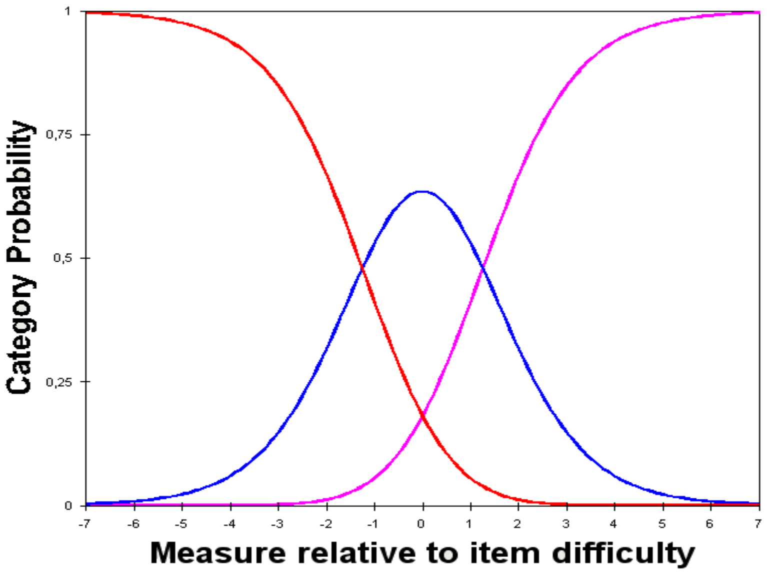
Variable maps

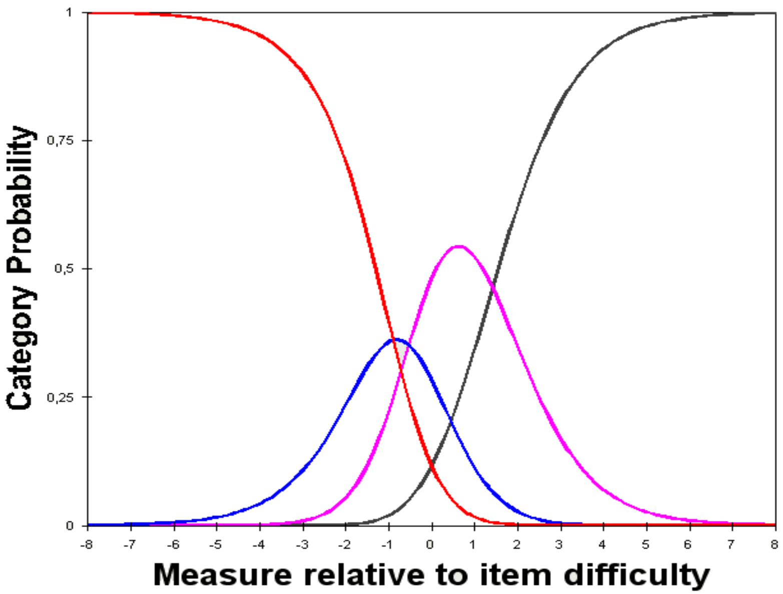
Comparison of three raters differing in severity level

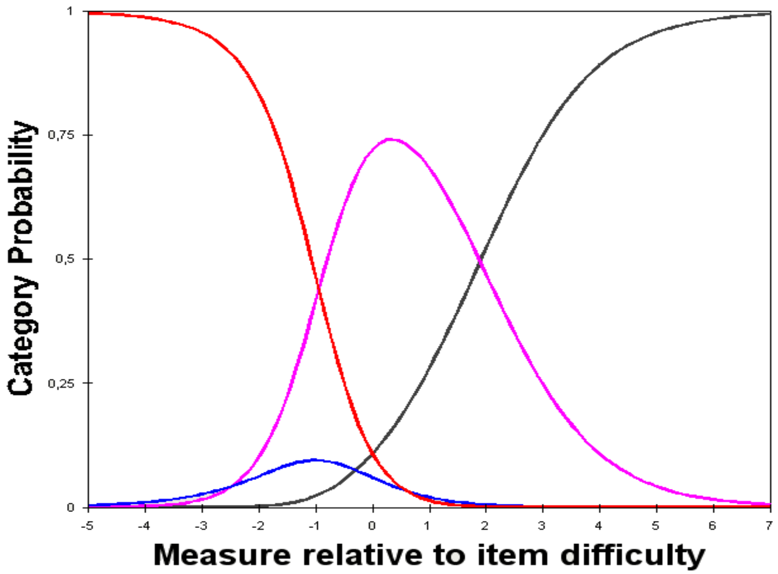
The difficulties of achieving each category in each item when evaluated by each rater are shown

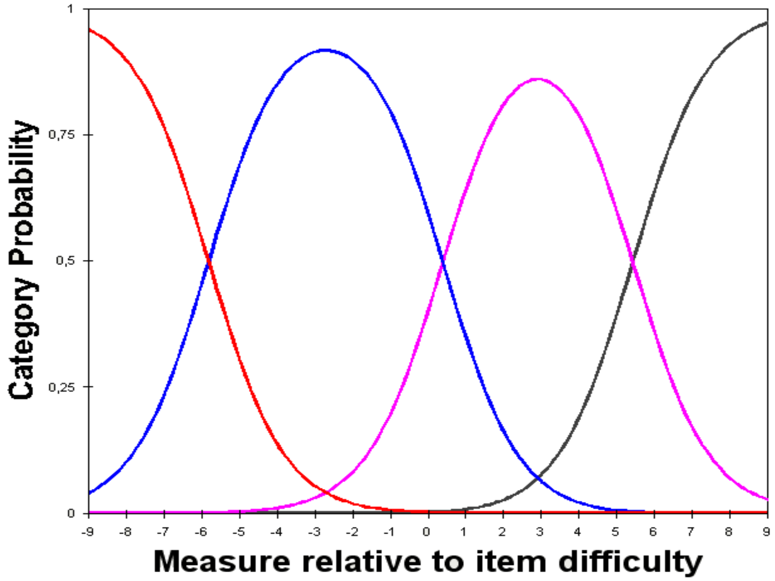
The raters differ essentially in the levels of severity they exercise when rating

Comparison of raters









Category probabilities curves for the rating scale when evaluated by different raters are shown.

The raters differ essentially in the consistency with which they apply rating scales.

Example of the research:
Questioning the grit construct: IRT analysis of the Russian version of the Grit scale

Tyumeneva Y.F., Kardanova E., Kuzmina Y. (2019). Grit: Two Related but Independent Constructs Instead of One. Evidence from Item Response Theory. *European Journal of Psychological Assessment*. Vol. 35. No. 4. P. 469-478

Grit scale

Grit construct is the ability to maintain long-term goals and overcome difficulties in the course of achieving them (Duckworth, Peterson, Matthews, &Kelly, 2007)

The original version of the scale includes two subscales:

- **Consistency of Interests (CI)**
- **Perseverance of Effort (PE)**

Duckworth, A. L., Peterson, C., Matthews, M. D., &Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087–1101.

Structure of the Grit scale (according to Duckworth, 2007)

Consistency of Interests	Perseverance of Effort
I often set a goal but later choose to pursue a different one*	Without irony, I am a hard worker
New ideas and new projects sometimes distract me from previous ones*	I have achieved a goal that took years of work
I become interested in new pursuits every few months*	I have overcome setbacks to conquer an important challenge
My interests change from year to year*	As a rule, setbacks discourage me
I have been obsessed with a certain idea or project for a short time but later lost interest*	I finish whatever I begin
I have difficulty maintaining my focus on projects that take more than a few months to complete*	At work, I am diligent

Each item was rated on a five-point Likert scale that ranged from 1 (*do not agree at all*) to 5 (*completely agree*).

Assessing Validity via Item Internal Structure

Contradictory findings

Several studies have confirmed the two-factor internal structure (e.g., Eskreis-Winkler, Shulman, Beal, & Duckworth, 2014; Maddi, Matthews, Kelly, Villarreal, & White, 2012; Pozzebon et al., 2013).

BUT

increasing amount of evidence that CI and PE subscales can reflect independent constructs rather than aspects of a single grit construct (e.g. Datu, Valdez, & King (2015) found that grit involves two distinct dimensions rather than a hierarchical construct).

What does "grit score" mean? To what extent a composite overall score makes sense as an indicator of any substantial single construct?

The purpose of the study

To present psychometrical arguments that CI and PE factors do not form the whole measure and instead should be considered independent, although still interrelated, unidimensional measures of different traits.

- Previously, psychometric properties of the Grit scale were assessed using the classical test theory (CTT)
- Rasch modeling has the potential to supplement CTT in a number of important ways

Model selection

- Three approaches in the item response modeling to the Grit scale:
  - ✓we can ignore the multidimensionality of the scale and apply a unidimensional model
  - ✓we can recognize multidimensionality and apply a unidimensional model to each dimension consecutively
  - ✓we can apply multidimensional models

Summary of Unidimensional and Multidimensional Model Scaling

Model	Number of Parameters	Deviance	Reliability	AIC	BIC
RSM	16	71,784.83	.74	71,816.83	71,908.51
PCM	49	71,293.49	.73	71,391.49	71,672.26

Note. RSM = Rating Scale Model; PCM = Partial Credit Model; Deviance = - 2 log likelihood; Reliability = person reliability index; AIC = Akiake information criterion; BIC = Bayesian information criterion.

Model	Number of Parameters	Deviance	Reliability	AIC	BIC
RSM	18	70,185.76		70,221.76	70,324.90
Dimension 1			.77		
Dimension 2			.73		
PCM	51	69,260.31		69,362.31	69,654.54
Dimension 1			.70		
Dimension 2			.81		

Note. Dimension 1 = CI scale; Dimension 2 = PE scale; other abbreviations are the same.

Multidimensional PCM fits the data much better than unidimensional models and multidimensional RSM

Summary of Consecutive Model Scaling

We recognize multidimensionality and apply a unidimensional model to each dimension consecutively

Model	Number of Parameters	Deviance	Reliability	AIC	BIC
RSM					
Dimension 1	11	42,866.75	.70	42,888.75	42,951.78
Dimension 2	9	26,900.78	.78	26,918.78	26,970.35
PCM					
Dimension 1	29	42,632.60	.70	42,690.60	42,856.77
Dimension 2	21	26,763.13	.79	26,805.13	26,925.46

PCM fits the data much better than the RSM

Multidimensional approach (two-dimensional PCM model) vs. consecutive approach (two unidimensional PCM models for two subscales)

- Under both approaches, the reliability coefficients for each dimension were almost the same.
- The correlation between respondents’ abilities under the two approaches was .99 for both subscales, and the correlation between the items’ difficulties was 1.00.

Conclusions:

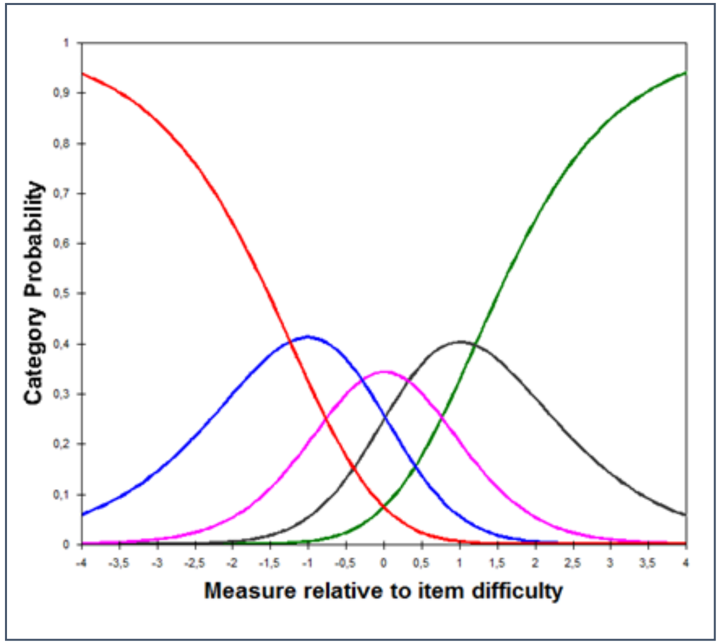
✓both approaches provide good explanations for the Grit scale data

✓the Consistency of Interests and Perseverance of Effort subscales reflect related but independent constructs rather than form the single construct Grit.

✓for further analysis, we can select any of these approaches.

Properties of subscales

Subscale	Rasch Reliability	Classical Reliability (alpha)	Ability Distribution Mean (SD)	Standard Error of Ability Measurement Mean (SD)
Consistency of Interests	.70	.75	.22 (.86)	.44(.09)
Perseverance of Effort	.81	.83	.90 (1.47)	.67(.13)



Category probability curves for the CI rating scale

Conclusions

- The Grit scale deals, in fact, with two different traits: consistency of interests and perseverance of effort
- Both subscales in the multidimensional approach exhibited good psychometric properties and quality in the 5-item response categories. Thus, they can be used as separate measures of different traits.

Thank you

ekardanova@hse.ru