Role of Modeling When Located Between Theory and Data

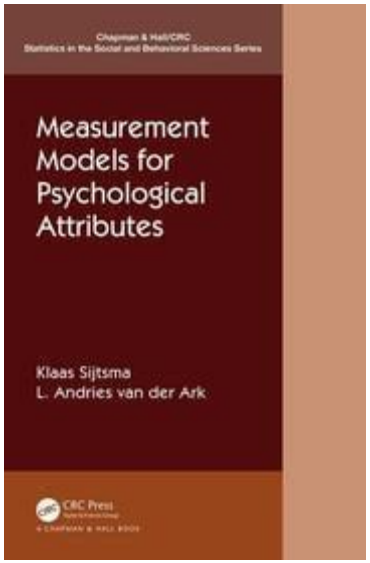
Klaas Sijtsma

Tilburg University
Understanding Society

Lecture on occasion of the BEAR Center Seminar, 2 November 2021

Sijtsma, K., & Van der Ark, L. A. *Measurement models for psychological attributes*. Boca Raton, FL: Chapman & Hall/CRC

Contents



Doesn't such an anthology of measurement models exist already? Existing books focus on

- One model: Mokken Model, Rasch Model, Latent Class Analysis, Cognitive Diagnostic Models
- One broader class: Classical Test Theory, Item Response Theory, Generalizability Theory, Factor Analysis
- Classical Test theory and Item Response Theory together, to show that IRT improves upon CTT
- Dimensionality: Unidimensional models, Multidimensional Scaling / MIRT
- Topics (within or across models): Adaptive Testing, Test Equating, Differential Item Functioning, Achievement Testing, Testlet Response Theory, Validity

Some older books have a broader scope:

- Torgerson (1958; *Theory and Methods of Scaling*)
- Coombs (1964; *A Theory of Data*)

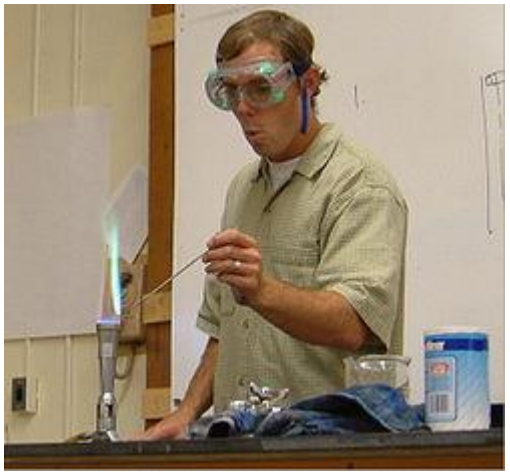
But they have very little overlap with our book; only pairwise comparison and proximity models overlap

Three-volume work by Krantz, Luce, Suppes & Tversky (e.g., Volume I, *Foundations of Measurement)* represents

- A totally different, non-psychometric and broader scientific perspective on measurement
- Based on an axiomatic approach
- That is often deterministic, and highly restrictive
- That was not (convincingly) adopted up by human sciences (and unsuited, I add)

Where does my fascination for measurement come from?

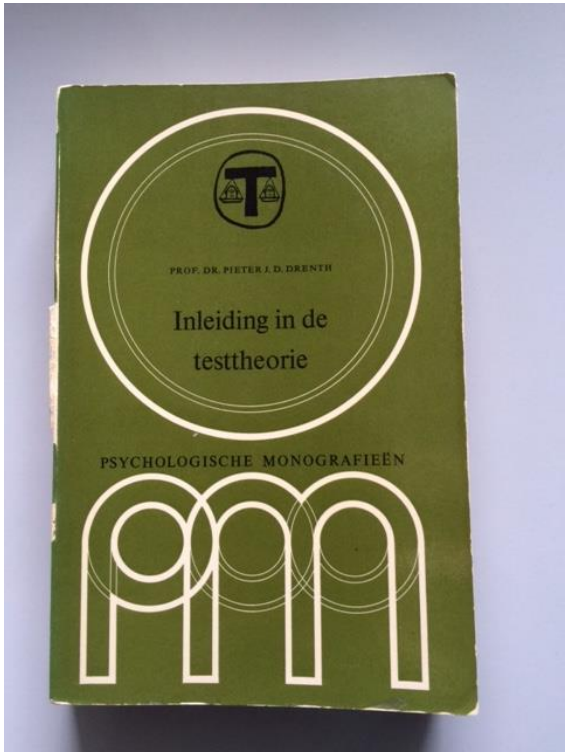
- In 1974, I started as a student of Pharmacy:
- Much lab work, determining
  - ✓ **quantitative** measurement (concentration chemical element, level of radioactivity, or strength of electrical current) → **AMOUNT**
  - ✓ **qualitative** measurement (chemical elements in a mixture) → **TYPE**



- A memory of measurement and early statistics (which is rather unlike physics in those days):
  - Ammeter for measuring electrical current
  - Typically, the needle oscillates across the scale
  - Estimate by observation most likely value, minimum value, and maximum value
  - Computations for all three values, produced "confidence interval"

- Boring for an 18/19-year old; changed to psychology

Next, I started psychology, and in my 2nd year read



- Amazed to find that psychologists measured attributes like intelligence and introversion
- And how they did this: With lists of problems one had to solve or questions one had to answer, and counts of number-correct or counts of credit points earned
- And not with apparatus picking up a physical phenomenon, and needles and scales showing a quantity
- Reliability concept replaced oscillation of needle and validity concept replaced physics theory and natural laws
- Psychological measurement was not without problems

This is how I became interested in measurement (if my memory isn't accurate, it's a good story anyway)

In the nonphysical sciences, measurement has always been problematic, and it has become increasingly evident to nearly everyone concerned that we must devise theories somewhat different from those that have worked in physics

—David H. Krantz, R. Duncan Luce, Patrick Suppes, & Amos Tversky (1971)

Psychometrics presents mathematical/statistical measurement models

Models are defined by *assumptions* about

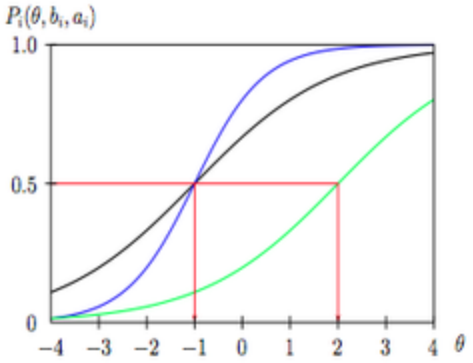
- Complexity of the measurement (i.e., dimensionality)
  - *Analytical reasoning (confounder: language skills?)*
  - *Electrical current/amperage (confounder: resistance?)*

- Internal structure of the measurement (e.g., local independence)

  Does measurement instrument pick up training effects, heat?

- Relation of items with the dimensions (e.g., response functions)

✓ *Do analytical reasoning tasks discriminate particular measurement levels (location parameter) and with what strength (discrimination parameter)?*



✓ *Is electrical voltage strong enough to make coffee grinder work?*

Other assumptions:

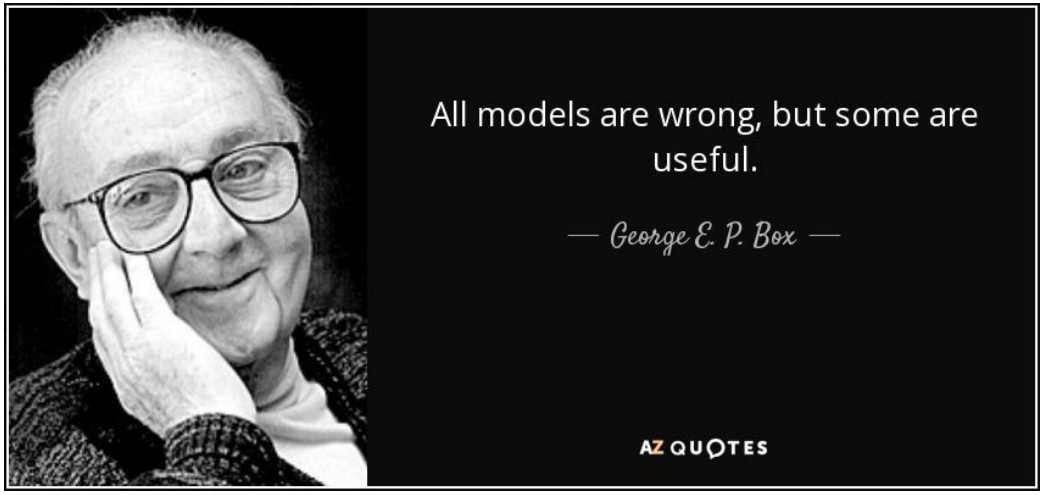
- Measurement error: random / systematic; distribution, correlations (classical test theory, factor analysis)
- Ordering of *types* (latent classes) rather than levels
  - ✓ *Developmental progression of solution strategies for fraction arithmetic*
  - ✓ *Evolutionary progression of species*
- Subattributes working together to solve a problem with a particular probability (cognitive diagnostic models)

Models imply *expected* structure of the data the researcher collected by means of the (preliminary, experimental) measurement instrument (test, questionnaire)

Goodness-of-fit research involves

- Comparing *expected* data (features) with *real* data (features), and
- Study the inconsistency or the misfit
- To ascertain whether we have a scale

Typical (i.e., unavoidable) result is **misfit**



Models are **idealizations**, that is, **simplifications**, of the phenomenon they try to describe or explain; one hopes they pick up the salient characteristics

Models *must* fail at describing the data; if they didn't, they would coincide with the data and provide no distinction between **main principles** and **details**

"Box" applied to measurement: Question is whether the misfitting model still provides a useful scale, e.g., having *predictive validity* (school, job, therapy)

In the social, behavioral, and health sciences models fail *notably*; what are causes?

Data collected in the human sciences are *messy*, meaning they contain

- Much *noise*
- Many *signals*, often weak, some mediocre, few strong

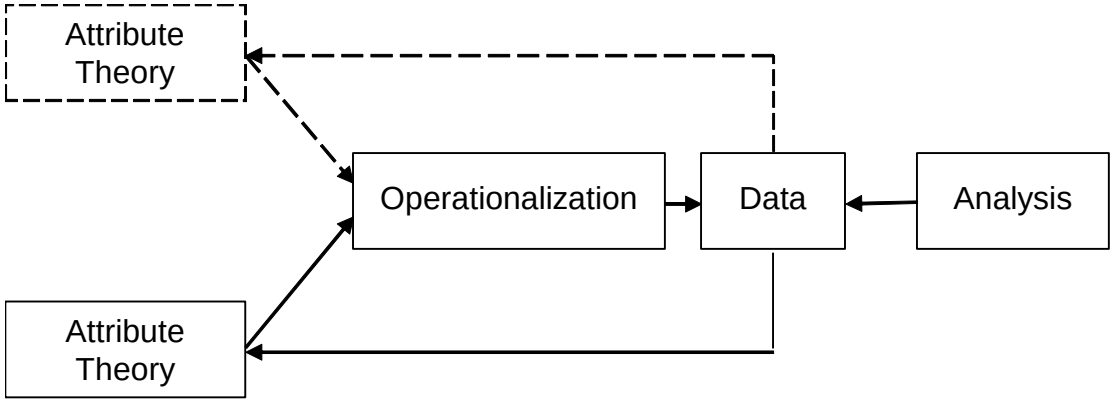
In measurement, causes are

1 Absence of well-founded ***theory*** about the target attribute (ability, trait, attitude), presence of partly-founded, incomplete theory (exceptions do exist)
2 Tendency of people to respond to being subject of an assessment procedure, known as ***participant reactivity***; in the human sciences, research object "talks back"
3 Random and systematic ***error sources*** beyond the researcher's control, other than imperfect attribute theory and participant reactivity

1—***Attribute Theory***; that is, theory about the attribute to be measured

Well-founded theory (solid box) provides *guidance* for operationalization, known as measurement prescriptions, providing link between concepts and behaviors

Absence of well-founded theory withholds guidance, researcher relies on experience, habit & tradition, educated guesses; measurement is off target to an unknown degree



Scheme for the construction of psychological measurement instruments

Analysis

- Analysis involves using measurement model to study goodness-of-fit model to data
- Data collected by means of preliminary test/questionnaire may or may not (or partly) reflect attribute
- Goodness-of-fit research (GoF) makes the most sense if one expects particular outcome, results are fed back to theory, help develop theory further and improve instrument
- GoF research makes less sense when expectation is not well articulated; then, outcome is little more than a clustering of some items, and in the absence of theoretical expectations, data exploration produces uncertain results for unidimensionality, monotonicity, invariant item ordering

2—***Participant Reactivity*** (Rosenthal & Rosnow, 1969)

Some examples:

- When asked about their introversion level, people *reflect* on their personality and this affects the answers they give (consistent pattern, ideal picture); person-dependent, unpredictable
- Someone nervous may (*unintendedly*) produce negatively biased results on an achievement test
- Another person may (*intendedly*) manipulate personality inventory to give socially desirable answers when they expect truthful answers to be disadvantageous
- *Demand characteristics* (Rosenthal & Rosnow, 1969): Features in the testing context inspire people to accommodate their responses to the perceived context; e.g., harsh item formulation may soften or harden responses of some people
- *Experimenter expectancy* (R&R, 1996): Test practitioner sends signals affecting some people; e.g., oral instruction to relax may signal that test is not important when it actually is, and may worsen test performance

Notice: Natural sciences pay much attention to realize *unobtrusive measurement*

3—Uncontrolled ***Error Sources***

Some examples:

- Noisy testing environment may negatively affect some people's performance
- Pleasant room with coffee and tea may have opposite effect on some people
- Temporary person-dependent mental state (bad sleep, good sleep; recent nasty or pleasant personal experiences)
- Language skills always influence test performance and increase individual differences

Do psychologists and psychometricians pay **enough attention** to these issues?

- *Psychologists* (and others) primarily focused on the *substantive* aspects of their research; statistics and psychometrics (measurement) are practiced *on the side*
- Used to be more focus on *theory development* and *measurement* & *standardization*; replaced with Internet platform data (Amazon MTurk) and e-health data (smart watches); where is the *standardization*?
- Researchers are inclined to engage in employing *researcher degrees of freedom* to optimize publication opportunities; is the situation deteriorating?

Cf. chemical measurement (ban pollution) and physical measurement (CERN); standardization has **Top Priority**

And ……

> *Statisticians*, like artists, have the bad habit of falling in love with their models
>
> —George Box

Not enough attention for substantive and methodological issues ……

The book chapters ……

Chapter 2 **Classical Test Theory and Factor Analysis**

Classical Test Theory for two items j and k does **not model** their **association**:

$$X_j = T_j + E_j \text{ and } X_k = T_k + E_k$$

True scores depend on item, origin item scores unknown; CTT is about **repeatability**

Factor Analysis does **model** the association between items:

$$X_j = b_j + \sum_{m=1}^{M} a_{jm}\xi_m + \delta_j \text{ and } X_k = b_k + \sum_{m=1}^{M} a_{km}\xi_m + \delta_k$$

M latent variables or factors ξ underlie performance on different items and tie them together; FA approach is about **modeling**, pulls **validity** in ……

Confounding of reliability and validity?

Chapter 3 **Nonparametric Item Response Theory and Mokken Scale Analysis**

Define weak(est) assumptions allowing **quantitative** person ordering:

- Unidimensionality; one latent variable θ (mathematical entity!)
- Local independence; no other influences
- $P(X_j \geq x|\theta)$ monotone in θ; higher scale value, higher response probability

$\Longrightarrow$ Ordering persons on total score $X = \sum_j X_j$ equal to ordering on θ (do not need θ)

Emphasis on

- Identifying dimensionality, subsets of items with different scales
- Estimating IRF, $P(X_j \geq x|\theta)$, to assess item strengths and weaknesses

Chapter 4 **Parametric Item Response Theory and Structural Extensions**

Typical is choice of parametric functions for $P(X_j \geq x|\theta)$; for example

$$P(X_j = 1|\theta) = \frac{\exp[\,\alpha_j\,(\theta - \delta_j)]}{1 + \exp[\,\alpha_j\,(\theta - \delta_j)]}$$

$$P(X_j \geq x|\theta) = \sum_{y=x}^{M} P(X_j = y|\theta) = \frac{\exp[\alpha_j(\theta - \lambda_{jx})]}{1 + \exp[\alpha_j(\theta - \lambda_{jx})]}$$

More structure, parameters take over from functions:

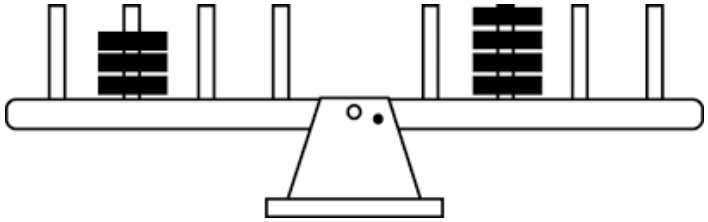
- Summary of statistical information more succinctly, lose more details
- Better connection to regular statistics
- Special cases of Nonparametric IRT, see Venn diagrams

Chapter 5 **Latent Class Models and Cognitive Diagnostic Models**

Define assumptions allowing **qualitative** person classification / taxonomy:

- Latent Class Models
- Cognitive Diagnostic Models—IRT Models and Latent Class Models

Proportional reasoning as an example



- Children use strategy to solve problems
- Strategies are incorrect, but
  - ✓ Produce "solutions"
  - ✓ Inform about development
- Identify strategies, assign children; total score X **uninformative**

CDMs define subattributes/skills and mechanisms to explain response probabilities

Chapter 6 **Pairwise Comparison, Proximity, Response Time, and Network Models**

Response time models

- In addition to correct/incorrect data, models include response times
- Other models include auxiliary information, like distraction data
- (Response times, distraction data, go back to 19th century psychophysiology)

My Hypothesis:

- Most information is in correct/incorrect data, and even more in the process data behind them

Network models

- No latent variables, observable variables affect one another; take depression, feeling low affects sleep quality, sleep deprivation affects social contacts, etc.
- Types of depression? Stages of depression?
- Connections with **qualitative** measurement? Observable classes?

Progress in psychological measurement must come from

- **Psychologists** engaging in ***theory development***, do away with *economic* / *fast results* approach to psychology, give in to *slow science*
- **Psychometricians** engaging in ***theory development***, integrating psychometrics again into psychology, give in to being a **quantitative psychologist**
- There are some (weak) signs this is (not) happening

Is this a revolutionary idea? Not at all, this is how psychometrics started, originated from psychology

Here are a few topics for discussion

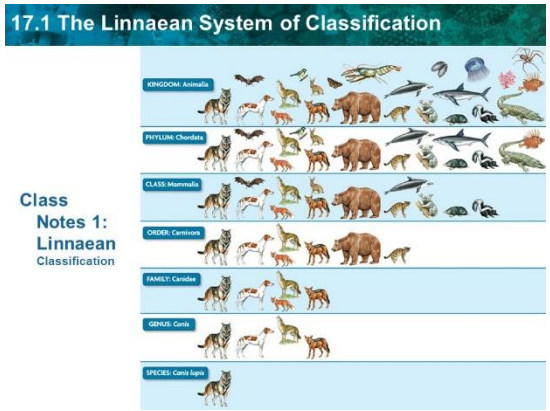
Psychology and other human sciences lack a “language” for understanding the phenomena they study

Biology

- Classification of species—Carolus Linnaeus

  Nomina si nescis, perit et cognitio rerum

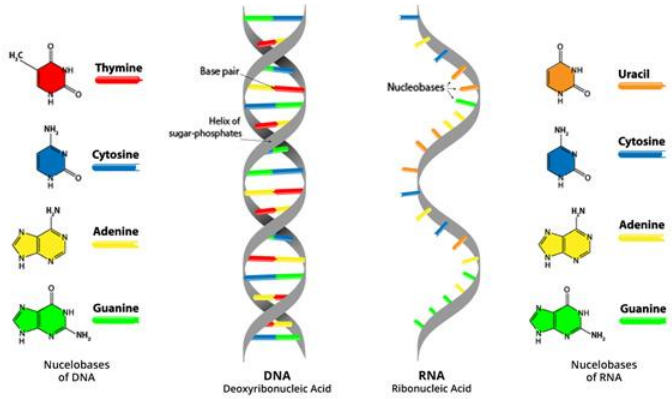
  -- Ignoring the names of things, we also lose the knowledge



- Structure of DNA, Deoxiribose nucleid acid—Watson & Crick

  The main repository

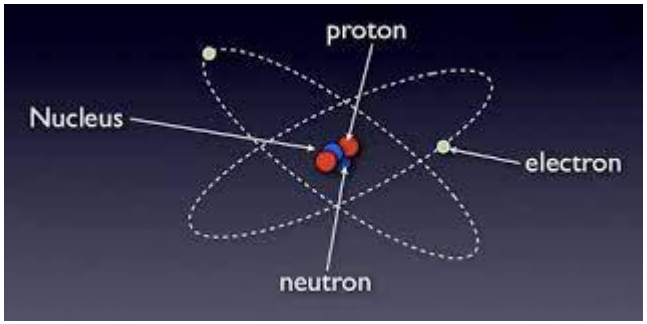
  of genetic information



Chemistry

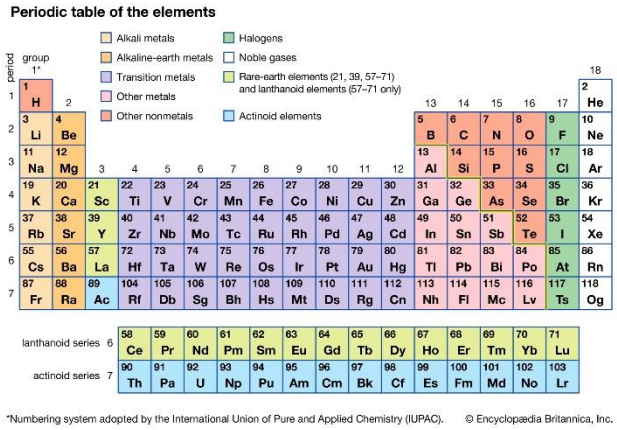
- Model of the atom

  Helps to understand the composition and properties of matter



- Periodic table of chemical elements

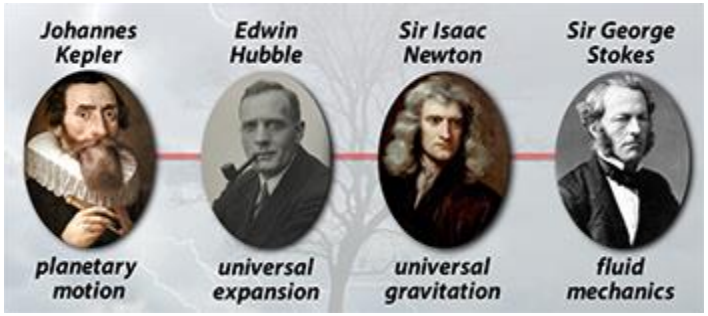
  Helps to understand why elements do or do not combine to form molecules



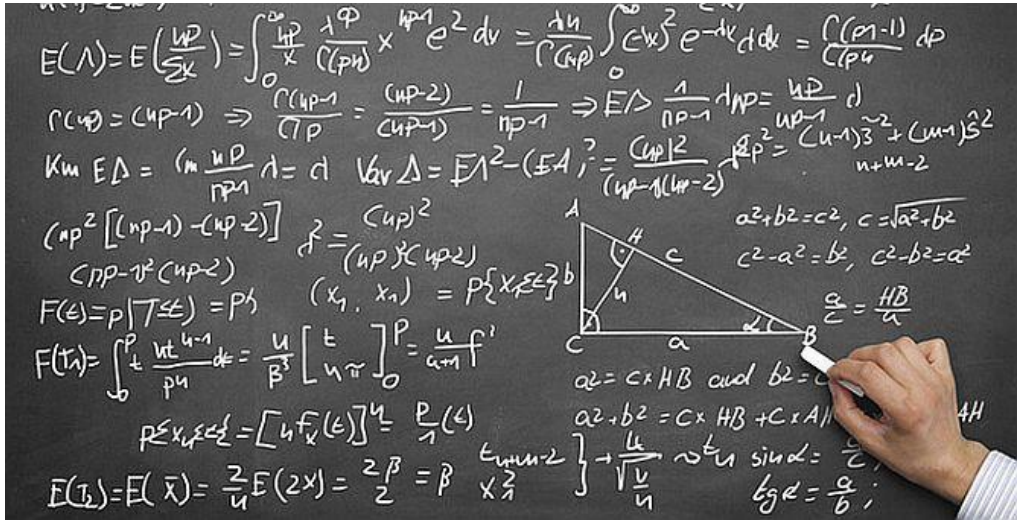
Physics

Mathematics accurately describes physical world, on planet Earth and the Universe

- Natural Laws

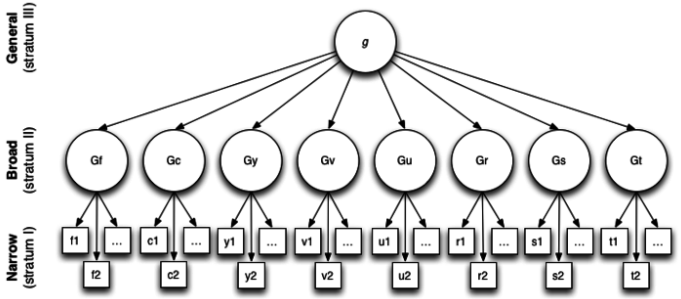


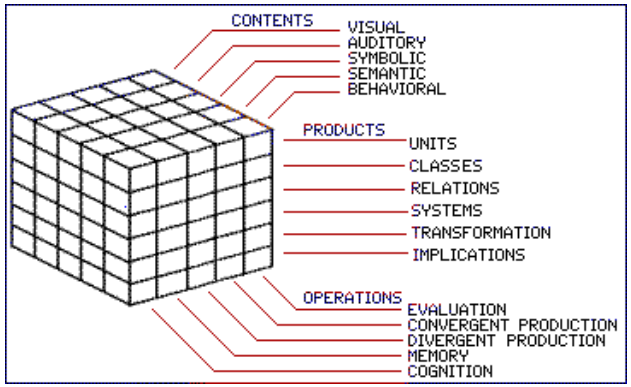
- Predictions from derivations



Psychology

Intelligence, various models, competitive but no crucial experiments





Personality, summary of “wealth” of traits in five overarching categories

Comments and questions

- Reminds me of categorization by Linnaeus; they help, but ……
- So far, no breakthrough opening new insights impossible previously
- Does Psychology invest too little in “slow science” and too much in “headlines”?
- Has neuropsychology helped psychology further along? Does brain imaging push psychology to the next stage?
- Do big data (Internet) and collecting data using smart watches lead to better measurement?

Different topics

- Do measurement models help the researcher thinking about useful measurement?

- Should psychological theories about attributes guide choosing or constructing appropriate measurement models?

- Is psychometrics better off developing back to quantitative psychology or further developing into statistics?


THE
END