How to Measure the "Objectivity" of a Test

Mark H. Moulton, Ph.D.

Director of Research, Evaluation, and Psychometric Development

Educational Data Systems

“Specific Objectivity” (Ben Wright, Georg Rasch)

Rasch found an analogy in physics

		Accelerations			
		1	2	3	4
	$a = f/m$	Mass 1	Mass 2	Mass 3	Mass 4
2	Force 1	2.0	1.0	0.7	0.5
3	Force 2	3.0	1.5	1.0	0.8
4	Force 3	4.0	2.0	1.3	1.0
5	Force 4	5.0	2.5	1.7	1.3
6	Force 5	6.0	3.0	2.0	1.5

In physics, “measurement” is not statistical

- You don’t need a representative sample from a population.
- You don’t need a normal distribution.
- You don’t need a lot of data
- You don’t need to worry about missing cells
- Each “force” is not a mean across a sample of masses
- Each “mass” is not a mean across a sample of forces

You DO Need…

- A method to estimate the hidden forces and masses
- A “reference mass” (e.g., a kilogram weight in Paris)
- Ratios between rows to be the same for each column
- Ratios between columns to be the same for each row
- If the ratios are *not* the same, why?
- If they are the same:
  - Data meets the requirements for objective measurement
  - Derived row/column parameters are “objective measures”
  - They are “objects”
  - A row object has the same meaning for all possible future column objects, and vice versa
  - Not just for this dataset
  - So long as you only accept objects that fit the model

- Dressed up in probabilities, this is the Rasch model

Additional Parameters

- Shortly after Rasch's publication (1960), Birnbaum (1968) proposed two additional item parameters:
  - "discrimination" – a parameter to estimate the discriminating power of each item
  - "guessing" – a parameter for MCQ items to estimate the effect of lucky guesses
- The new model "fit" real-world data better
- BUT it abandoned Rasch's "specific objectivity"

Ben Wright Objected!

The Birnbaum model is designed to imitate data, to be faithful to the data as well as possible, to accept any kind of data, whatever may come up. However it is contrived primarily for MCQ response curves. Quite different from that is the Rasch model which is not designed to fit any data, but instead is derived to define measurement. The Rasch model is a statement, a specification, of the requirements of measurement - the kind of statement that appears in Edward Thorndike's work, in Thurstone's work, in Guttman's (1950) work. Rasch is the one who made the deduction of the necessary mathematical formulation and showed that it was both sufficient and necessary for the construction of linear, objective measurement. It is also nice that there are sufficient statistics for these parameters, because that's a useful and robust device for getting estimates. The Birnbaum model has loose standards for incoming data. It hardly ever objects to anything because it's adjusted to adapt to whatever strangeness there is in the data. The Rasch model has tight standards. The two models are opposites - one loose, the other tight - in the standards they set for the data they will work with. (Wright, 1992)

How One Becomes a Psychometrician

- I was (am) a philosopher. Objectivity connotes:
  - Cognitive independence
  - Truth, reality
  - Many philosophers question whether objectivity is possible, or even desirable
  - *"What is truth?", said jesting Pilate, and would not stay for an answer.* (Francis Bacon)
- Ben's claim of "objective measurement" surprised me -- I was skeptical.
- I *did* stay for the answer.

The Situation

- "Objectivity" as a goal is *local* to the Rasch community
  - Item Response Theory ignores it, or defines it incorrectly.
  - Machine Learning and AI lack the concept except as a notion that "overfit is bad".
- All numerical methods would benefit, since objectivity:
  - improves predictive accuracy
  - improves reproducibility and generalizability
  - favors scientific theory over dumb data fitting
- But how do we compare different models and methods?
  - There is no universal objectivity statistic, not even in Rasch! Just "fit to the Rasch model". I'm about half way to a statistic.

First Things: What is Objectivity?

- To the degree the following three conditions are met, objectivity exists:
  1. The “reality” in question consists of objects.
  2. The “data objects” correspond to “real objects”.
  3. The “model” requires its “real objects” and “data objects” to *behave* like objects as a condition of fit, within its constraints.
- A *high* objectivity statistic (e.g., 1.0) should reflect the presence of *all three* conditions.
- A *low* objectivity statistic (e.g., 0.0) should reflect that one or more of the conditions has not been met.

Reality Objects

Data Objects

Model Objects

Report Objects

	1	2	3	4	5
A	1	2	2	0	2
B	2	3	2	5	0
C	4	3	0	0	5
D	1	0	5	4	0
E	0	0	4	4	2
F	0	4	4	3	2
G	4	0	1	0	3
H	4	3	4	0	2
I	5	1	0	3	3
J	5	1	3	1	1
K	2	1	2	1	1
L	5	1	0	3	0
M	4	0	3	2	3
N	1	0	1	5	1
O	0	3	0	1	1
P	0	0	1	5	2
Q	1	3	1	5	5
R	0	2	0	1	1
S	3	4	2	0	1
T	1	5	2	5	1
U	5	1	2	2	2
V	5	4	1	5	5
W	0	0	2	3	0
X	4	1	3	2	1
Y	2	3	4	2	5
Z	2	1	1	0	3

β_n

δ_i

C_j

F_k

- Abilities
- Difficulties
- Probabilities
- Measures
- Predictions

What is an “object”?

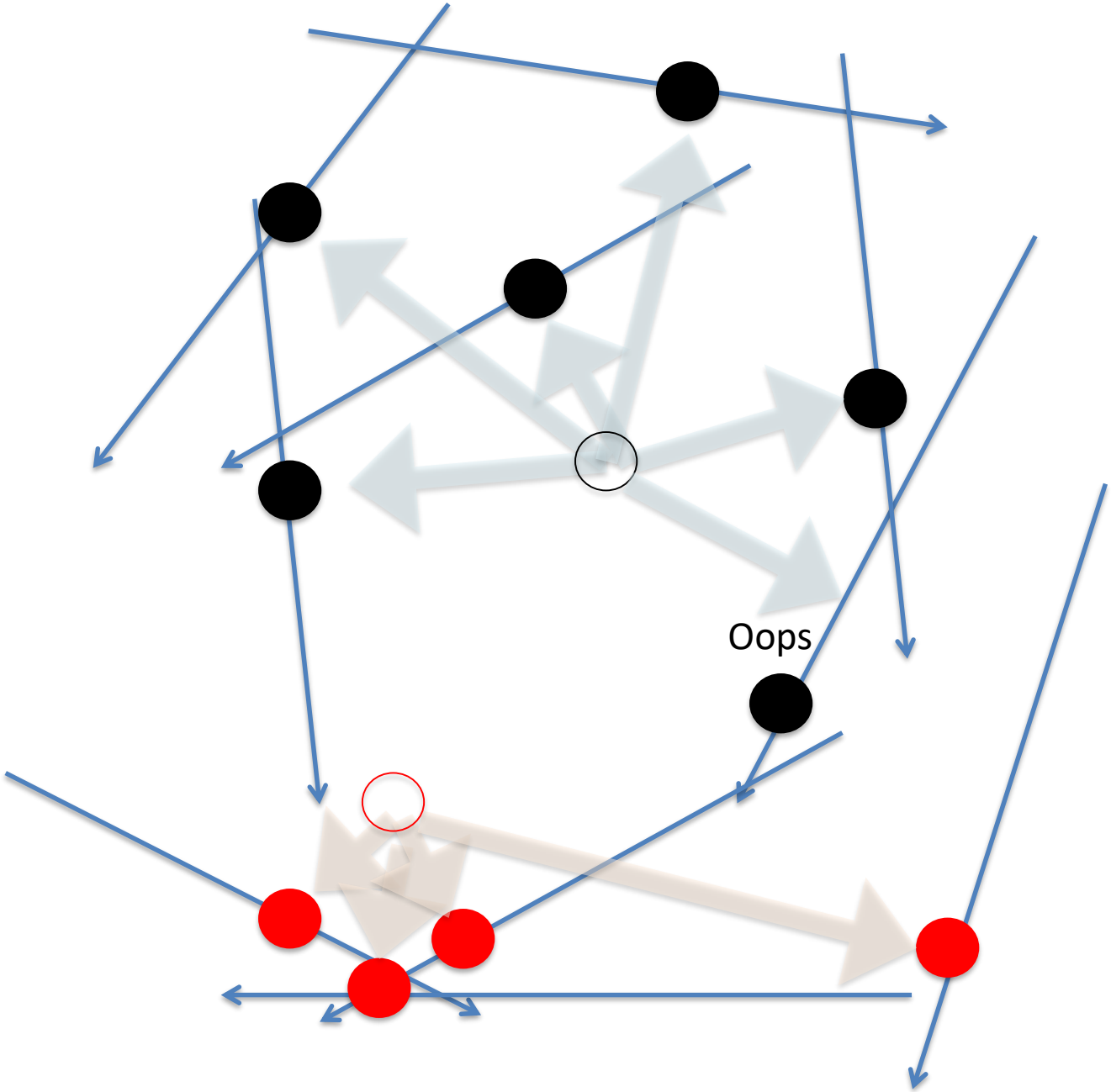
- From *ob-* + *jacere* (to throw): *to throw before or against, to put in front of something*
- Objects “present themselves”, manifest distinctness, yet are invisible (latent) in themselves.
- Some philosophical properties:
  1. lead to “accurate” expectations and predictions
  2. have unitary or “stable” (invariant) meanings across situations
  3. are “independent” of other objects
  4. are in some sense “true”
- These philosophical properties suggest mathematical properties that may be used to derive a practical statistic.

Mathematical Properties of Objects

- Accuracy
  - Ability to predict values of missing cells (basis of the 2009 Netflix Prize)
- Stability
  - An object's parameters calculated from one section of a dataset should equal those calculated from any other section.
  - This is equivalent to saying objects should be *unitary*. They should occupy one, and only one, position in *n*-dimensional space.
- Independence (local item independence)
  - The parameters associated with one object should have no forced relationship with any other.
- A small standard error (proximity to "truth")
  - Differences between cell estimates calculated by a model and the "true" values for those cells (not observed) should approach zero.

My Summary Definition
(for those late nights)

*An object of analysis is **objective** as such to the degree that it can be associated with multiple independent interactions **N** > **D** with objects **I** of a given type where the interactions manifest as values that, taken together, imply one and only one position for that object in the **D**-dimensional space in which the interactions take place.*


Oops

To calculate objectivity, we need...

- A labeled array of observed values **X**
  - Assumed to equal a hypothetical "true" array **T** plus random noise **S**: **X** = **T** + **S**
  - The array can have any number of facets, but let's assume two (row objects × column objects).
- A methodology *M* that models **X** such that:
  - *M*(**X**) yields parameters for each row and column object. (Call the parameter arrays **R** and **C**.)
  - *M*(**R**, **C**) yields an estimates matrix **E** that is the same shape as **X**: **E** = *M*(**R**, **C**).
- *M* must support the following:
  - **X** can contain missing values.
  - *M* fills in the missing cells with values from **E.**
  - Parameters in **R** can be calculated from **C** and **X**; parameters in **C** can be calculated from **R** and **X** (aka, "anchoring").

Eligible Methodologies *M*

- All Rasch models
- All IRT models (2PL, 3PL, GPC)
- NOUS (the model I use)
- Some matrix decompositions (not SVD)
- Multiple Regression (limited)
- Factor analysis (probably not)
- Bayesian models (I don’t know)
- Neural networks (I think so)

Prediction Accuracy

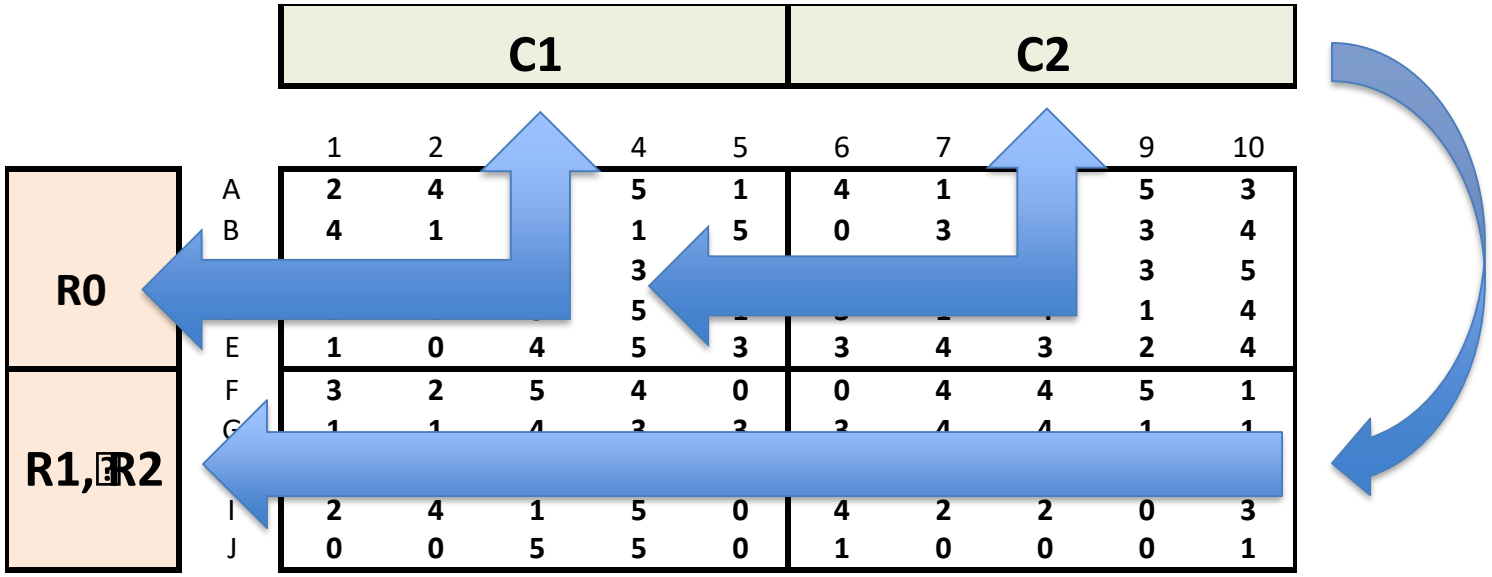
1. In observation array **X**, make a set of cells "pseudo-missing". (The remaining cells are called the "training dataset".)
2. Apply methodology *M* to **X**[*non-missing*] to obtain **E**.
3. Correlate the observations and estimates for the pseudo-missing cells:

$$Accuracy = correl(\mathbf{X}[missing], \mathbf{E}[missing])$$

Stability

1. Shuffle the rows and columns of **X**, either randomly or according to something like ability/difficulty then divide into four quadrants.
2. Use the top half to compute column (item) parameters **C**. Split **C** into **C1** and **C2**.
3. Apply **C1** to the lower-left quadrant of **X** to estimate row (person) parameters **R1**.
4. Apply **C2** to the lower-right quadrant of **X** to estimate parameters **R2**.
5. Correlate **R1** and **R2** (which should be equal if objective):

Stability= *correl*(**R1**, **R2**)


C1
C2
R0
R1,
R2
1 2 4 5 6 7 9 10
A 2 4 5 1 4 1 5 3
B 4 1 1 5 0 3 3 4
3 3 5
5 1 4
E 1 0 4 5 3 3 4 3 2 4
F 3 2 5 4 0 0 4 4 5 1
I 2 4 1 5 0 4 2 2 0 3
J 0 0 5 5 0 1 0 0 0 1

Tricky for 2-PL, 3-PL

- It's not quite straightforward to calculate person parameters given a set of item parameters
- Need to apply a special procedure to get **R1** and **R2**.
- Procedure:
  - Build two "conversion tables" for **R1** and **R2**
  - These relate all possible person parameters (thetas) to their corresponding raw scores
  - Calculate raw scores for **R1** and **R2**
  - Look up the relevant thetas and correlate them

Independence

1. From **X**, calculate **E** = *M*(**X**)
2. Calculate residuals array **Res** = **X** – **E**
3. Calculate squared correlations between each pair of columns of **Res**. Take their mean and subtract from one.

   Independence = 1 – *mean*(*correl*(**Res**[*I*], **Res**[*J*])^2)

4. If the column objects are fully independent:

   mean(correlations) ≈ 0; *Independence* ≈ 1

The Objectivity Statistic

- Currently used as an array-level statistic, but in principle can be calculated for individual objects.
- Geometric mean of the three statistics, with $0 \leq \textit{Objectivity} \leq 1$:

$$\textit{Objectivity} = (\textit{Accuracy} \times \textit{Stability} \times \textit{Independence})^{\frac{1}{3}}$$

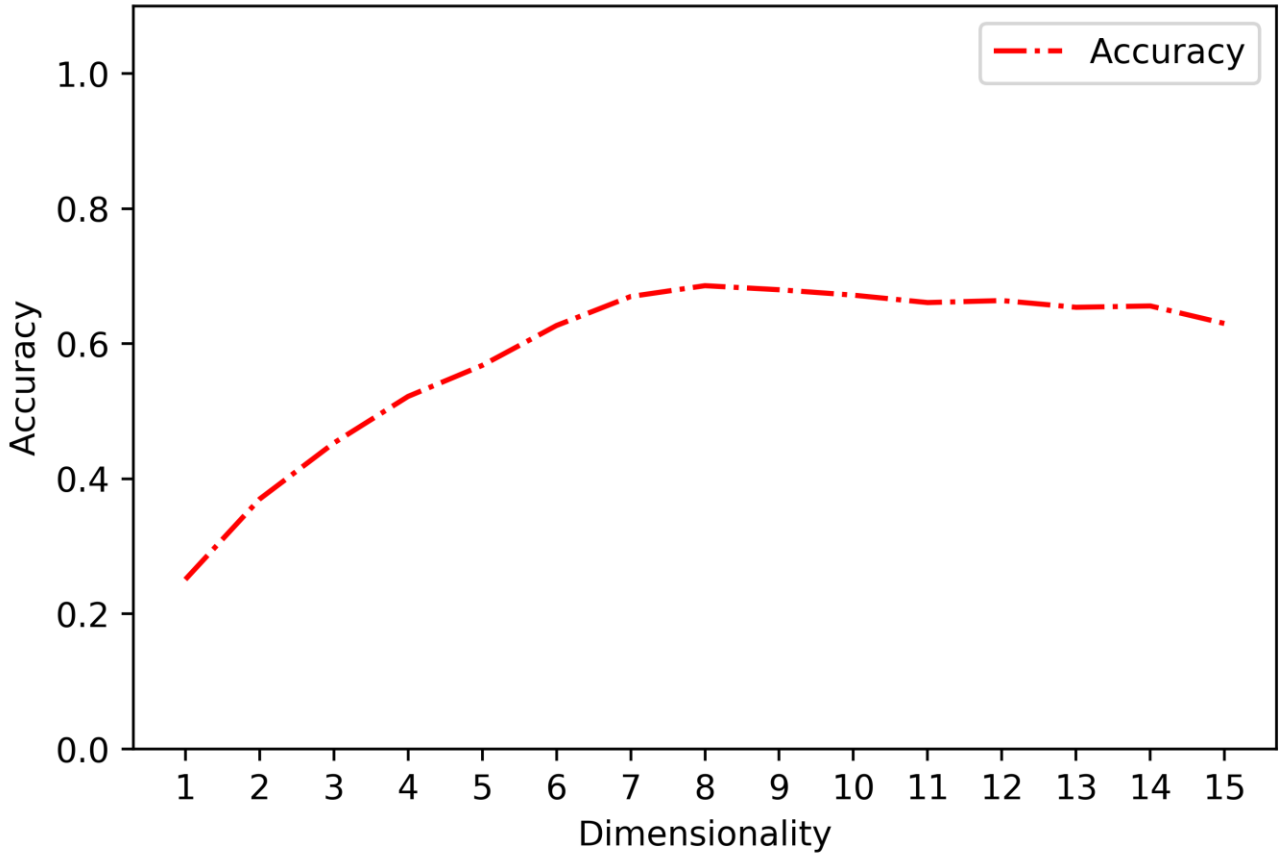
An Example (NOUS)

- My colleague Howard Silsdorf and I developed NOUS in 2003.
- NOUS is an alternating least squares (ALS) matrix decomposition adapted for psychometrics, with a Rasch-like objectivity requirement:

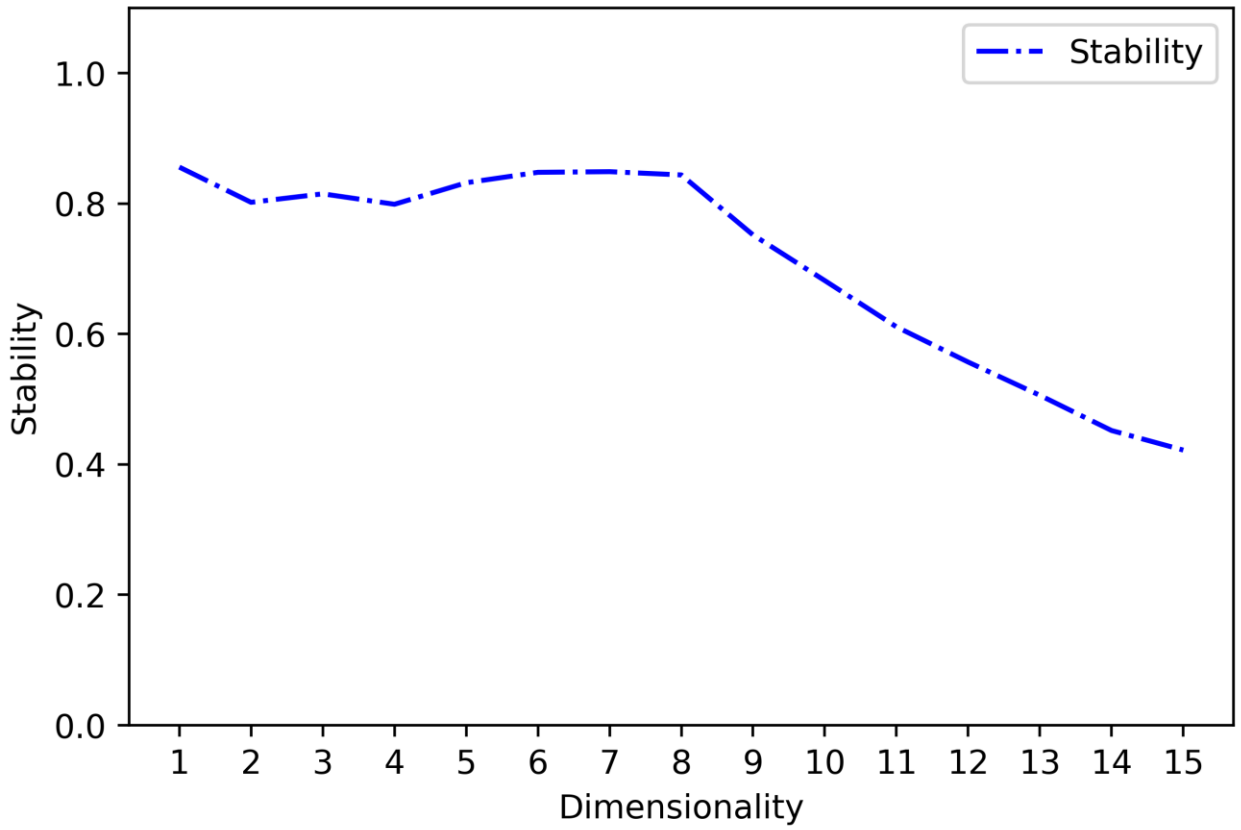
  E = **RC**

  (Cell Estimates = Row Matrix (abilities) * Column Matrix (easinesses)
- Handles highly multidimensional data ("within-item")
- Each dimensionality defines a separate NOUS model to be evaluated according to its objectivity statistic.
- I ask: which model (i.e., which dimensionality) yields the highest objectivity statistic?
- NOUS currently uses *Objectivity* = (*Accuracy* × *Stability*)^{½} .
- Simulation: 500 rows, 100 cols, 8 dimensions, 4-category responses, typical noise

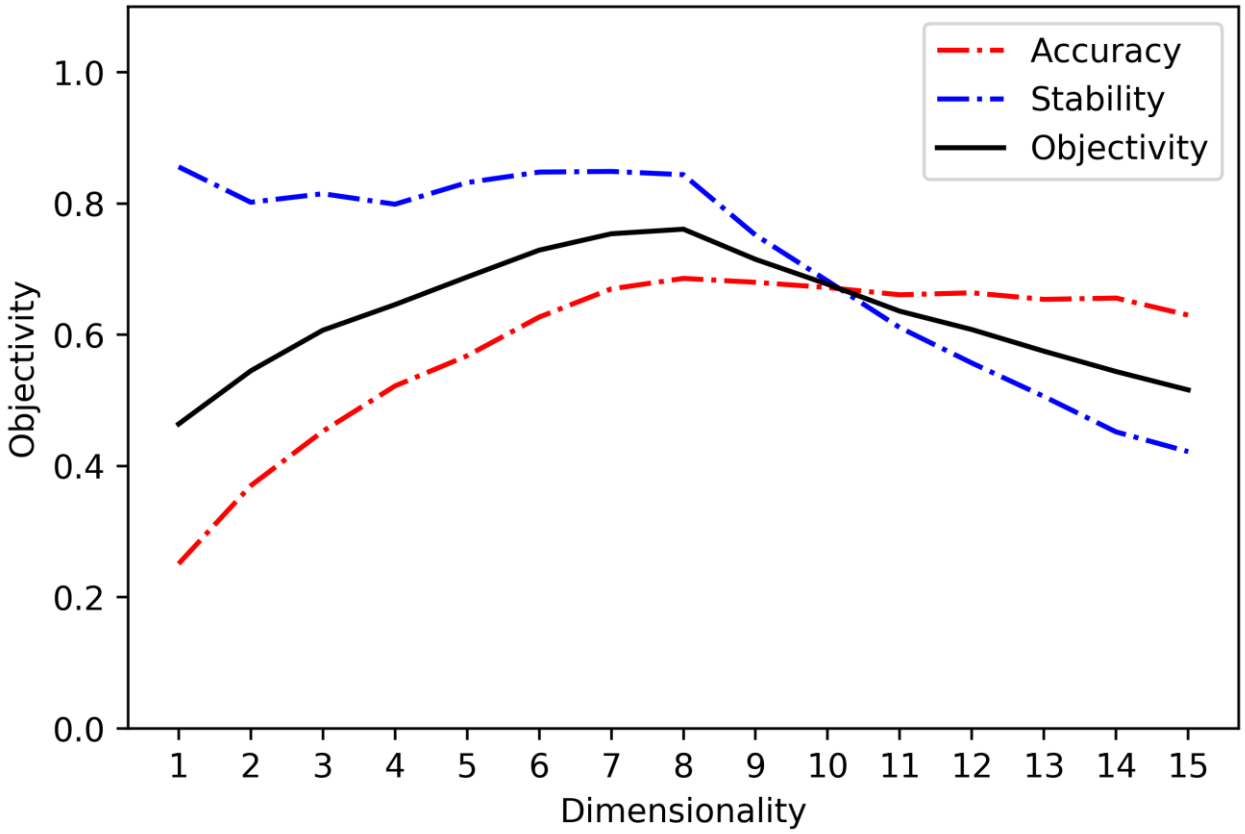
Accuracy (predict missing cells)



Stability (parameter reproducibility)



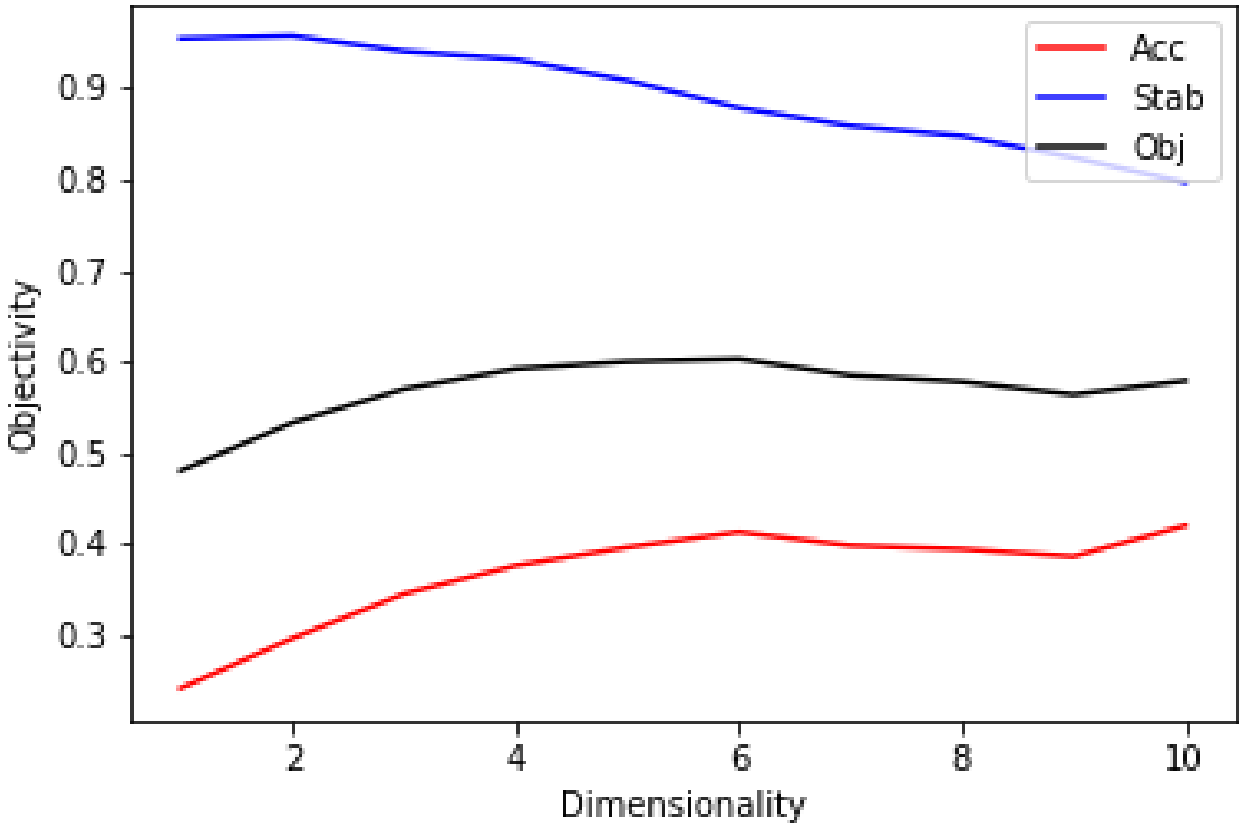
Objectivity



Complementarity

- Accuracy is most sensitive to *underfit.*
- Stability is most sensitive to *overfit.*
- Combining them yields a clear peak, identifying the optimal model.

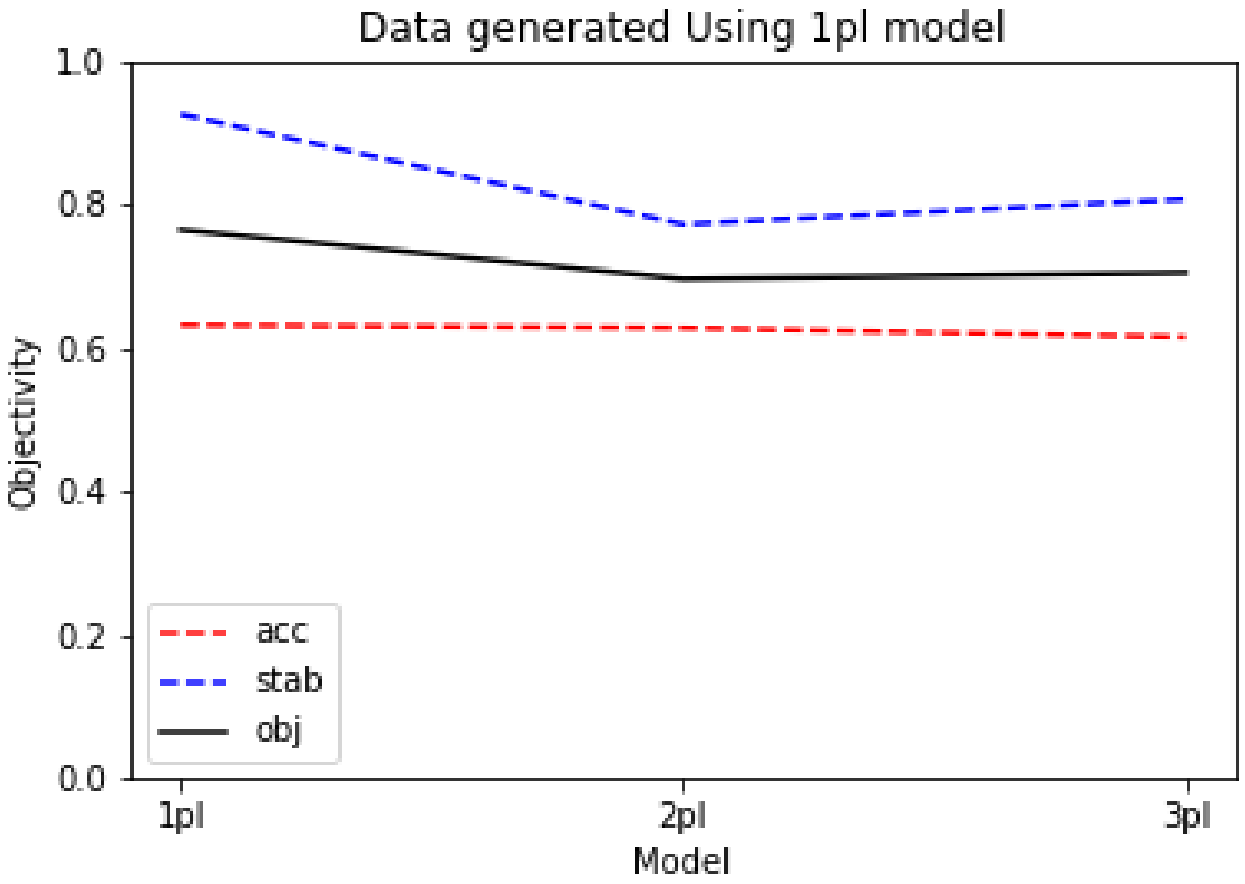
A Recent Dyslexia Analysis



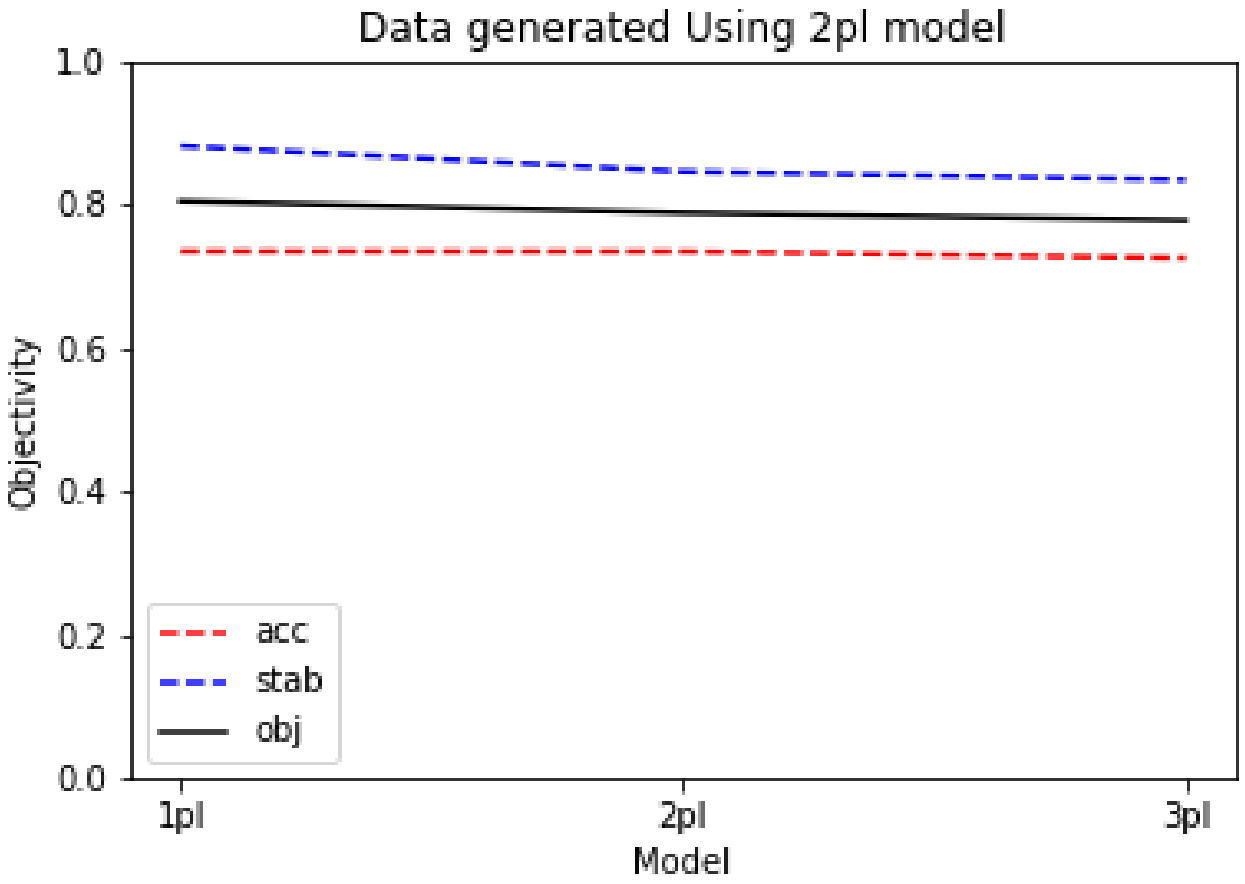
It’s not just about dimensionality

- We just evaluated 15 models that differ in dimensionality on NOUS, nothing else.
- But objectivity can *also* be used to evaluate:
  - Best IRT model for a given dataset, for example:
    - Rasch (1PL)
    - 2PL
    - 3PL
  - Other statistical methodologies
  - Datasets and subsets of datasets

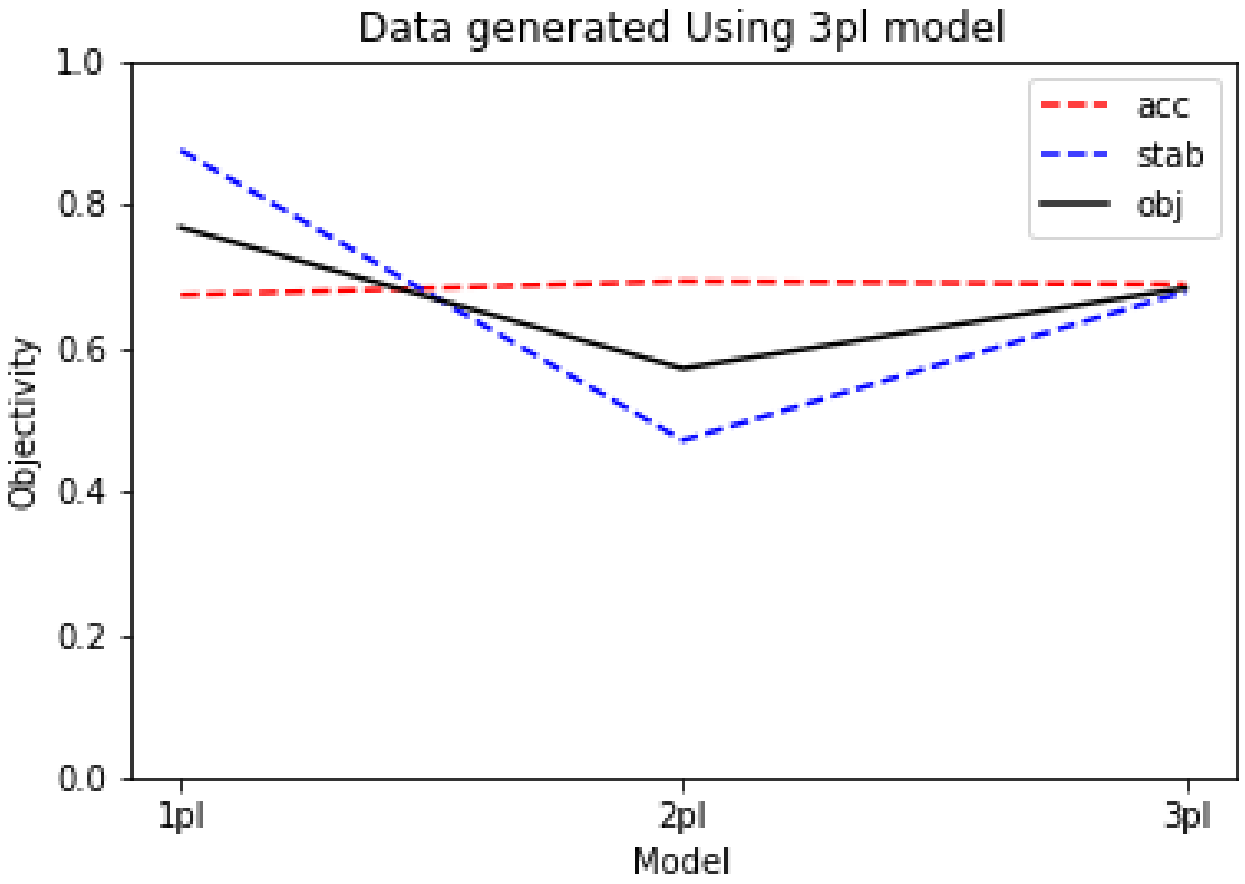
Compare Using 1PL Data



Compare Using 2PL Data



Compare Using 3PL Data



Upshot

- The Rasch (1PL) model was as objective, or more objective, than the 2PL and 3PL models for these simulations
- Even when I generated data to conform to the 2PL and 3PL models.
- The improvement in fit (not shown) between cell observations and expected values for the 2PL and 3PL did not mean the resulting parameters were any more objective.

Deducing Objectivity by Layer

- Say we have a low objectivity stat. What does it mean?
- We diagnose by process of elimination:
  - Is the model valid for the dataset in question? Do the objects meet its requirements? Do we need more parameters?
  - If the model is okay, is there something wrong with the data? Was it incorrectly recorded? Did some externality interfere?
  - If the data is okay, it is possible our proposed objects are not really objects at all, i.e., not unitary.
  - A person may not be a person.

Software

- My software, Damon, automatically calculates objectivity to choose best dimensionality.
  - But it is not a canonical IRT model
  - Though similar algorithms are used a lot in machine learning
- For the 1PL, 2PL, and 3PL
  - I wrote a script in R based on the “ltm” R package
  - But it’s rough. Contact me if interested.
- For other IRT models
  - I hope to write an R script based on the “TAM” package, which handles a wealth of IRT models.

Not the Future

- This is NOT the *Objectivity* formula of the future:
  - Too awkward and slow
  - Not standardized (choice of missing cells, choice of data partitions)
- But it works
  - Identifies best dimensionality
  - Nicely lines up with intuitive ideas about objectivity
- Prediction: The "final" *Objectivity* formula will be based on standard error, like *Reliability*.
  - The trick is deriving a universal standard error formula that works across all models and datasets.

Concluding Thought: Objects in Reality

- Can objects even exist in “the real world”?
  - On the one hand “objects” are purely mental or mathematical abstractions, not physical.
  - Yet we are forced to assume that “objective reality” is comprised of objects, else we could not do measurement or science.
  - So is objective reality in some sense mental?
- This is an ancient and perplexing question:
  - Intelligibility of the world (Plato, Aristotle, Kant, Einstein)
  - Principle of Sufficient Reason (Leibniz)
  - Monads (Pythagoras, Lucretius, Giordano Bruno, Leibniz)
  - Science (assumes intelligibility and gets lucky)
- Working hypothesis:
  - The world is a mix of chaos and objects.
  - Chaos forms into objects, objects disintegrate into chaos
  - But science and measurement can only look at the world, or small sections of it, to the degree it consists of objects. It is blind to the rest.
  - Even so, why are there objects at all?

Thank you!

Mark H. Moulton

Educational Data Systems

markhmoulton@gmail.com