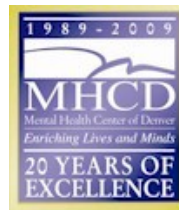


How analogies can save the day when it comes to explaining difficult statistical concepts to stakeholders


Or: The strange case of the friendly analogy

***Antonio Olmos, Kate DeRoche
and CJ McKinney***

Mental Health Center of Denver



Presented at the American Evaluation Association Conference
November 13, 2009

- 
-
- A very important part of the success of any evaluation depends on the continuous communication and feedback between the evaluation team and the program stakeholders
 - This communication depends on how well the two parties understand each other

The challenge

- How do you communicate some important parts of the evaluation when your stakeholders do not have the background?
 - The dialogue can quickly become at best, a monosyllable exchange (i.e., yes, no, grunts)
 - Worst case: a monologue where the evaluator “as the expert in the field,” is dominating the conversation.

Potential options

- Reduce the complexity of the analysis/instruments/outcomes so they're easy to understand during the design stage, and easy to analyze and interpret later.
 - Sometimes what may be the easier solution is not necessarily the best solution.
 - This simplification of the analysis/instruments/outcomes may make the evaluation meaningless as far as generalizations, extrapolations, etc.

MHCD's option

- We describe to our stakeholders the key concepts they need, to understand our end of the evaluation process (statistical analysis or instrument design), in a way that they can relate to
 - Able to follow along, making not only meaningful suggestions, but also important decisions about items/data
 - Understand the extents and limitations of the tools in a way that then they can explain to other stakeholders or in other forums
 - Creates a deeper appreciation for the products and stronger ownership of the evaluation
 - Produces champions that can praise the virtues - and limitations- of evaluation



How do we do it?

Some examples

Our approach to data analysis

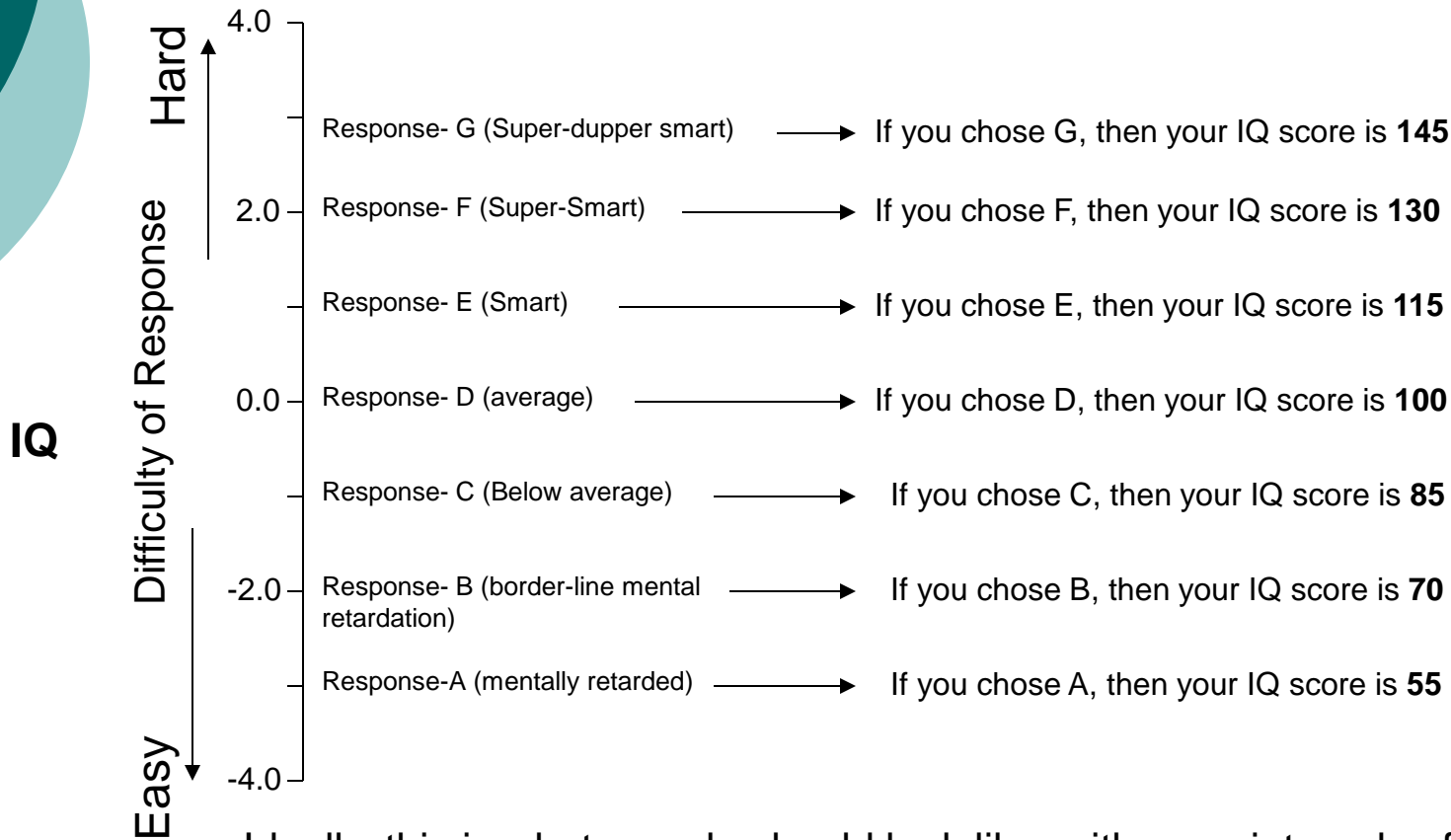
- When doing data analysis 2 different approaches:
 - Hypothesis guided (Columbo): you have an hypothesis and you use data to try to prove or disprove it
 - Data guided (NUMB3RS): You use all sort of analysis to help you understand what the data is saying
- We will be presenting a little bit of both



What is IRT anyway?

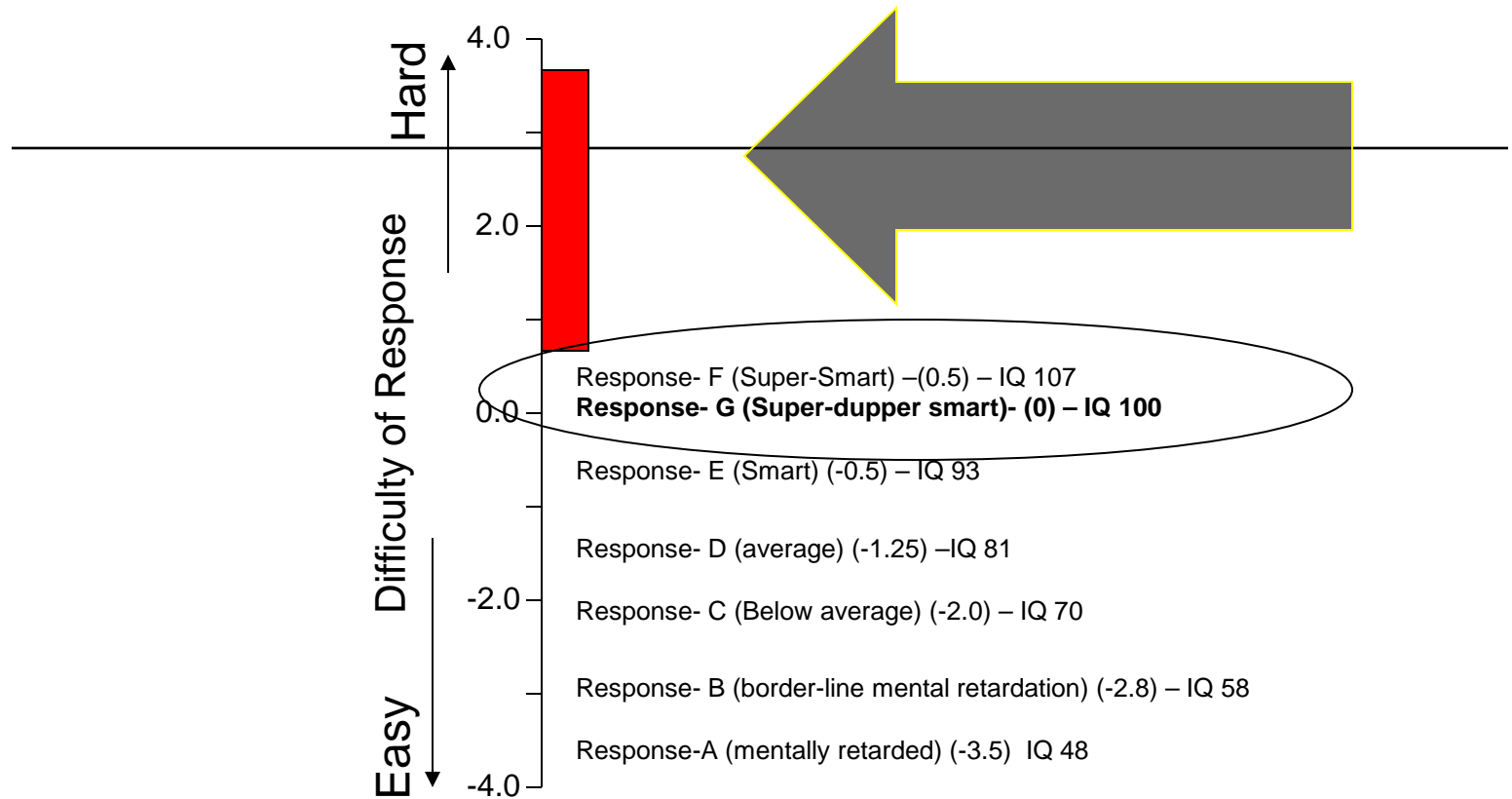
Scaling Example with IQ

We are assuming that **A** is a very easy question (i.e. what is your name?) and **G** is a very hard question (what does floccinaucinihillipilification mean?) It is assumed that if you get a higher response correct, then you got all of the responses below correct, even though you may or may not have been tested for the items below (i.e. if you get **D** correct, then you got **C**, **B**, & **A** correct). The higher the response you get correct, the higher score you receive.



Ideally, this is what a scale should look like, with even intervals of potential responses across the scale. We also want to see the scale range from +3 to -3, so that we can measure IQ scores from 145 to 55.

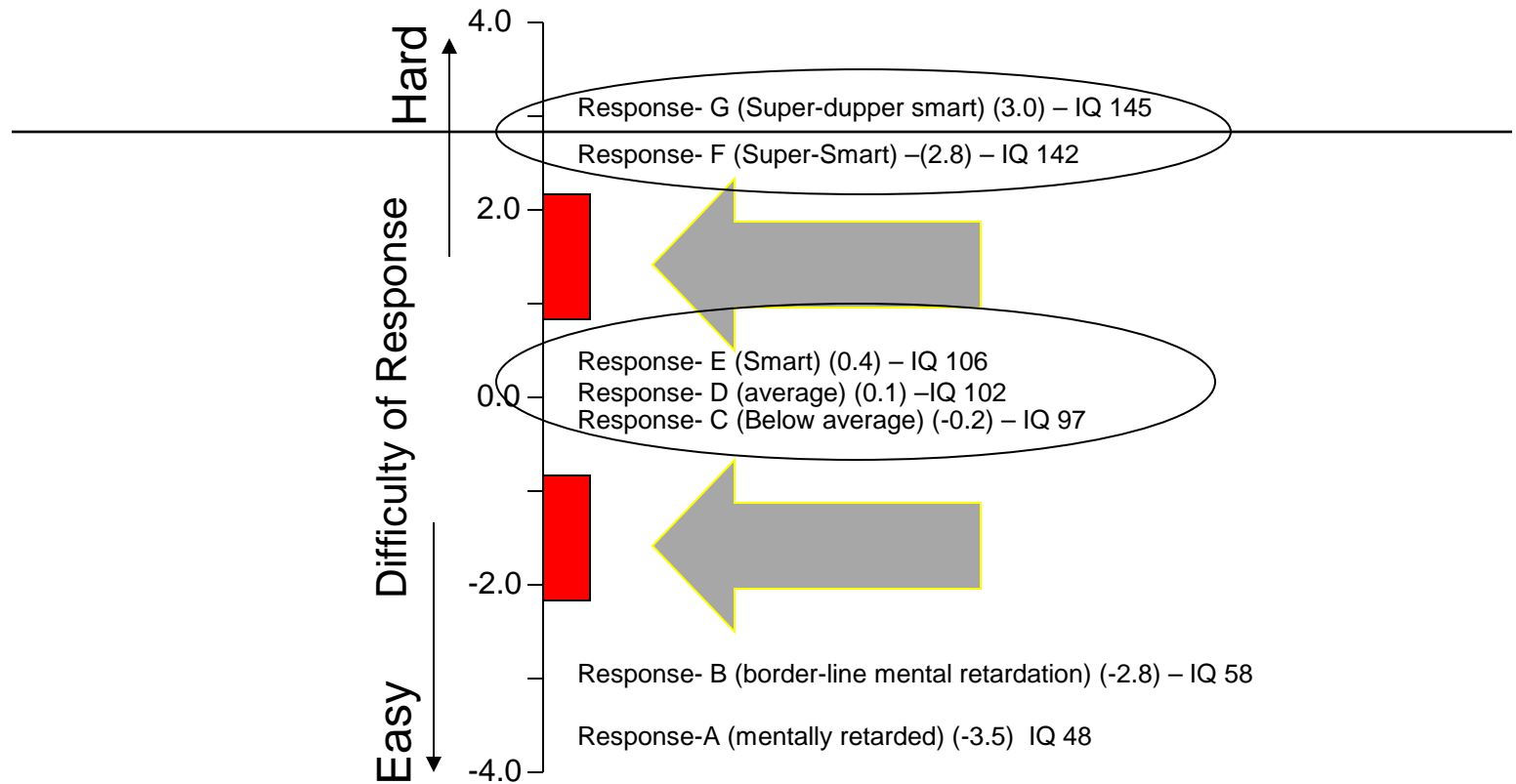
Examples of Problems with Scale (IQ)



Problem 1: The scale does not contain any responses above $(+0.5)$ suggesting that the highest IQ we can measure is 107. Therefore, we will not be able to know how much smarter someone with an IQ score higher than 107 (i.e. 130) might be

Problem 2: Notice that the order of easy to hard goes A,B,C, D, E, G, then F, suggesting that F and G are out of order. Therefore, a response of what we think is hard (only for supper-dupper smart people) is really not that hard and will only produce an IQ score of 100, not 145 as assumed.

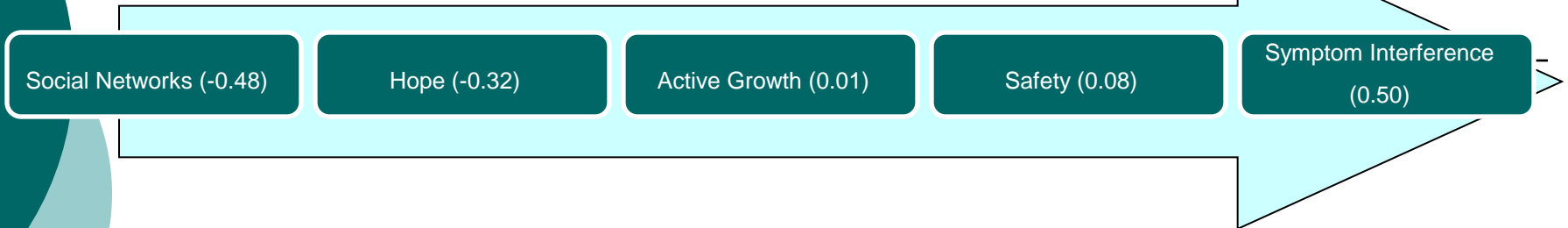
Examples of Problems with Scale (IQ)



Problem 3: Again we see gaps, but not they are within the scale (not just at the top or bottom). This suggest that there are no items able to measure an IQ score between 141 to 107, and between 96 and 56; therefore people are not able to receive score in this area. **WE CANNOT HAVE LARGE GAPS**

Problem 4: We see that items are clumping together (i.e. E, D & C). This means that responses E, D, & C are basically measuring the same thing, and only discriminate between 106 to 102 and 102 to 97. All 3 items are measuring the same thing and are not necessary. We could remove D. **WE DO NOT WANT CLUMPS OF RESPONSES**

Order of Difficulty on the CRM V3.0



- The easiest recovery domain is an increase in **social networks** and **hope**
- As the domains increase in difficulty the number of consumers that get a high score in this domain decreases,
 - For example, if a consumer has a high score in safety they will also have a high score in active growth, hope and social networks because these markers are easier to endorse for our consumers
- The hardest recovery domain to achieve is **symptom interference**. That is, most of our consumers who score high on symptom interference will also score high on all other recovery domains

Now, we can ask interesting questions...

- The knowledge we've gained through our instruments is showing its worth
 - Show the impact of our programs
 - Development of studies to explore some ideas emerging from the data (importance of hope)
 - Comparison of different approaches to therapy
 - Cost-effectiveness studies
 - Changes in clinical practices



Regression Discontinuity

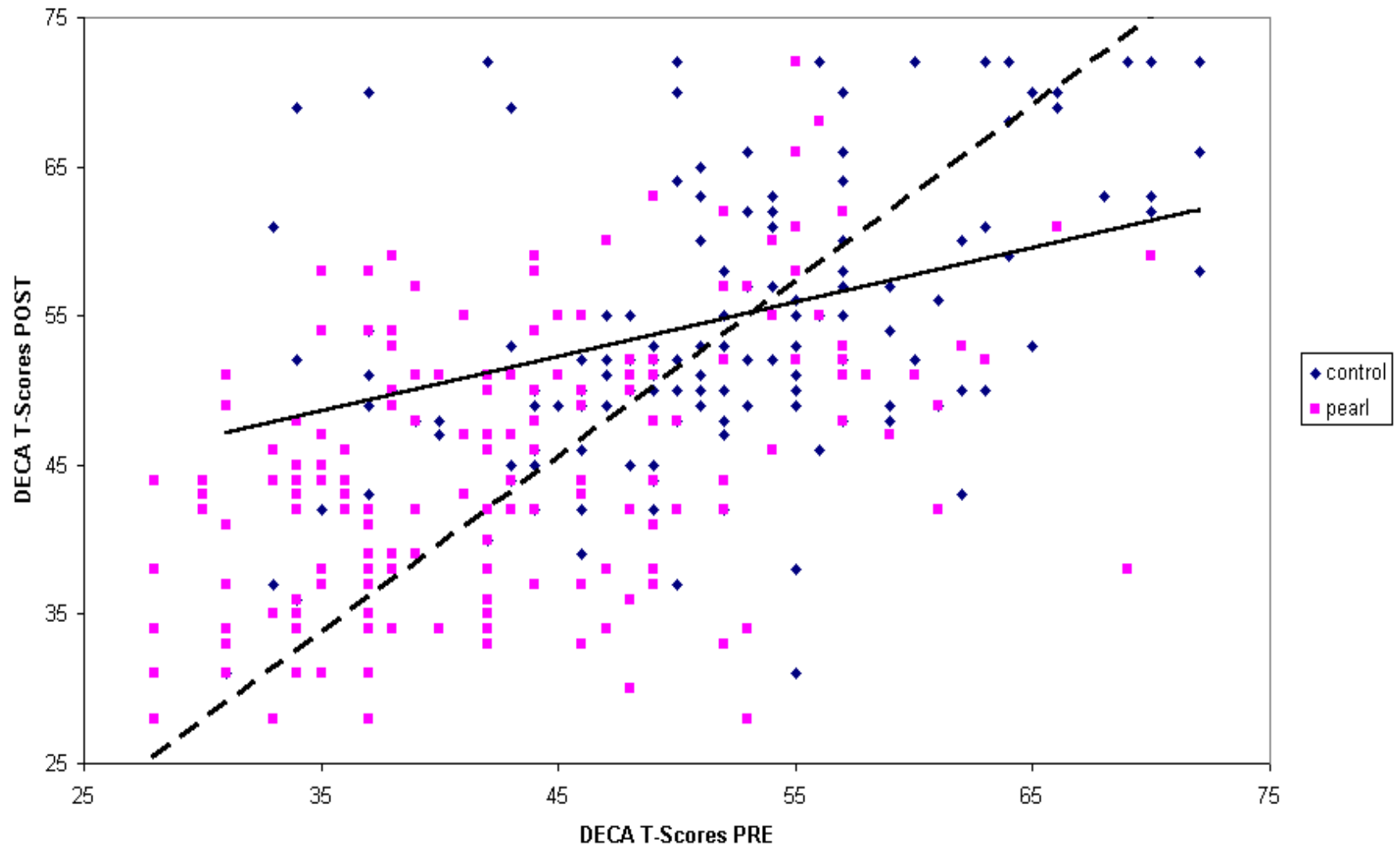
DECA-- Regression discontinuity design

- Research designs where clients can be randomly assigned to control and treatment are considered **gold-standard** in research (help eliminate threats to validity)
- However, from an ethical point of view, this approach prevents clients in need from receiving a treatment that could be even life-saving.

DECA-- Regression discontinuity design (cont)

- In a basic RD design, children are measured in some criteria that will allow an evaluation of their severity. Those with a low severity score -at baseline- are assigned to a control group; those with a high severity score are assigned to the treatment group.
- Thus, using DECA scores, we can assign children to a “**control group**” and compare them to a “**treatment group**”

Regression Discontinuity Design Total Protective Factors





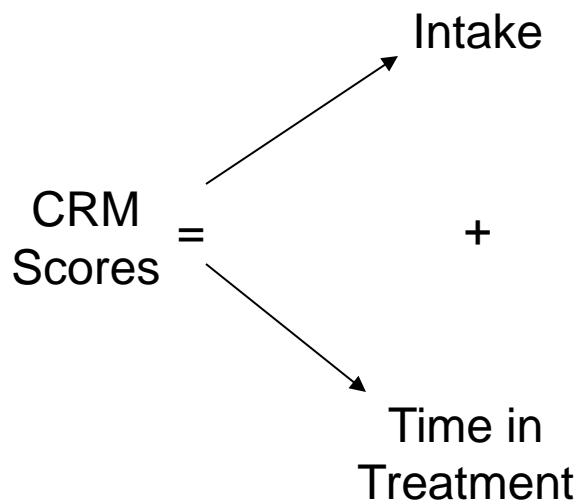
Multi-level models

a.k.a Hierarchical Linear models

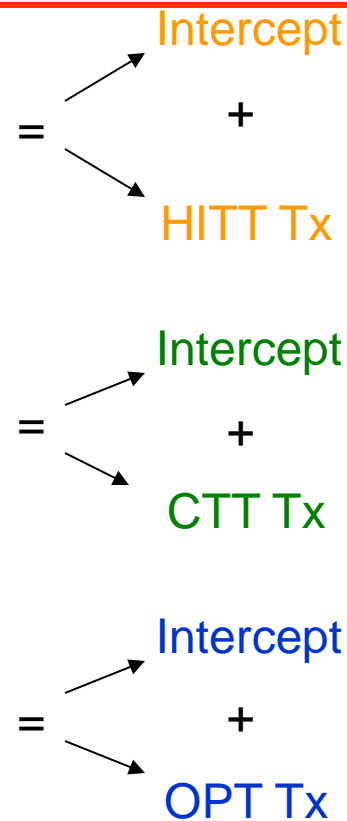
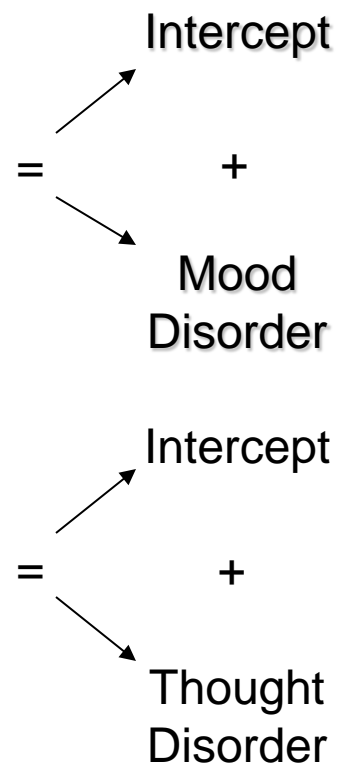
Example of Multilevel modeling concepts

Consumer Level Effect

Typical SLR Model

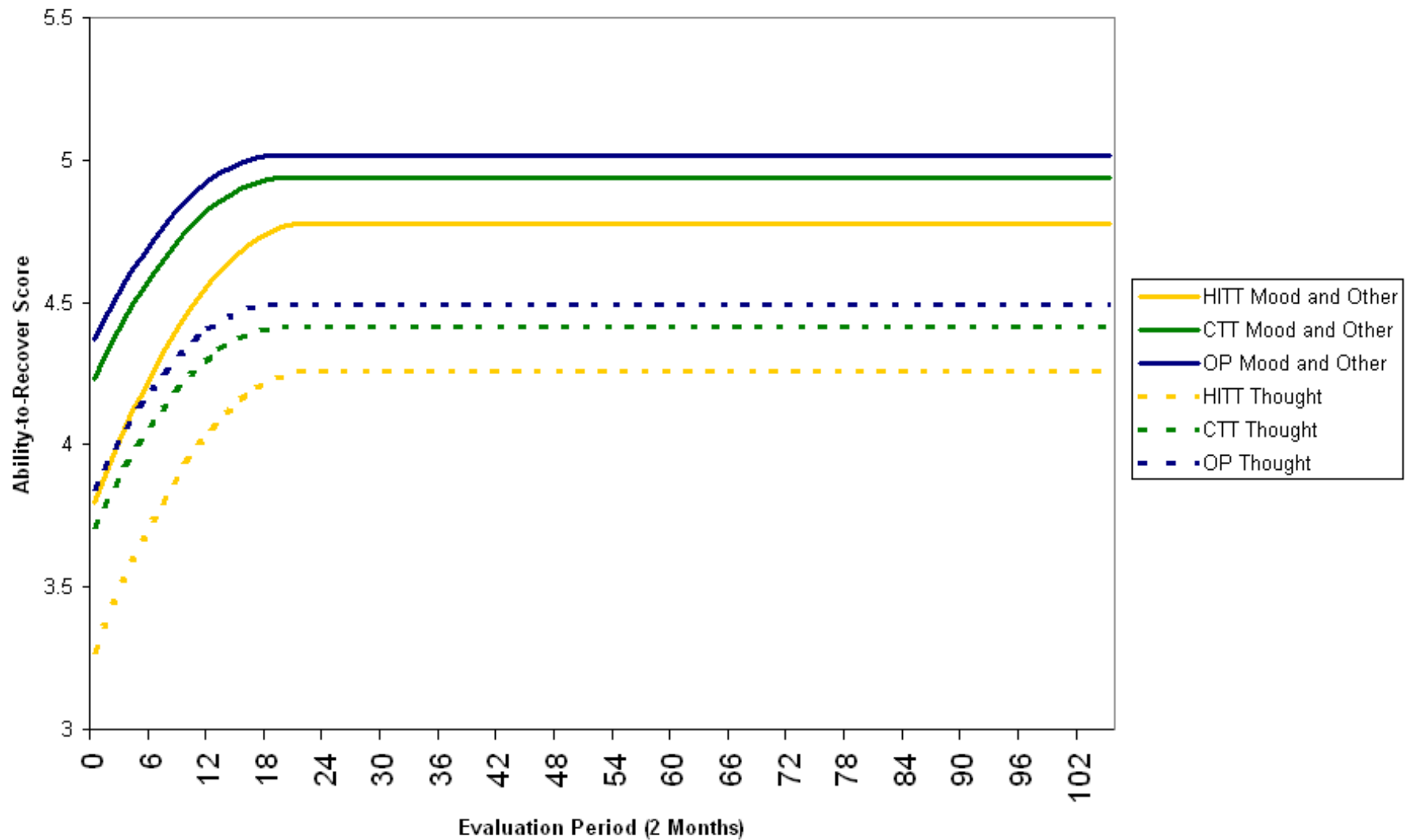


System Level Effect



Higher Level
Effects

Estimated Changes in Recovery Marker Scores Over Time





Quality Control charts

Using information coming from
Multi-level models

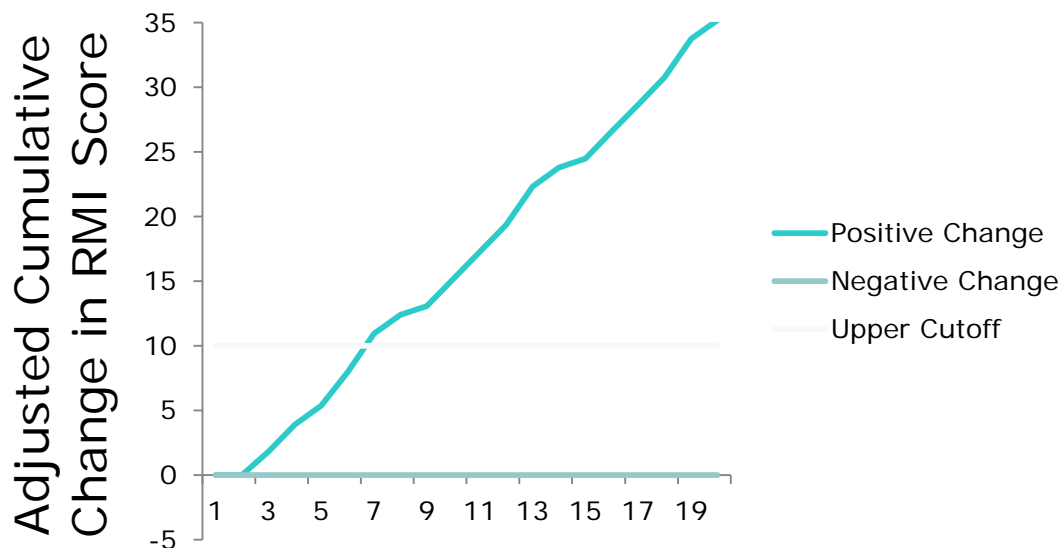
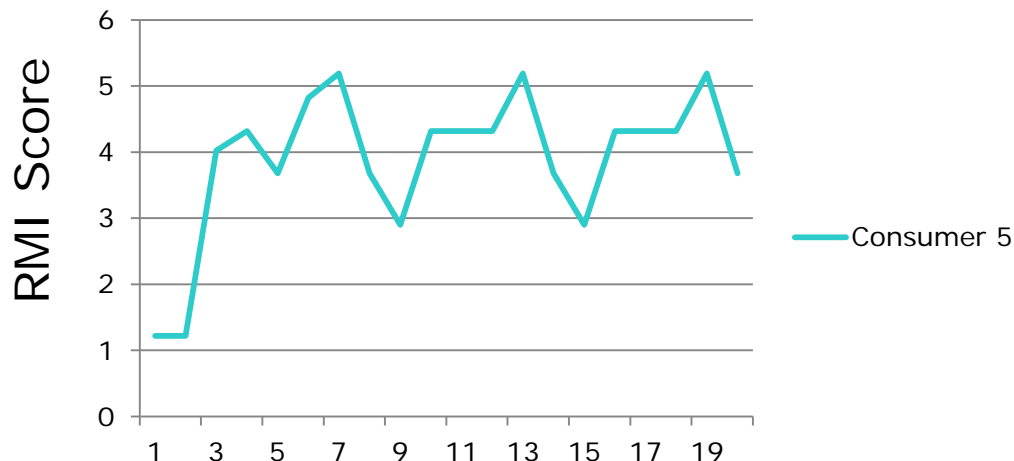
This consumer shows an example of how a true change does occur but may be misinterpreted as being stable.

Though change may not be readily seen,

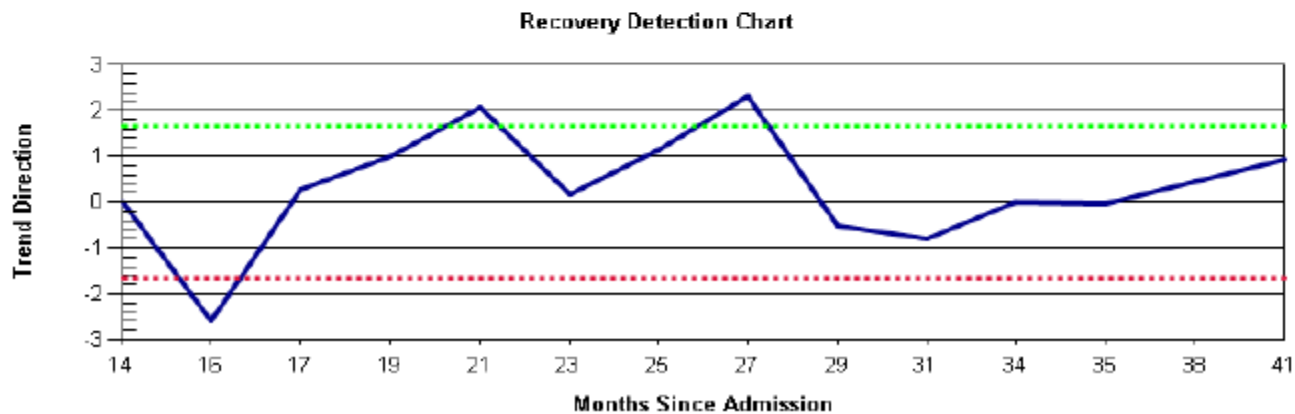
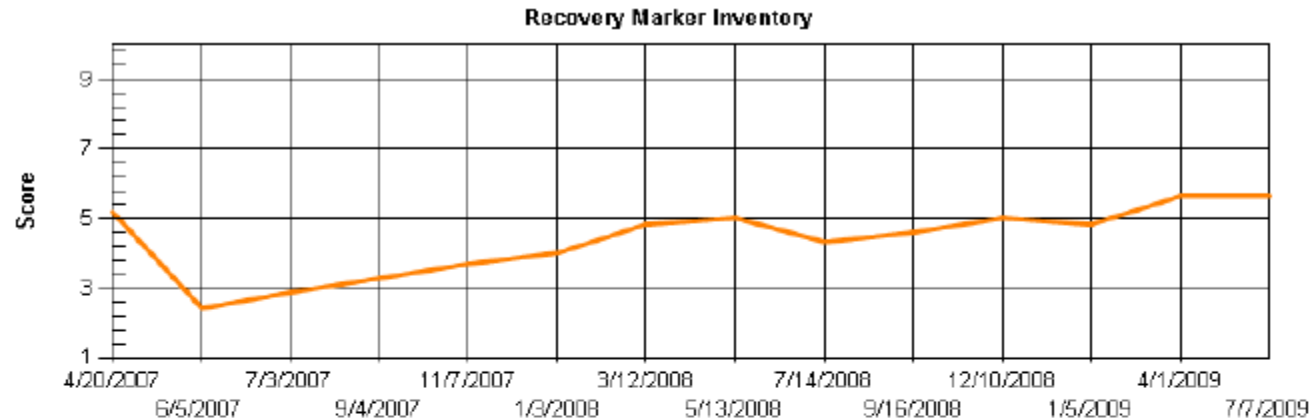
the Adjusted Cumulative Change chart shows that a true shift in the RMI has occurred.

Hence therapy focusing on further recovery supports may be indicated.

Consumer 5

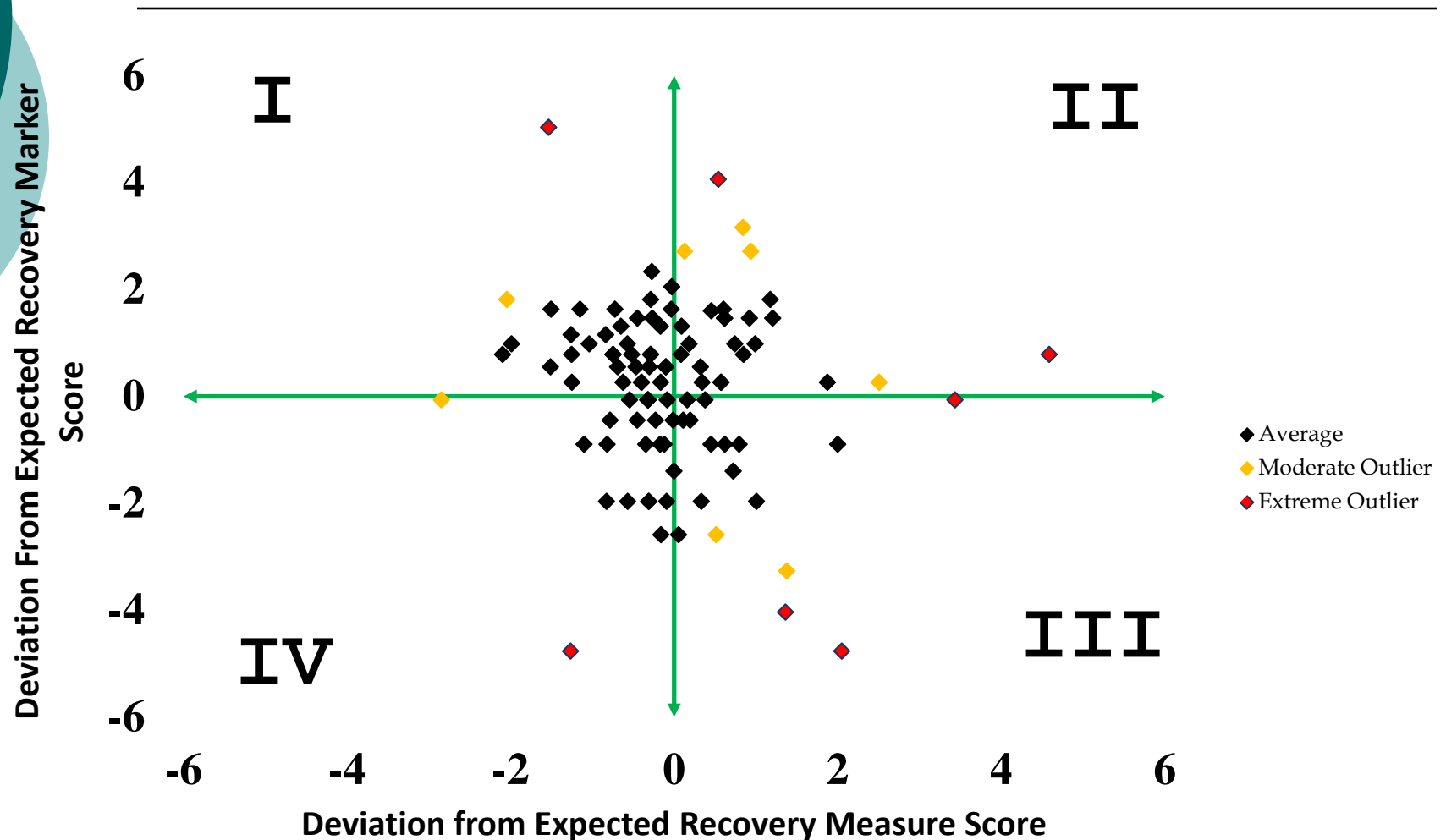


Example of charts being tested



Recovery versus Environmental factors

Chart ML-corrected (Team level)

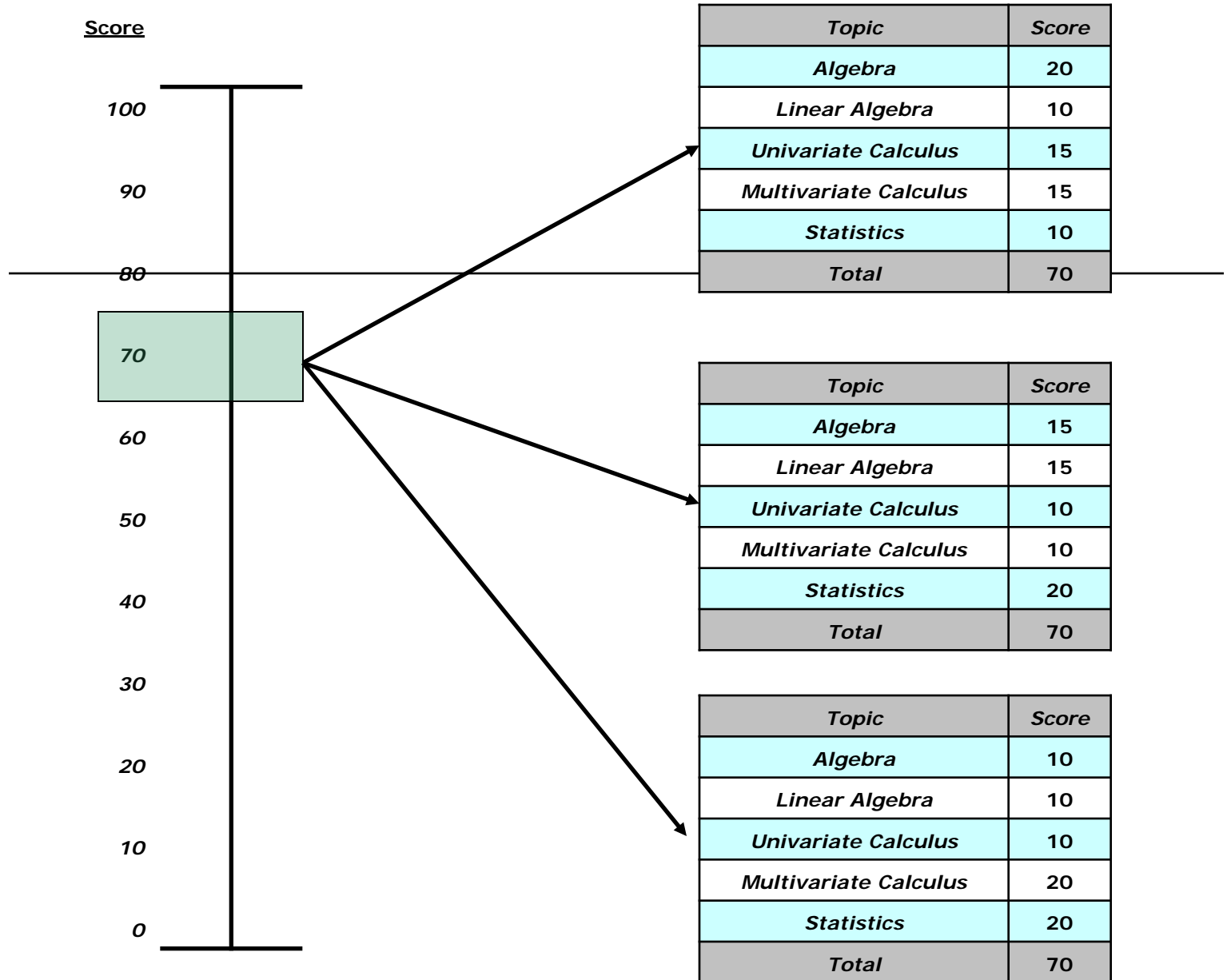




Using IRT to set benchmarks

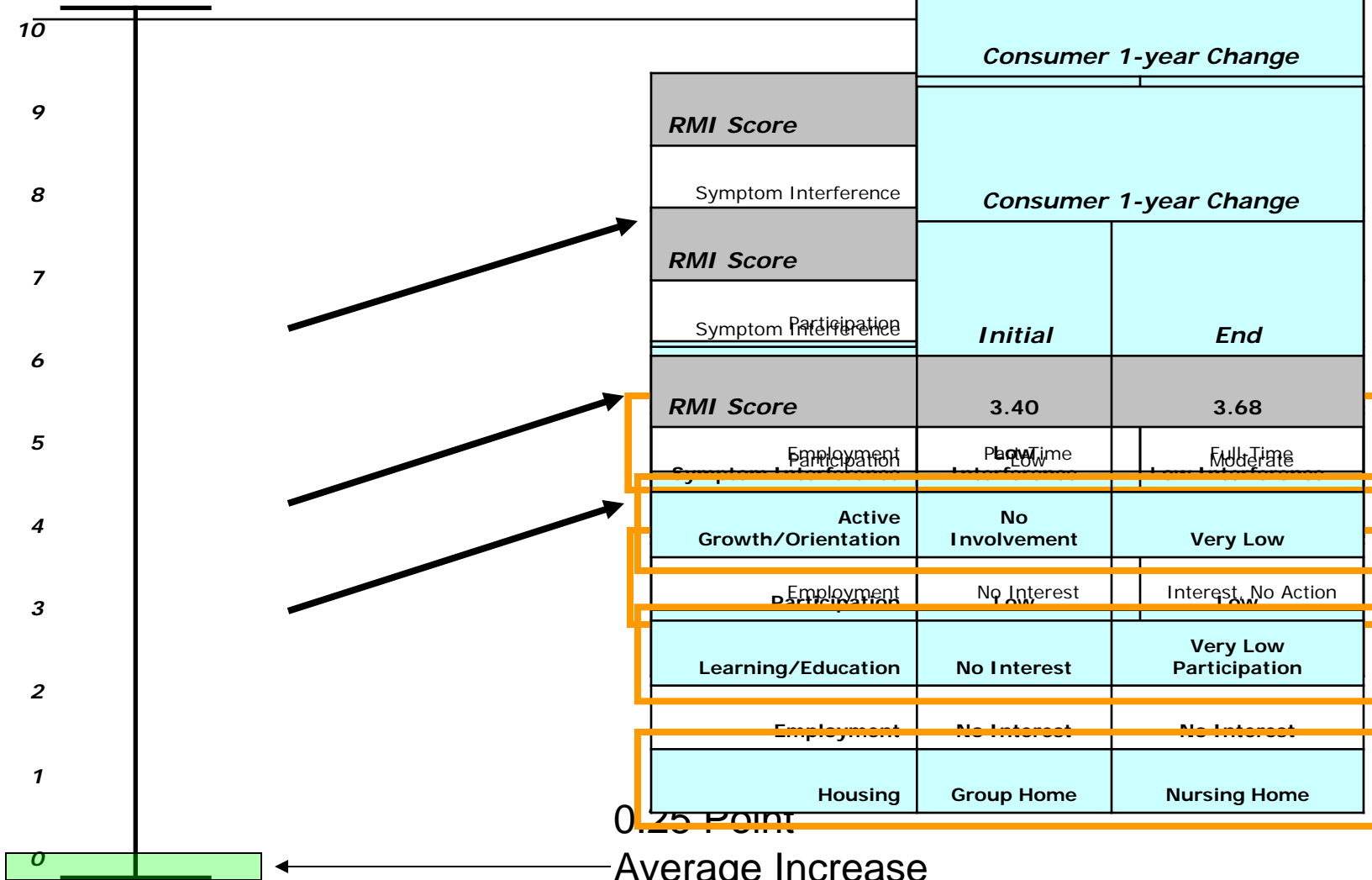
Math Test

Same score, but different pattern of responses for each student.




Environmental Recovery Support Factors

RMI Score





What we like to share with you

- 
-
- Rather than dumbing things down, take some time to teach your stakeholders about complex methods
 - You don't have to spend the next 4 years teaching IRT, MLM, and Regression Discontinuity
 - All you need is a good set of analogies that make sense to your stakeholders



For copies of our presentations/papers,
please visit our websites

<http://www.outcomesmhcd.com>

<http://www.reachingrecovery.org>

<http://www.mhcd.org/AboutUs/EvaluationAndResearch.html>

Antonio.Olmos@mhcd.org

Kathryn.Deroche@mhcd.org

Christopher.Mckinney@mhcd.org