

Relating Automatic Spoken Spanish Test Scores to the ILR Scale


29 October 2004

**East Coast Organization of Language Testers (ECOLT)
George Washington University**

Jennifer Balogh, Jared Bernstein, Isabella Barbier, Elizabeth Rosenfeld

**Ordinate Corporation
Menlo Park, California**

Presentation

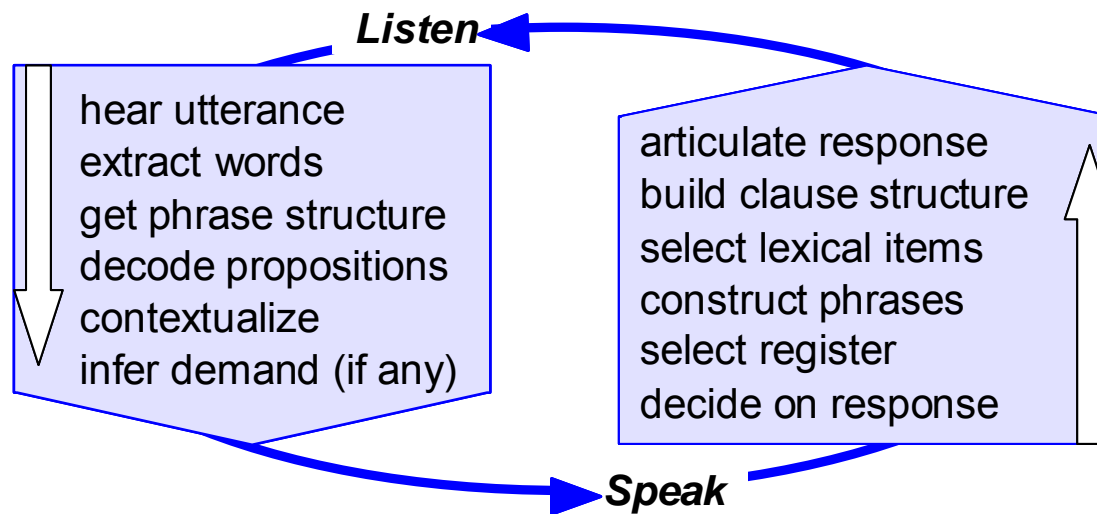
- **Spoken Spanish Test (SST) Description** 
- **Relating SST to ILR scale**
 - **Concurrent validity using ILR scale**
 - **Predicting ILR scores**

Description of SST

- Computerized Spoken Spanish Test
 - Taken over the telephone
 - 15 minutes to complete
 - Landline phone
 - Automated administration and scoring
 - Uses speech recognition technology
 - Scores available on secure web site

SST Construct

- Measures facility in spoken Spanish
 - Ease and immediacy in understanding and producing appropriate conversational Spanish.

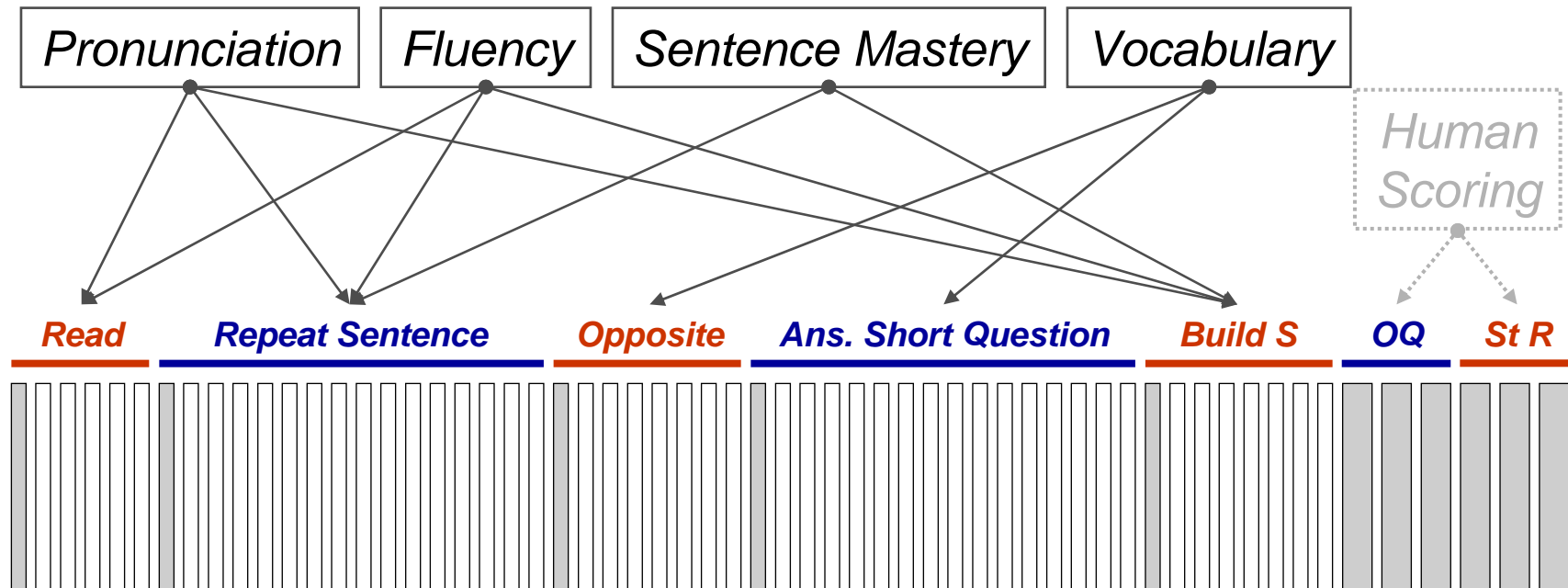


Adapted from Levelt, 1989

SST Design

Test Part	Task Type	Example
Part A	Read Aloud	Julio había recibido de regalo una hermosa bicicleta último modelo. <i>Julio was given the latest model of a beautiful bicycle as a gift.</i>
Part B	Repeat Sentences	El joven camina por la calle. <i>The man walks along the street.</i>
Part C	Say the Opposite	alto <i>high</i>
Part D	Answer Short Questions	¿Cuántas patas tiene un perro? <i>How many legs does a dog have?</i>
Part E	Build Sentences	te / María / ama <i>you / Maria / loves</i>
Part F	Answer Open Questions	¿Prefiere usted vivir en la ciudad o en el campo? Por favor explique su selección. <i>Do you prefer to live in the city or the countryside? Please explain your choice.</i>
Part G	Retell Stories	Tres niñas caminaban a la orilla de un arroyo cuando vieron a un pajarito con las patitas enterradas en el barro...

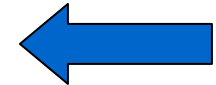
SST Design and Scoring Logic



$$SST = (30\% \text{ Sent.M}, 20\% \text{ Vocab}, 30\% \text{ Fluency}, 20\% \text{ Pron})$$

Presentation

- **Spoken Spanish Test (SST) Description**
- **Relating SST to ILR scale**
 - **Concurrent validity using ILR scale**
 - **Predicting ILR scores**



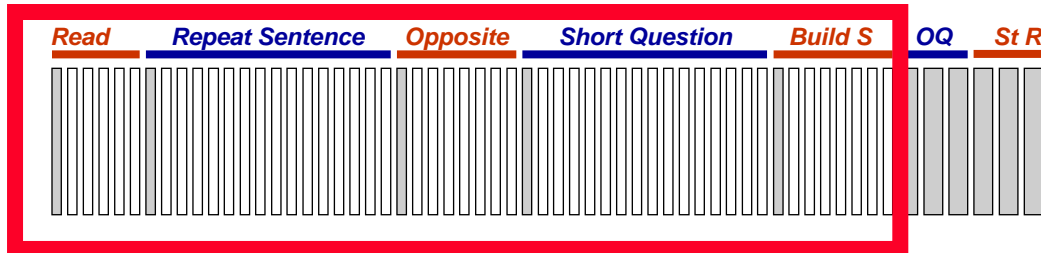
Validity Framework

- State argument
- Assemble evidence
- Evaluate most problematic assumptions
- Restate argument (repeat cycle)

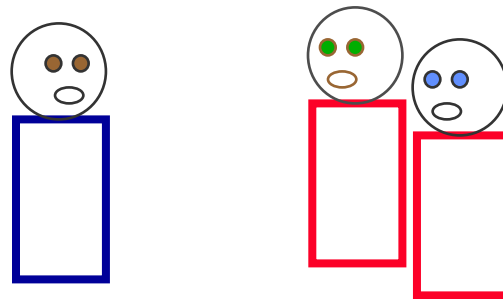
ARGUMENT:

SST scores will be highly correlated with human ratings (ILR scale)

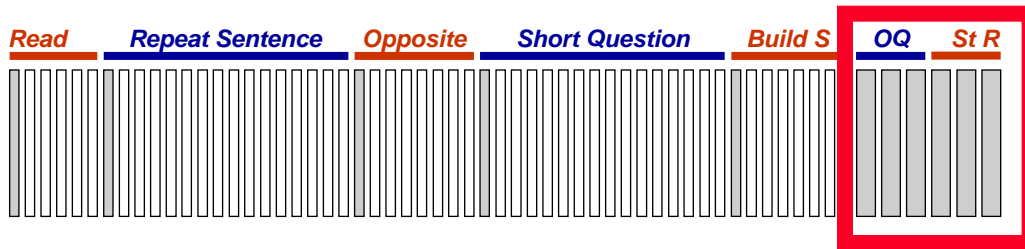
Concurrent Validity Evidence



*SST
Machine Scores*

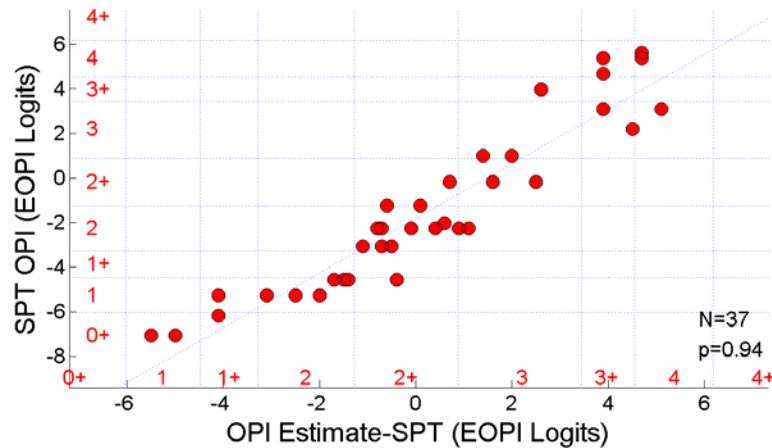


*ILR-SPT
Human Interview Scores*



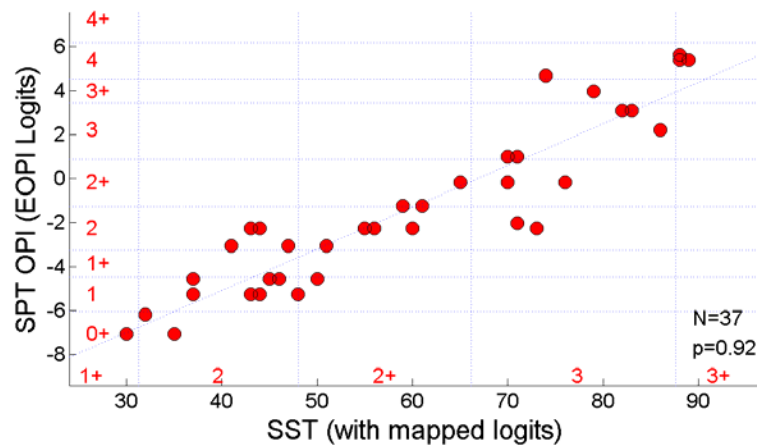
*ILR-SPT Estimates
(2 human raters per)*

SPT OPI (SPT Interviews)



SPT OPI ~ ILR Estimate-SPT

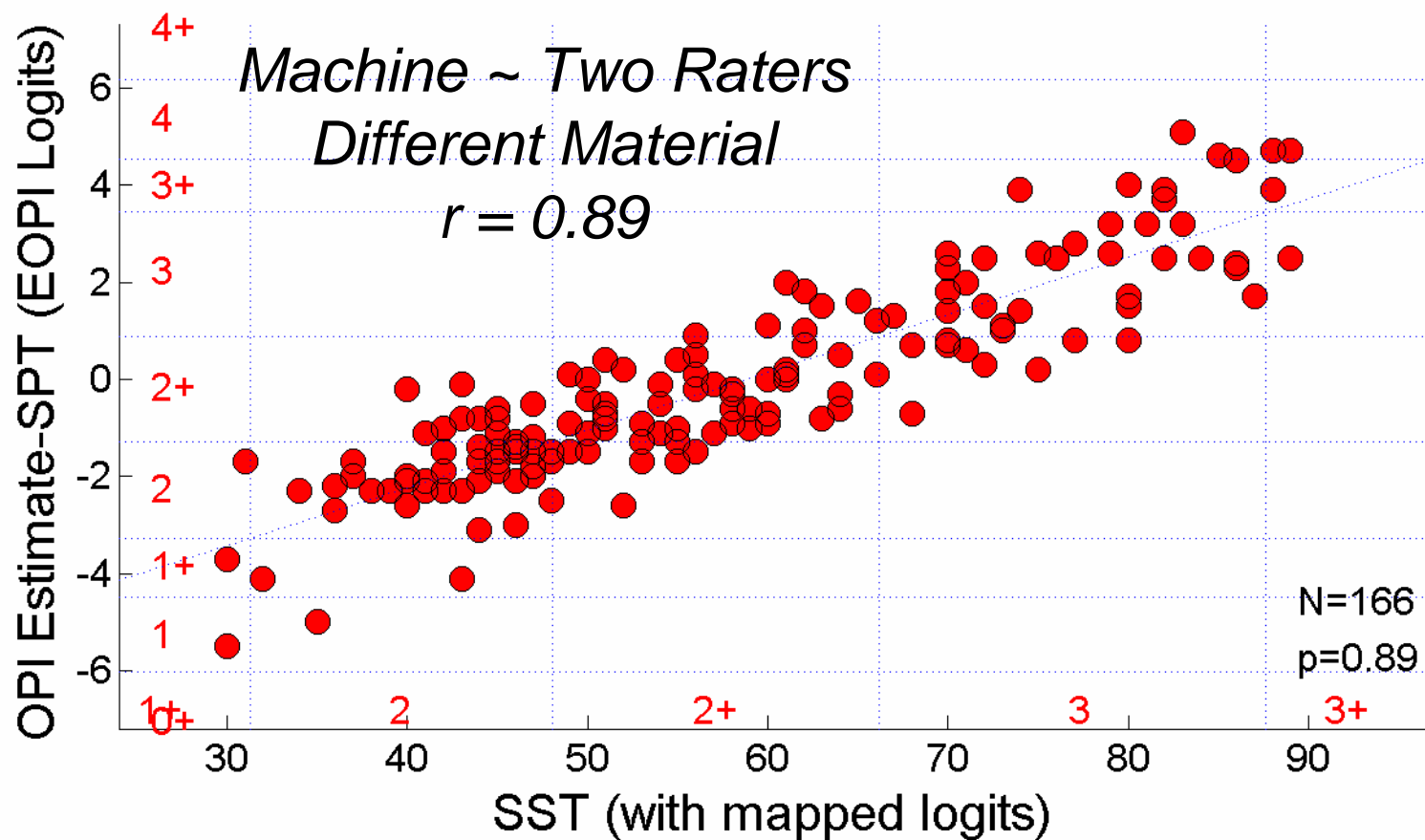
*Same Two Raters
Different Material
 $r = 0.94$*




SPT OPI ~ SST

*Two Raters ~ Machine
Different Material
 $r = 0.92$*

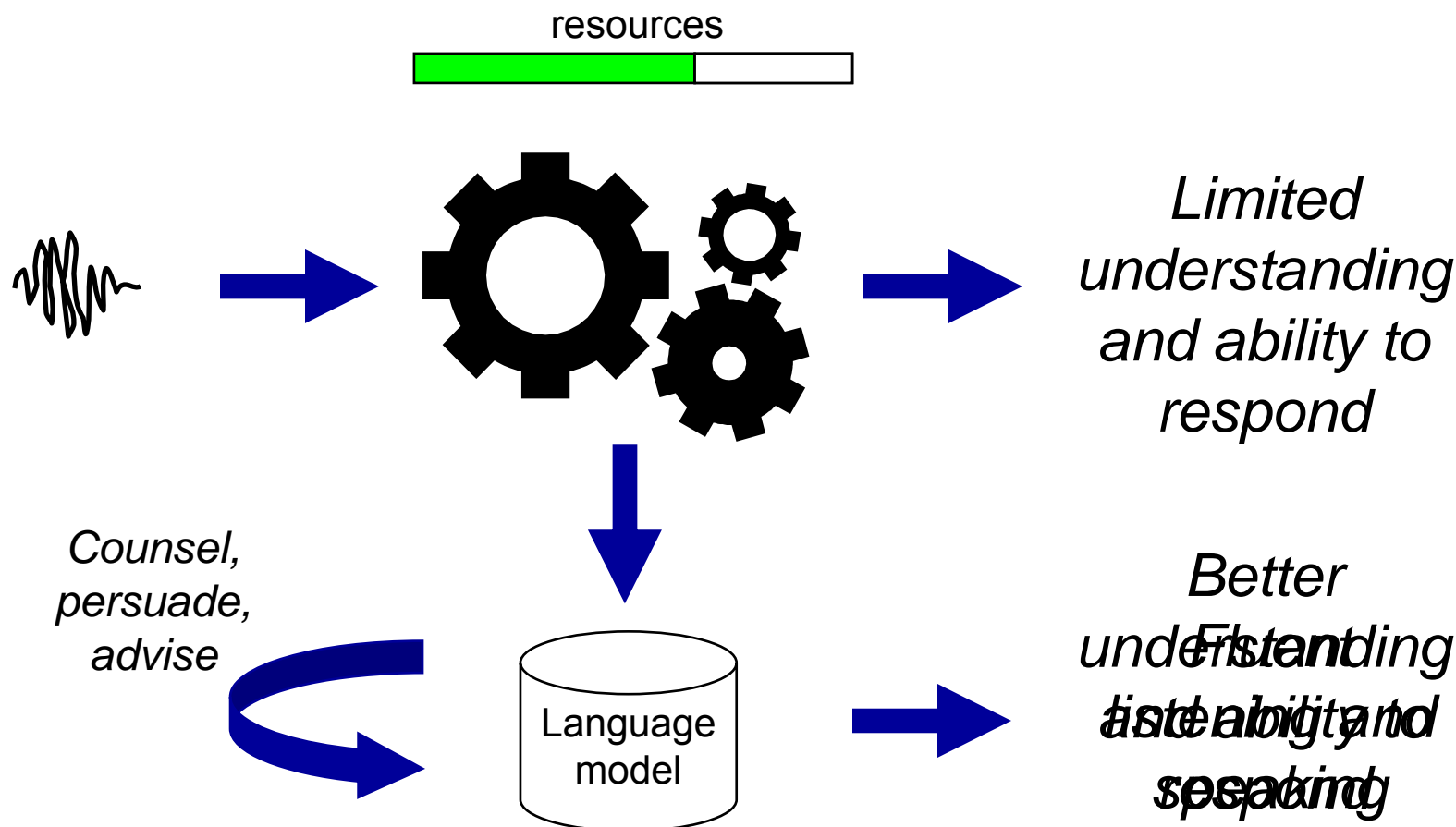
SST ~ ILR Estimate-SPT



Validity Framework

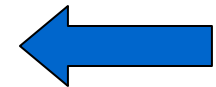
- State argument
- Assemble evidence
- Evaluate most problematic assumptions 
 - Why are correlations so high when constructs are different?
- Restate argument (repeat cycle)

Theory of Language Proficiency: Automaticity



Presentation

- **Description of Spoken Spanish Test**
- **Relating SST to ILR scale**
 - **Concurrent validity using ILR scale**
 - **Predicting ILR scores**



Argument

SST scores will accurately predict ILR lower bound scores for military use

1. Methodology
2. Evidence

Predicting ILR Scores from SST Scores

1. Express ILR scores in logits

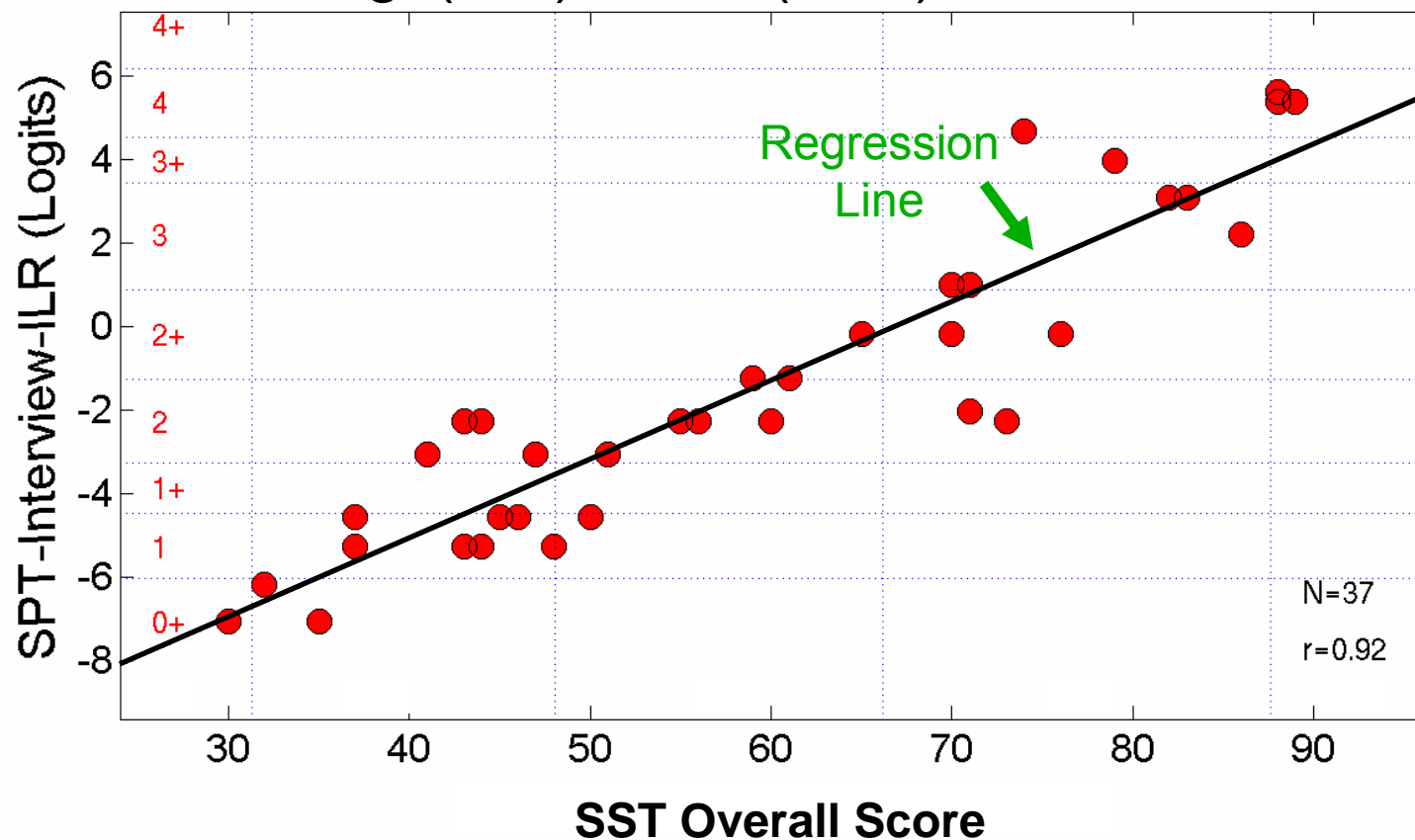
Mapping based on IRT analysis of ILR estimates

Double scoring of 6 responses (same 2 raters)

2. Generate regression equation

Predicting ILR Scores from SST Scores

$$\text{logit(ILR)} = 0.19(\text{SST}) - 12.69$$



Predicting ILR Scores from SST Scores

1. Express ILR scores in logits

Mapping based on IRT analysis of ILR estimates

Double scoring of 6 responses (same 2 raters)

2. Generate regression equation

$$\text{logit(ILR)} = 0.19(\text{SST}) - 12.69$$

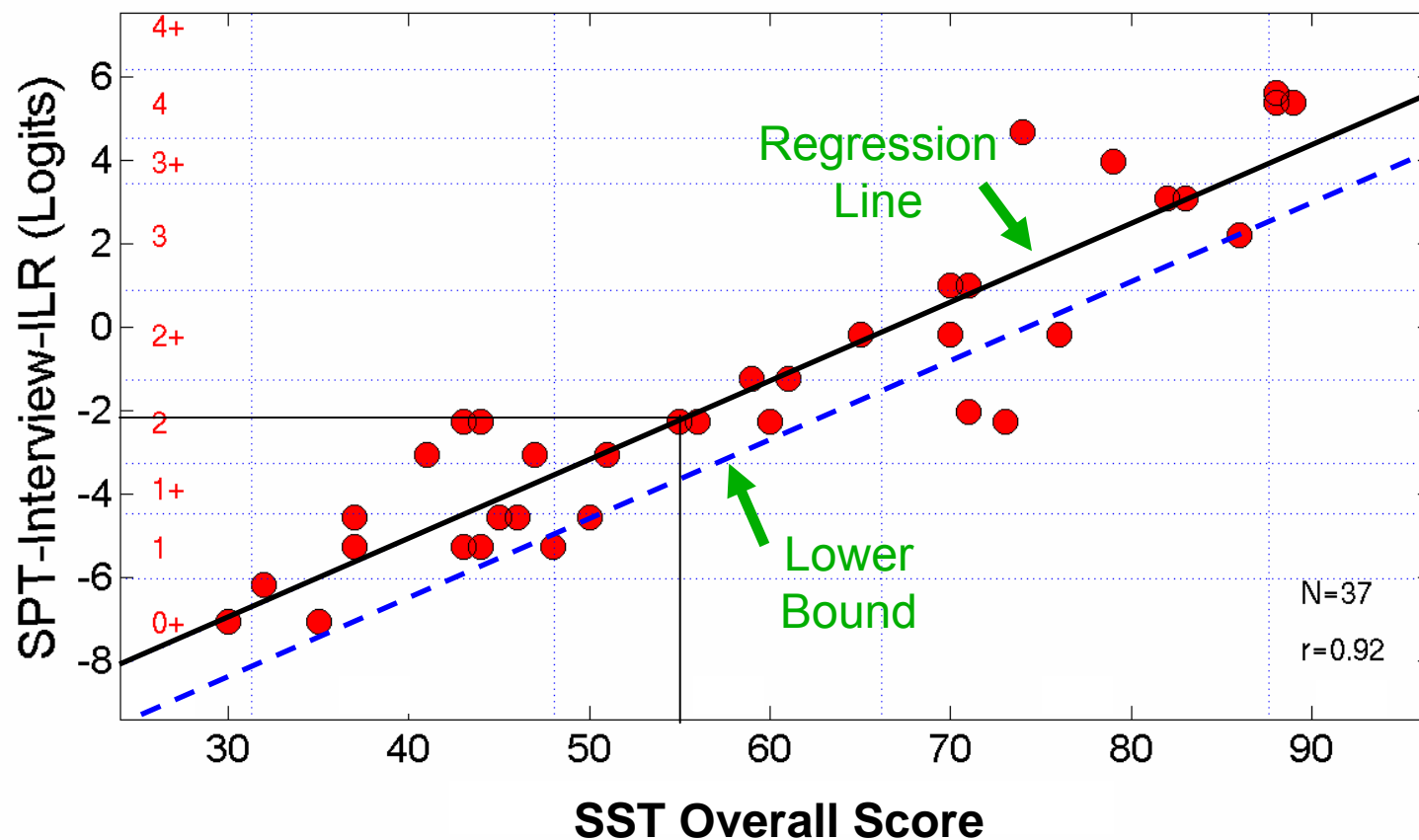
3. Convert logits to ILR scale

Use thresholds from FACETS analysis

Predicting ILR Scores from SST Scores

$$\text{LowerBound(ILR)} = \text{ILR} - (\text{t-score})(\text{standard error of the estimate})$$

For 80% confidence, 36 df: $t = 0.85$ (one tailed)



Concordance Table

SST Overall Score	Best Estimate of ILR Score	\geq ILR Score with 80% Confidence
20	0	0
21 - 35	0+	At least 0+
36 - 43	1	At least 0+
44 - 49	1+	At least 1
50 - 55	2	At least 1+
56 - 60	2	At least 2
61 - 66	2+	At least 2
67 - 71	2+	At least 2+
72 - 77	3	At least 2+
78 - 80	3	At least 3

Validity Evidence

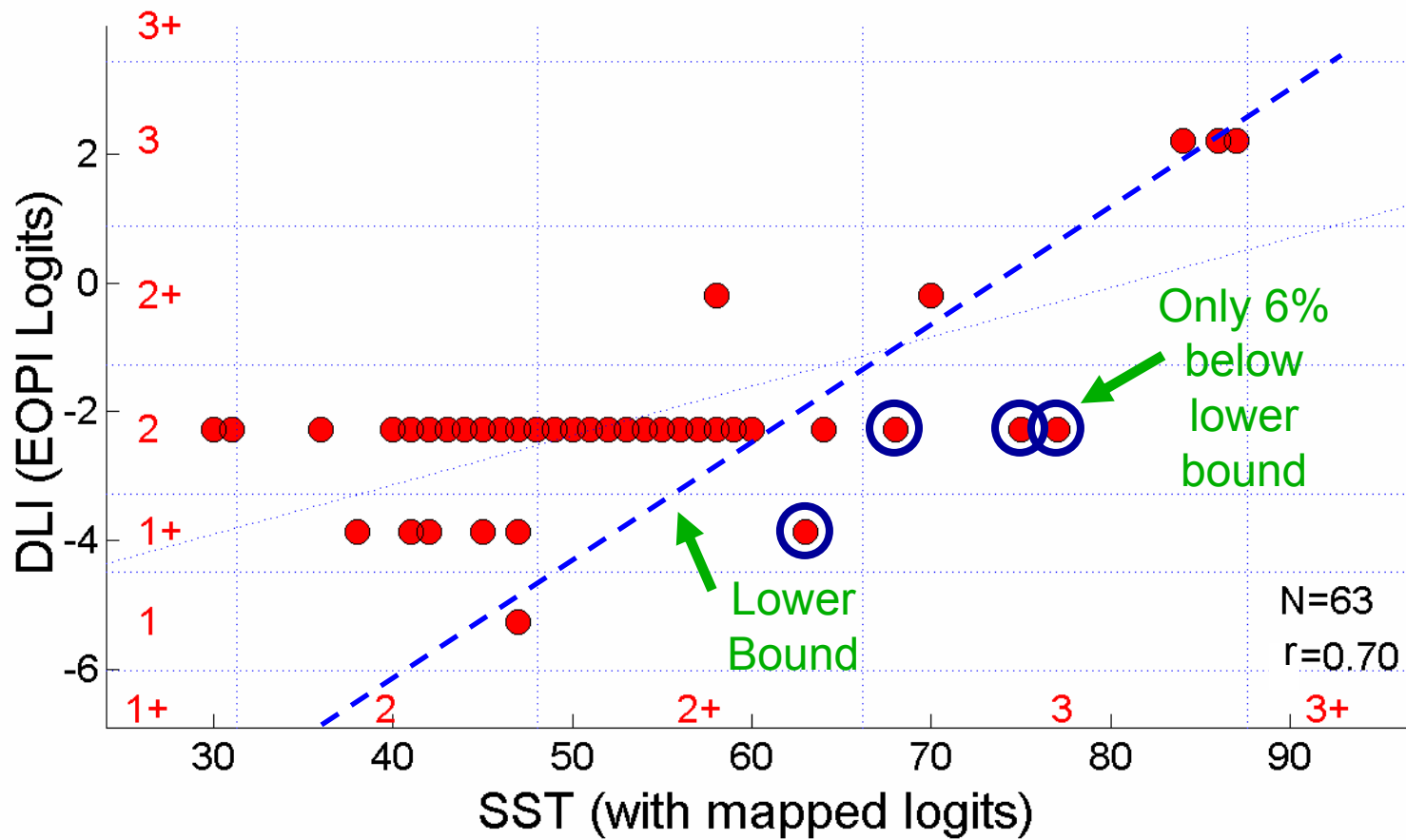
Validate lower bound prediction

- 92% of observed ILR SPT interview scores \geq lower bound
- 92% of observed ILR SPT estimates \geq lower bound

What about data not used to generate scores?

DLI OPI data

Validity Evidence: DLI OPIs



Conclusions

- SST scores are highly correlated with human ratings on the ILR scale

Automaticity theory explains why correlations are high even though constructs are different

- SST scores accurately predict ILR lower bound scores for military use

Lower bound cut-off scores at 80% confidence account for 92% of observed scores