

To Thine Own Self Be True: A Five-Study Meta-Analysis on the Accuracy of Language-Learner Self-Assessment

Troy L. Cox, PhD

Associate Director of Research and Assessment
Center for Language Studies
Brigham Young University

“To Thine Own Self Be True.”

-Polonius

- As a busy body in Hamlet known for his platitudes that he himself did not follow, a close reading should really be:
 - *Beware of listening to men who give counsel*
 - **You’ve been warned.**



A Tale of Five Studies

- Brown, A., Dewey, D. & Cox, T. (2014). Assessing the Validity of Can-Do Statements in Retrospective (Then-Now) Self-Assessment. *Foreign Language Annals*, 47(2), 261-285.
- Nielson, J., Dewey, D., & T. Cox (2016) *Second-language self-assessment: The influence of video demonstrations on their accuracy*. Paper at the Georgetown University Roundtable on Linguistics (GURT). Washington, DC, USA.
- Summers, M., Cox, T. & McMurray, B. (2016) *To Thine Own Self Be True: How Well Do Can-Do Statements Predict Ability?* Paper presented at 2016 ACTFL annual convention of American Council on the Teaching of Foreign Languages. Boston, MA, USA
- Cox, T., Bown, J., Bell, T & Evans, J. (2017) Does the question language in advanced L2 reading proficiency assessments make a difference? In S. Gass & P. Winke (Eds.) *Foreign Language Proficiency in Higher Education* (working title). Springer Publishing, New York.
- Peterson, J. & Cox, T. (2017). *Performance Self-Appraisal Calibration of ESL Students on a Reading Comprehension Multiple-Choice Assessment*. Thesis.

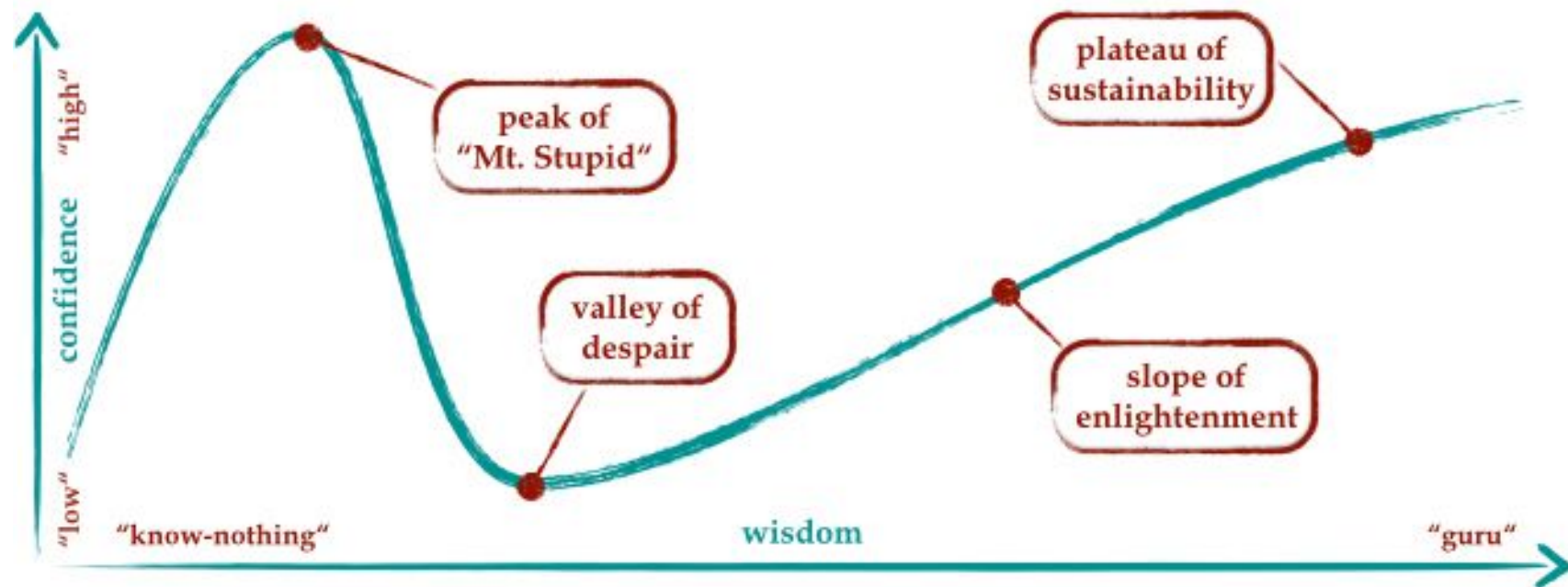
A Table of Five Studies

	Native Language	Second Language	Instrument
Speaking	English	Russian	Self-Retrospective Statements based on Can-Do Subheadings
	English	Spanish	Can-Do Statements with Video Exemplars
Speaking & Writing	Spanish, Chinese, Korean, Portuguese, Russian, French & Others	English	Survey questions based on the Can-Do Statements
Reading	English	Russian	Confidence slider after each question. Passages in Russian with questions in both English and Russian
	Spanish, Portuguese, Japanese, Korean, Chinese, & Others	English	Confidence slider after each question. Passages and questions in English

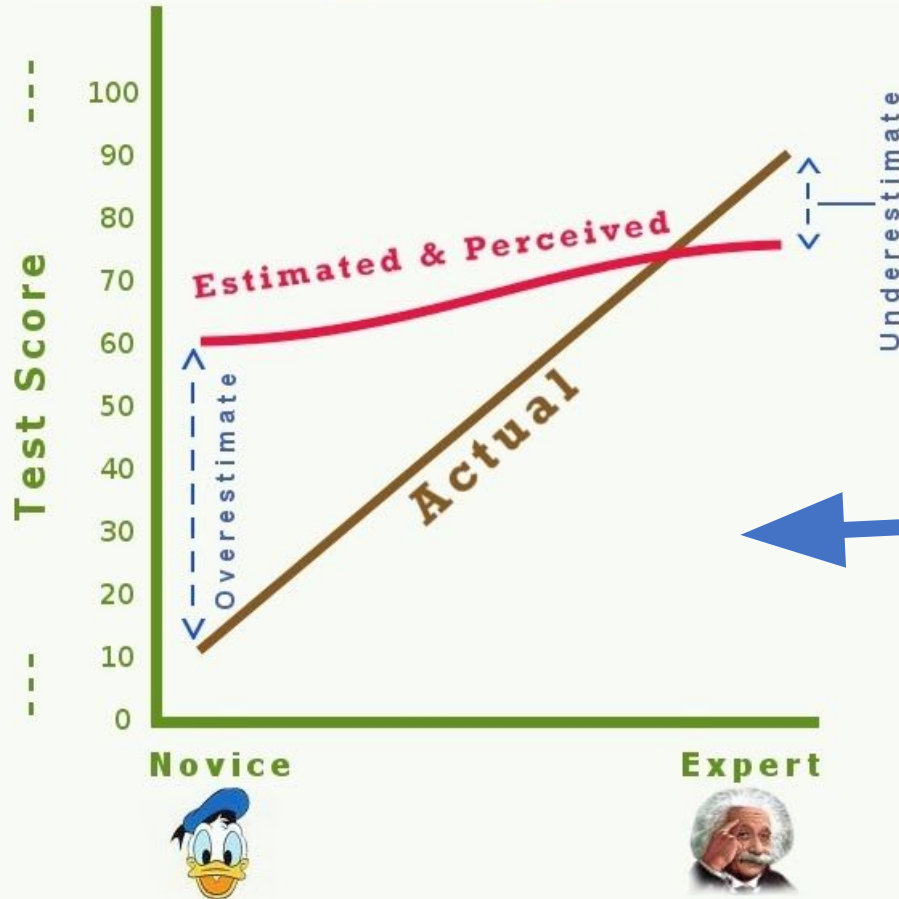
Self-Assessment in General

- Can take less time than traditional tests
- Problems with cheating and test security can be minimized
- Learner motivation, autonomy and self-regulation can be increased
- Can correlate fairly well with objective measures (.60-.90)
- Correlations tend to be higher for more objective disciplines (math, science, etc.)
- Dunning-Kruger Effect
 - Accuracy of self-assessments tends to increase with proficiency
 - The more experience a person has with a task, the better they self-assess
-

Dunning-Kruger effect



Dunning-Kruger Effect



Imposter
Syndrome

Self-Assessment in Language Studies

- There are mixed results
 - Correlations ranging from .20 to .90
- Correlations are lowest for reading comprehension
- Cultural background can affect accuracy of ratings
- The more specific the wording of items, the more accurate the self-evaluations tend to be

Timing of Assessment

- We are often overly confident when we have less experience
- Learners' confidence can decrease as the event gets closer
- Learners' estimates are more accurate when they have had experience
- Learners' rules change over time

Background: Description of NCSSFL–ACTFL Can Do Statements

ACTFL Standard	Description
Distinguished	I can communicate reflectively on a wide range of global issues and highly abstract concepts in a culturally sophisticated manner.
Superior	I can communicate with ease, accuracy, and fluency. I can participate fully and effectively in discussions on a variety of topics in formal and informal settings. I can discuss at length complex issues by structuring arguments and developing hypotheses.
Advanced High	I can express myself freely and spontaneously, and for the most part accurately, on concrete topics and on most complex issues. I can usually support my opinion and develop hypotheses on topics of particular interest or personal expertise.
Advanced Mid	I can express myself fully not only on familiar topics but also on some concrete social, academic, and professional topics. I can talk in detail and in an organized way about events and experiences in various time frames. I can confidently handle routine situations with an unexpected complication. I can share my point of view in discussions on some complex issues.
Advanced Low	I can participate in conversations about familiar topics that go beyond my everyday life. I can talk in an organized way and with some detail about events and experiences in various time frames. I can describe people, places, and things in an organized way and with some detail. I can handle a familiar situation with an unexpected complication.
Intermediate High	I can participate with ease and confidence in conversations on familiar topics. I can usually talk about events and experiences in various time frames. I can usually describe people, places, and things. I can handle social interactions in everyday situations, sometimes even when there is an unexpected complication.
Intermediate Mid	I can participate in conversations on familiar topics using sentences and series of sentences. I can handle short social interactions in everyday situations by asking and answering a variety of questions. I can usually say what I want to say about myself and my everyday life.
Intermediate Low	I can participate in conversations on a number of familiar topics using simple sentences. I can handle short social interactions in everyday situations by asking and answering simple questions.
Novice High	I can communicate and exchange information about familiar topics using phrases and simple sentences, sometimes supported by memorized language. I can usually handle short social interactions in everyday situations by asking and answering simple questions.
Novice Mid	I can communicate on very familiar topics using a variety of words and phrases that I have practiced and memorized.
Novice Low	I can communicate on some very familiar topics using single words and phrases that I have practiced and memorized.

Background: Vertical Scale (Wright) Map

Logits

3

2

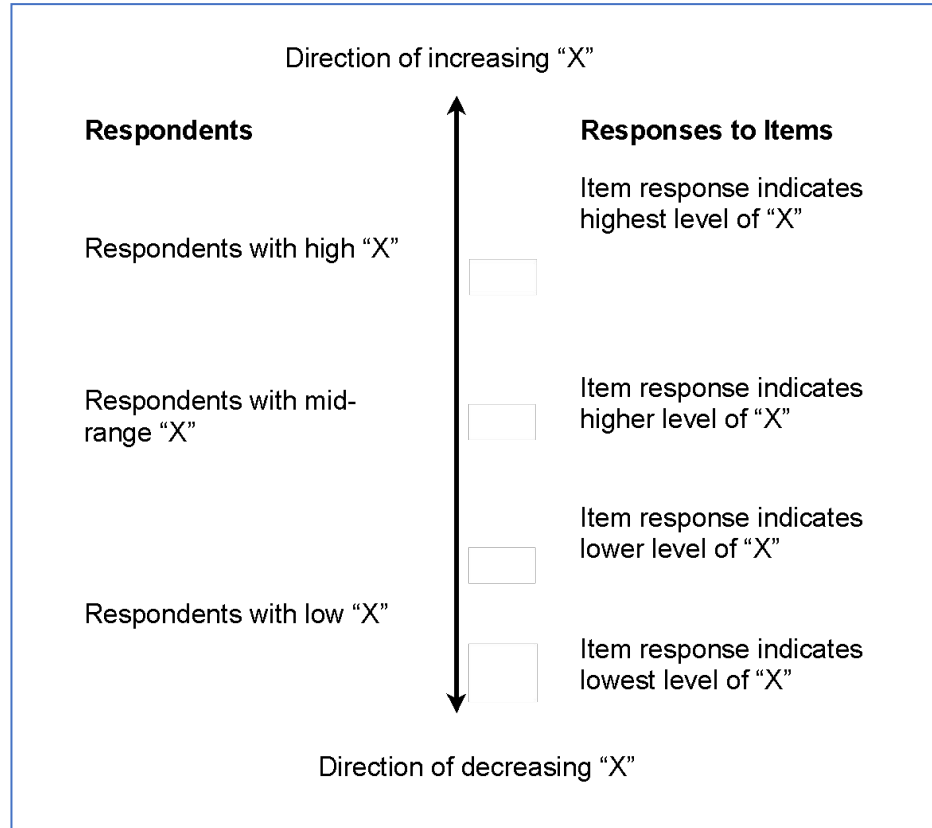
1

0

-1

-2

-3



When people and items have the same logit (log odds) the probability of a correct response is 50%

Study 1

Speaking (English —> Russian)

Self Retrospection

Tony Brown, Dan Dewey, & Troy Cox

Participants

- Upper-level Russian learners who had participated in internships abroad (2006-2014)
- Learners had completed OPIs before and after their internships
- 64 learners contacted and asked to complete the self-retrospective survey
- 36 learners responded (27 male & 9 female)

Statement Examples

I could support my opinions clearly and precisely and construct hypotheses.

	Could not do this even with extensive preparation	Unsure as to whether I could or could not do this	Could do this with extensive preparation	Could do this with minimal preparation	Could do this without any preparation
Pre-Internship	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Post-Internship	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

I could discuss complex information in debates or meetings.

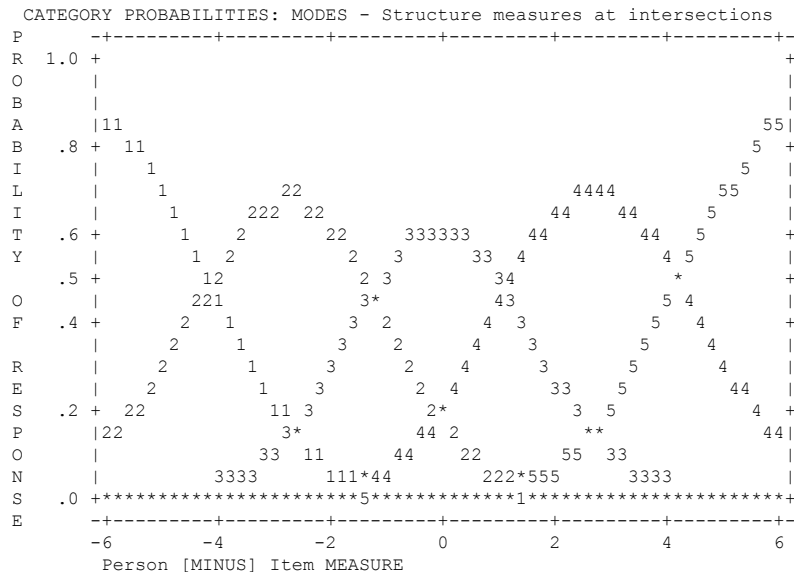
	Could not do this even with extensive preparation	Unsure as to whether I could or could not do this	Could do this with extensive preparation	Could do this with minimal preparation	Could do this without any preparation
Pre-Internship	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Post-Internship	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Research Question 1:
What is the reliability of the
self-assessment instrument used in this
study?

Does the scale function?

Does it separate persons and items?

Scale Diagnostic



Category	Absolute Frequency	Relative Frequency	Average Measure	Outfit	Threshold	SE
1-Can't Do	17	1%	-2.98	.98		
2-Unsure	149	6%	-.30	1.61	-4.13	0.26
3-W/ Ext prep	544	22%	.60	.80	-1.24	0.11
4- W/ Some prep	1070	43%	2.80	.86	1.13	0.06
5-No Prep	720	29%	4.91	1.02	4.24	0.07

Vertical Scale

Research Question 1:
What is the reliability of the self-assessment instrument used in this study?

Item	THEN Person Ability Estimate Mean = 1.88, SD = 2.03	NOW Person Ability Estimate Mean = 3.76, SD = 1.83	Item Difficulty Parameter Mean = 0, SD = .17	Scale
7		9-BAY 9-ANT 7-RAY 10-FAR 9-MAR 8-SPE		5
6	6-RAY 8-BAY 3-RIC			
5	3-MAR 3-MAH 3-MIL	10-MIL 9-RIC 9-KUN 3-BRA		
4		9-WAL 9-BAL 3-TIF 9-DAL 3-ASH 3-JOH 9-AND 10-NEW		4
3	8-KUN 3-JOH 3-OLS 9-LEM 7-ANT 3-WAL	3-PER 9-CAL 3-HAZ 7-YOU 10-ROB 9-TAK 10-LEM 9-CHA		
2	7-PER 7-GOD 9-CAL	9-LAY 3-ASH 6-TIF 7-BAL 3-DAL		
1	9-NEW 8-TAK 3-MOW	7-HAZ 6-AND 6-CAR 8-WAH 3-CAR		3
0	3-ROB 7-SIM 7-VER			
-1	7-WAH 6-BRA			
-2				2
-3	6-HEP			
-4				1

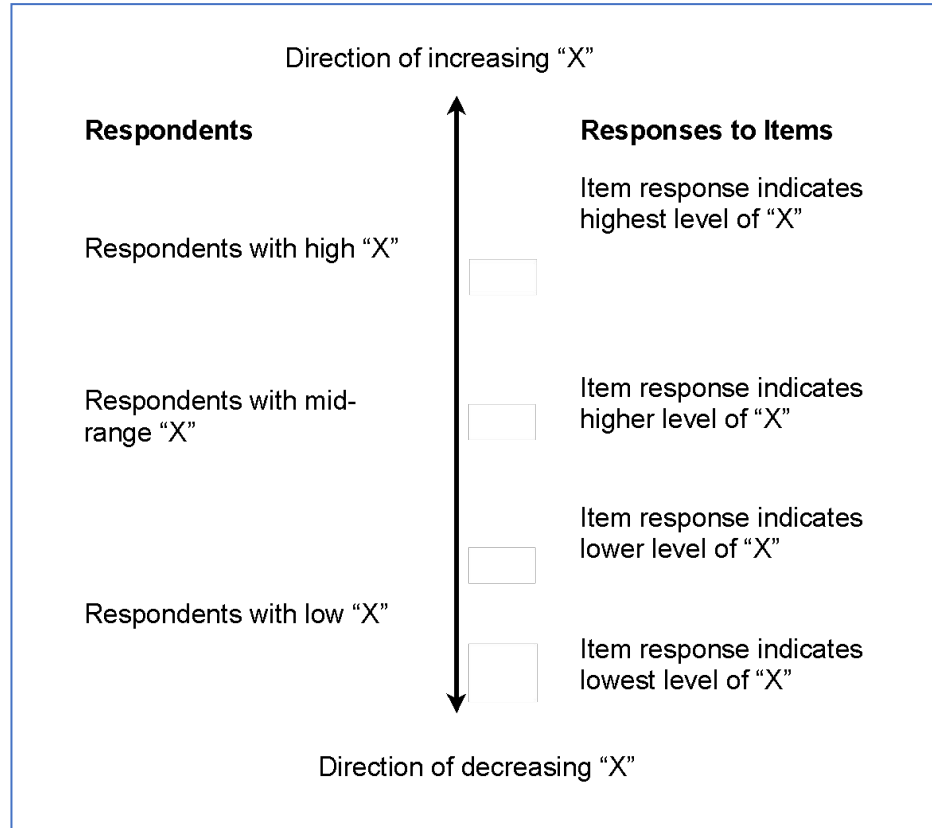
Then-Now Self-Assessment Vertical Scale Map

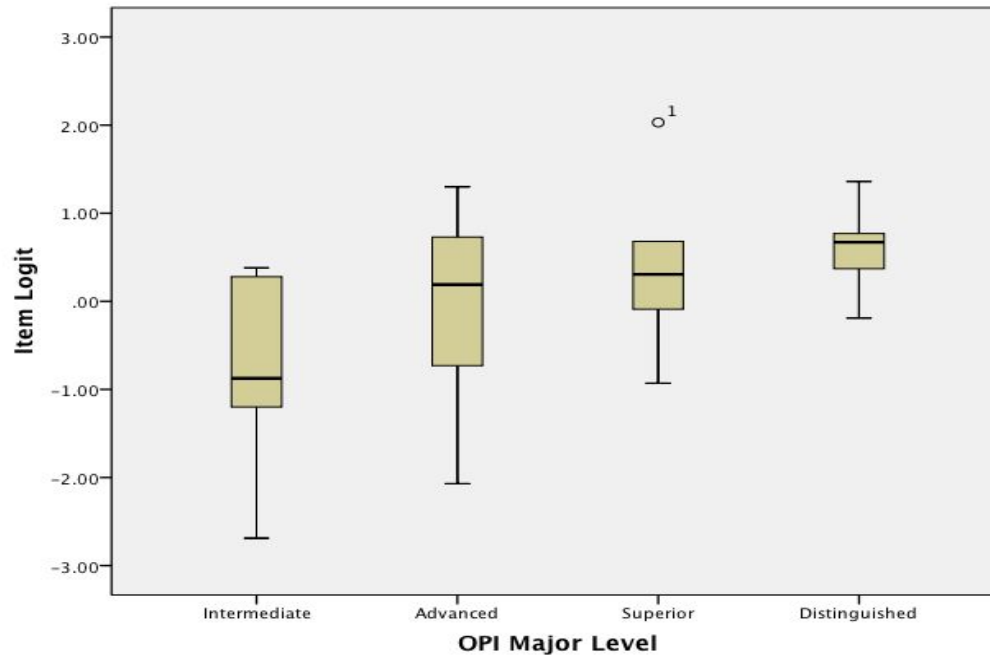
	Students	Items
Separation Reliability	.96	.95
Separation Strata	5.57	4.52

Research Question 2:

To what extent do the survey items ascend in a hierarchy of difficulty levels based on the ACTFL speaking proficiency guidelines?

Do the Can-Do statements fall in place with the construct map?





N = 6

Mean = -.83

N = 19

Mean = -.02

N = 6

Mean = .38

N = 5

Mean = .60

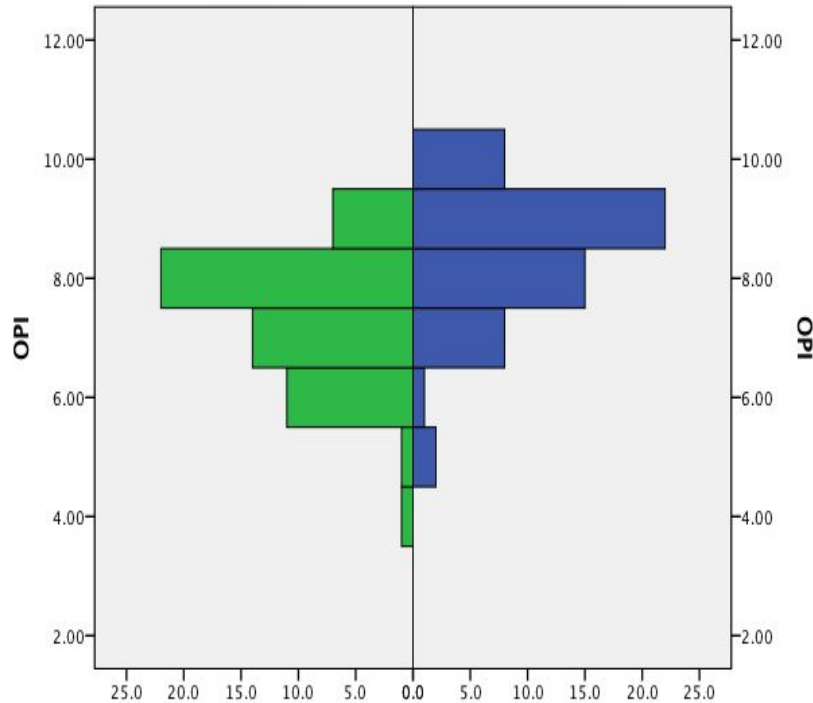
An independent measures ANOVA found that the differences were **NOT** statistically significant ($F = 2.36$, $df = 3$, $p = .09$)

Research Question 3:

What is the predictive validity of self-assessment items in predicting an OPI score?

- Did OPI scores change over time?
- Did self-assessment change over time?
- What is the relationship between Then-Now scores OPI ratings?

Did OPI scores change over time?—YES

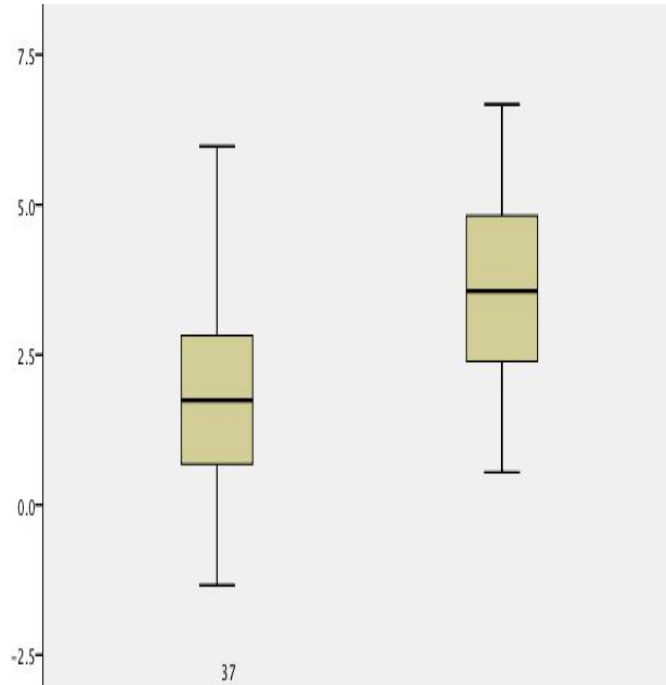


Wilcoxon Matched Pairs
Signed Ranks Test

$Z = -5.57$ $p < .001$,
41 instances of the
subjects scoring higher
on the post-test.

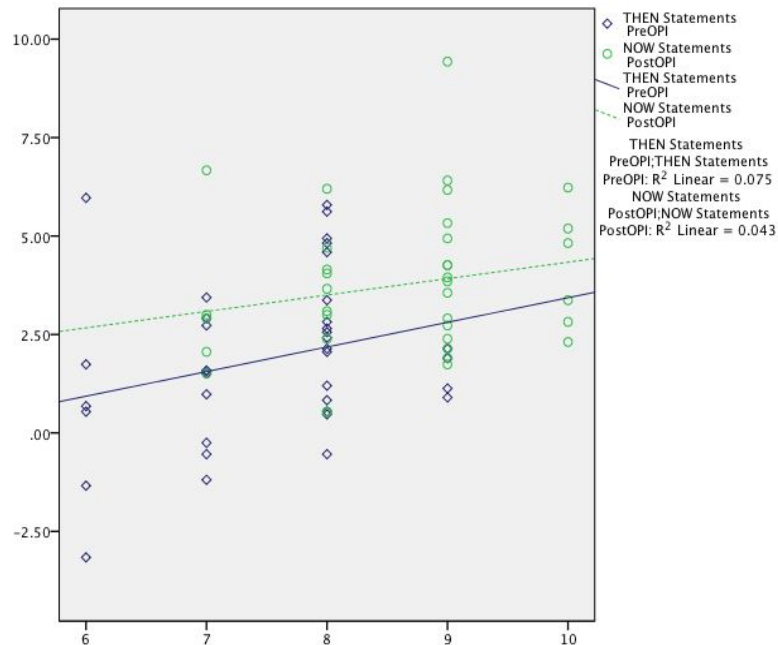
There were 12 instances
in which subjects had
the same rating on the
pre and post and only 2
instances in which a
student scored lower on
the post-internship OPI.

Did Then-Now scores change over time?—YES



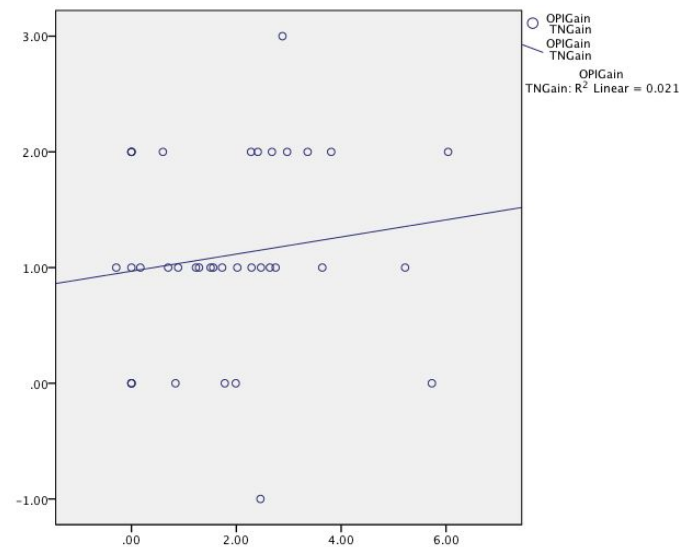
Paired Samples T-Test.
The difference of the means was -1.88 (sd = 1.64, 95% CI -2.43, -1.33) resulting in $t = -7.00$, $df = 36$, $p < .001$

What is the relationship between Then-Now scores OPI ratings?



	N	Spearman's Rho	P (one-tailed)	Effect Size
Then and Pre-OPI	37	.27	.06	Small to Medium
Now and Post-OPI		.21	.11	Small to Medium

What is the relationship between Then-Now score gain OPI rating gain?



	N	Spearman's Rho	P (one-tailed)	Effect Size
ThenNow Gain and OPI Gain	37	.21	.10	Small to Medium

Research Question 3:

What is the predictive validity of self-assessment items in predicting an OPI score?

The relationship between self-assessment and OPIs are slight with a small to medium effect size. Self-assessment can provide some useful information, but is insufficient to replace external assessment.

OUR RESULTS

- Are learners overly confident because they still have insufficient experience to make accurate judgments?

Study 2

Speaking (English —> Spanish)

Will videos make a difference?

John Nielson, Dan Dewey, & Troy Cox

Instrument

- Items Consisted of Statement with Examples
- Progressed across 7 sublevels from Intermediate Low to Superior
- Each sublevel had 3 items
- Adaptive logic used in administration
- **No-Video Survey**
 - 21 plain can-do items
- **Video Survey**
 - 21 items with can-do plus video
 - Videos come from ACTFL recordings of OPI's

No-Video Example

I can compare and contrast life in different locations and in different times.

EXAMPLES

- explain how life has changed since I was a child and respond to questions on the topic.
- compare different jobs and study programs in a conversation with a peer.
- explain how technology has changed our lives while discussing this topic with another.

Not at all



With great difficulty



With some difficulty



Easily



Quite easily



I can ask and answer questions on factual information that is familiar to me.

EXAMPLES

- geography
- history
- art
- music
- math
- science
- language
- literature

Not at all



With great difficulty



With some difficulty



Easily



Quite easily



Video Example



I can compare and contrast life in different locations and in different times.

EXAMPLES

- explain how life has changed since I was a child and respond to questions on the topic.
- compare different jobs and study programs in a conversation with a peer.
- explain how technology has changed our lives while discussing this topic with another.

Not at all



With great difficulty



With some difficulty



Easily

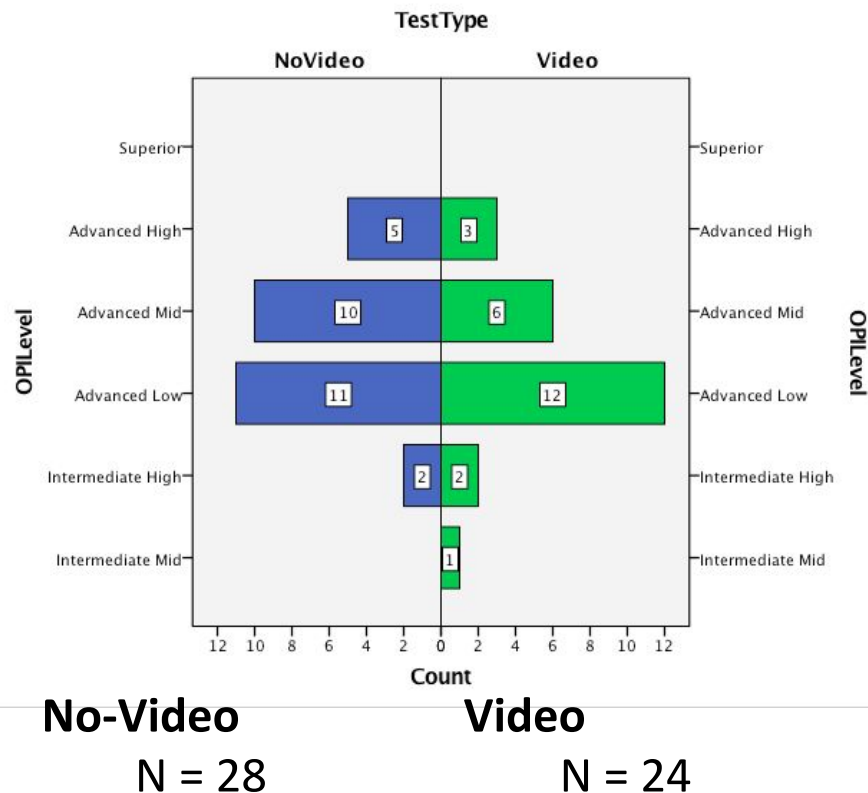


Quite easily



Participants

- Qualtrics survey sent to 322 Spanish students who had been scheduled through the CLS to take an OPI within the last year.
- Randomly assigned to No-Video or Video group.
- 68 (21%) started the survey.
- 54 (17%) completed the survey
- 2 responses excluded due to missing OPI scores.

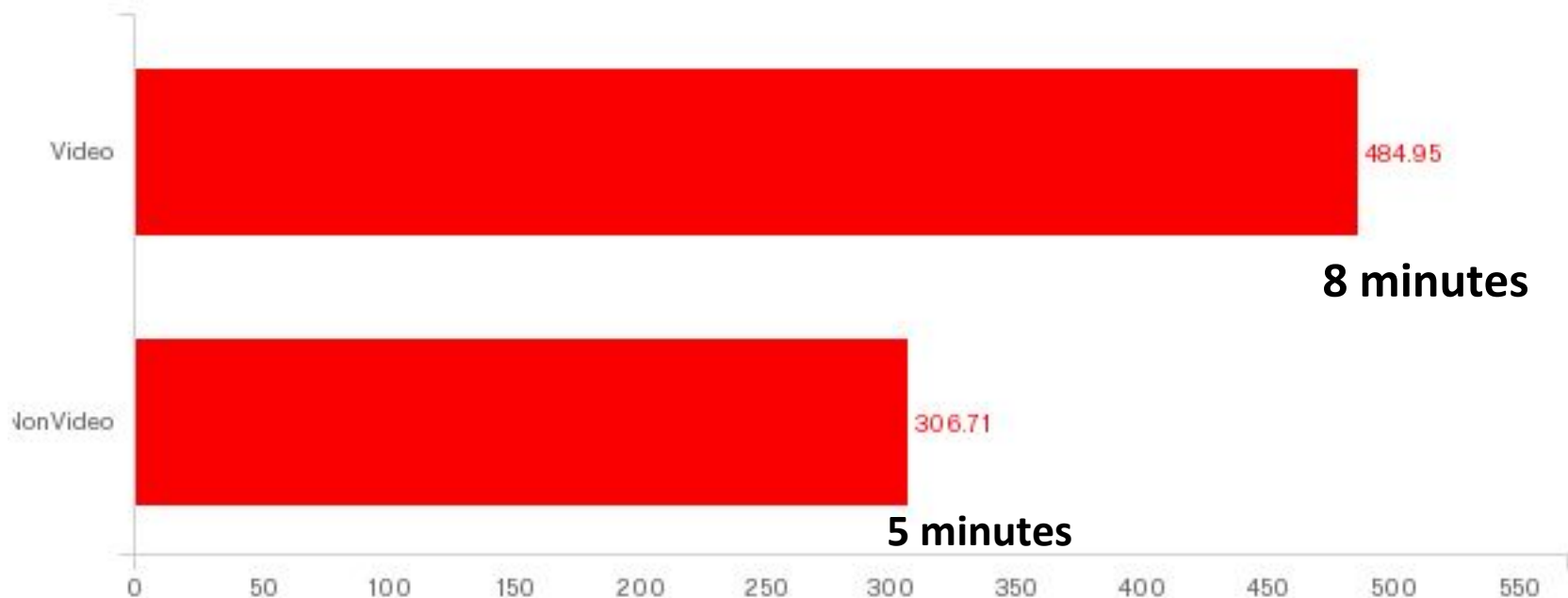


Average Time

Video Group = 22

No Video Group = 31

Filtered with 1 Hour Max (7 people [6 Video, 1 No-Video] excluded).



Research Questions:

To what extent are a No-Video and Video can-do self-assessment survey comparable in terms of...

- a. rating scales?
- b. Instrument reliability?
- c. Intended vs. actual item difficulty?
(Intended ACTFL level and item logit)
- d. Predictive validity of person ability?
(OPI level and person logit)

RQ: To what extent are a No-Video and Video can-do self-assessment survey comparable in terms of...

Score Card: No-Video vs. Video Rasch Analysis	No-Video	Operator (=, < or >)	Video
Rating Scales?			
Instrument Reliability?			
Intended vs. Actual Level Difficulty? (Intended ACTFL level and item logit)			
Predictive Validity of Person Ability? (OPI level and person logit)			

Score Card

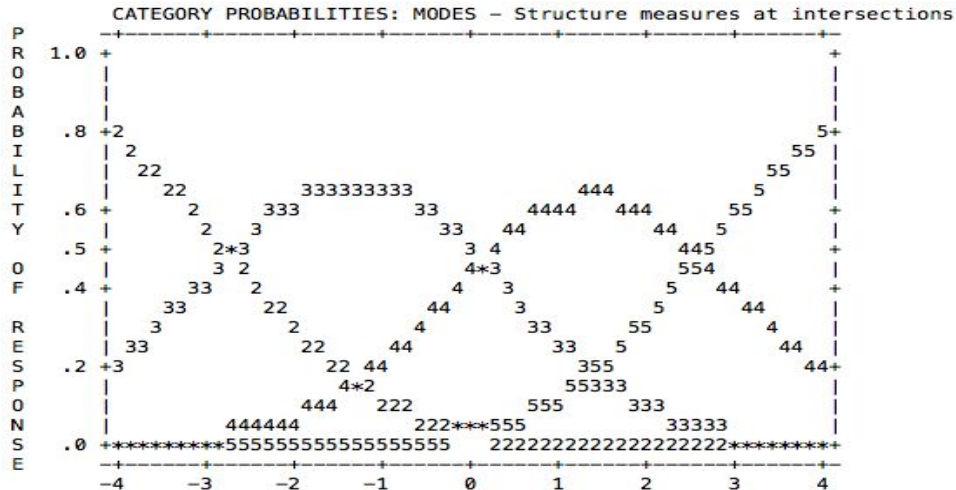
No-Video Rating Scale Diagnosis

Not
Great Diff
Some Diff
Easily
Quite Easily

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	OBSVD %	AVRGE	SAMPLE EXPECT	INFIIT MNSQ	OUTFIIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
2	2	22	5	-2.14	-1.92	.80	.82	NONE	(-3.87)	2
3	3	105	25	-.30	-.33	1.05	1.08	-2.72	-1.32	3
4	4	161	38	1.44	1.40	.88	.93	.12	1.38	4
5	5	131	31	3.23	3.28	1.11	1.12	2.60	(3.76)	5
MISSING		204	33	1.54						

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.



Relative frequency
categories 2 thru 5 had at
least 10 responses BUT
2 only accounted for 5% of
the observed scores.

Average measures and
thresholds increased
monotonically.

Spacing between thresholds
evenly distributed.

Outfit mean squares did not
exceed 2.0

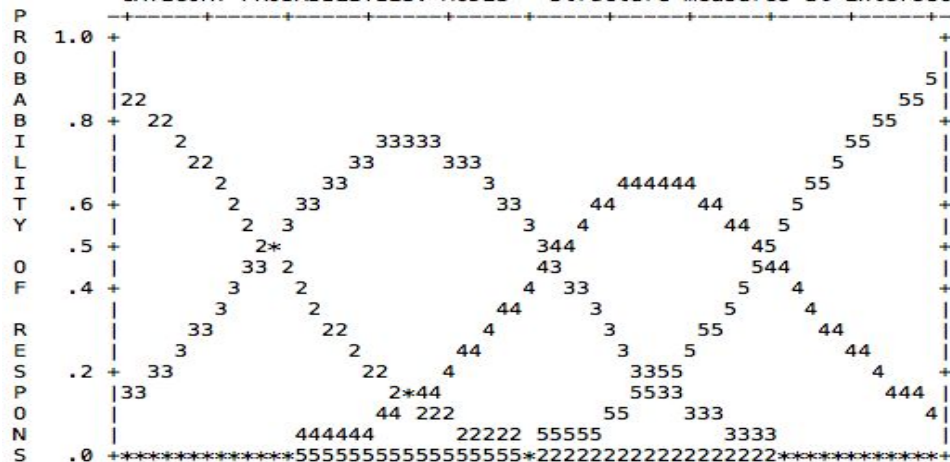
Video Rating Scale Diagnosis

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY	OBSERVED	OBSVD	SAMPLE	INFINIT	OUTFIT	ANDRICH	CATEGORY		
LABEL	SCORE	COUNT	%	AVRGE	EXPECT	MNSQ	MNSQ	THRESHOLD	MEASURE
2	2	18	6	-2.30	-2.47	1.11	1.11	NONE	(-4.35)
3	3	96	30	-.57	-.55	1.07	1.07	-3.23	-1.49
4	4	122	38	1.50	1.56	.92	1.09	.28	1.62
5	5	82	26	3.64	3.57	.90	.93	2.95	(4.10)
MISSING		165	34	.97					

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.

CATEGORY PROBABILITIES: MODES - Structure measures at intersections







relative frequency
categories 2 thru 5 had at
least 10 responses BUT
only accounted for 6% of
the observed scores.

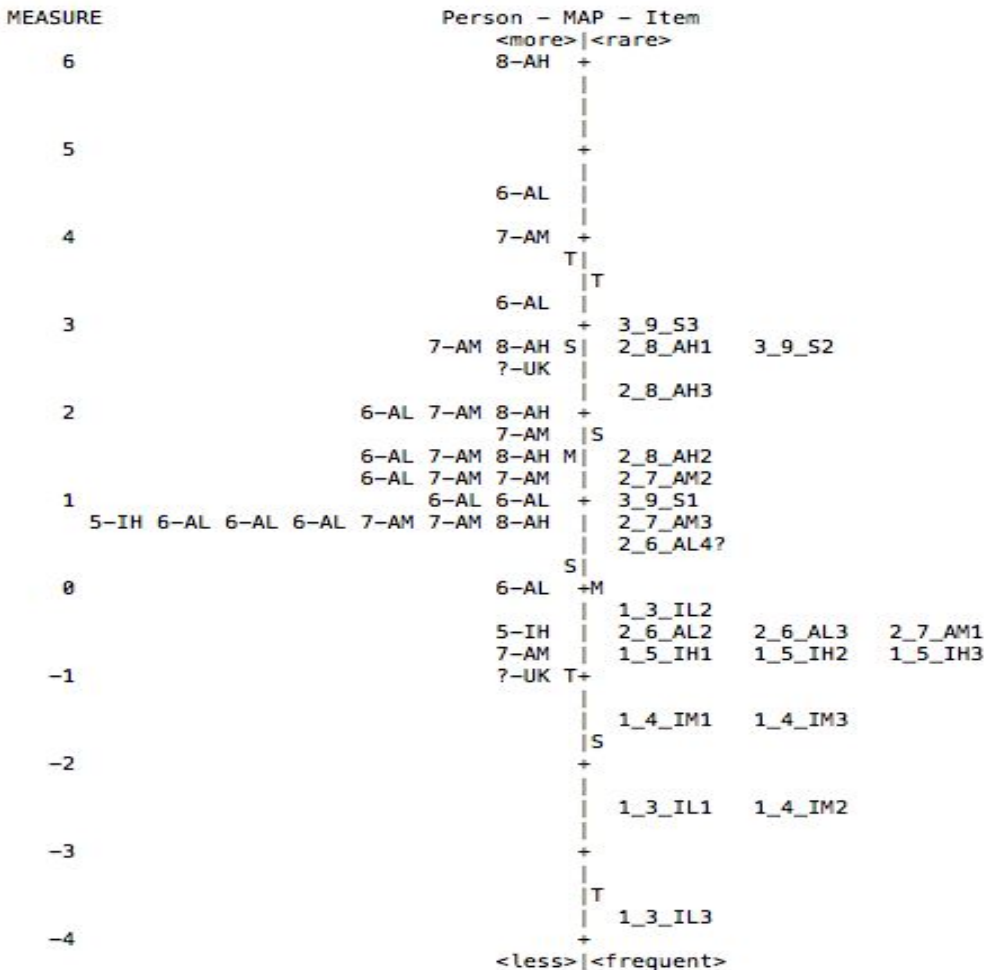
average measures and
thresholds increased
monotonically.

spacing between thresholds
evenly distributed.

Outfit mean squares did not
exceed 2.0

RQ: To what extent are a No-Video and Video can-do self-assessment survey comparable in terms of...

Score Card: No-Video vs. Video Rasch Analysis	No-Video	Operator	Video
Rating Scales?	 	=	 



No-Video Reliability

Person reliability
separation = .83

Cronbach Alpha = .69
(approximate due to
missing data)

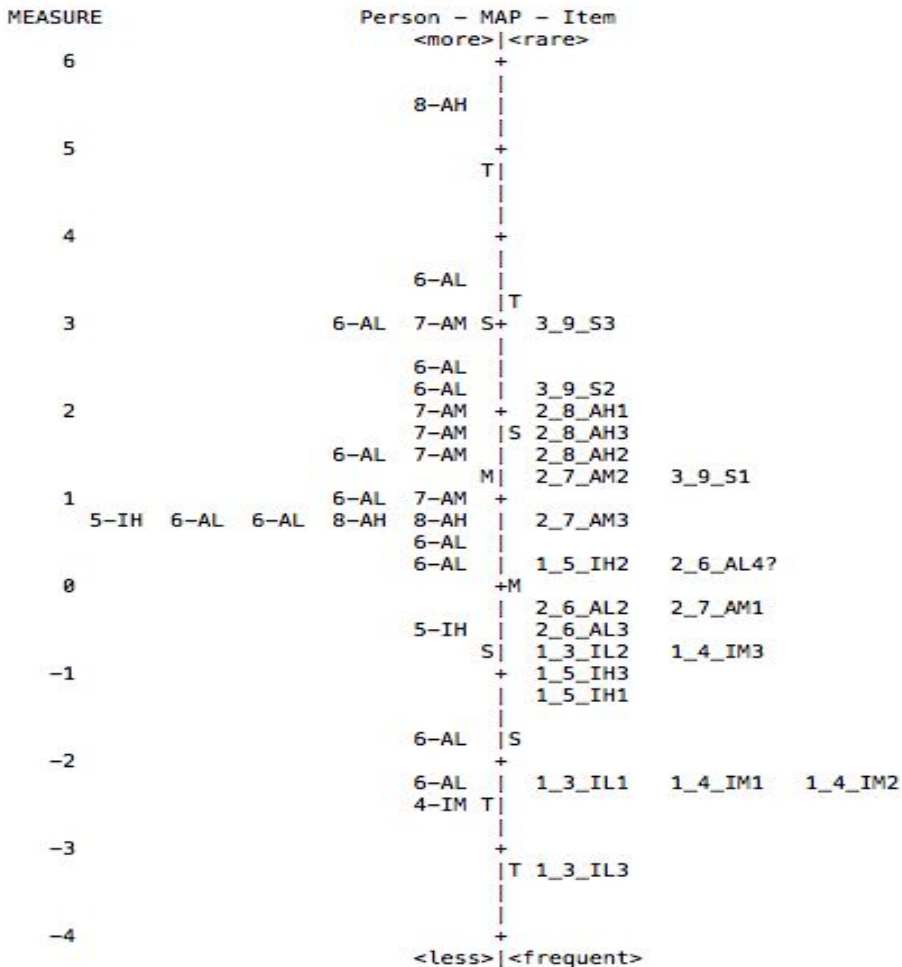
Item reliability
separation = .93

Video Reliability








Person reliability
 separation = .91

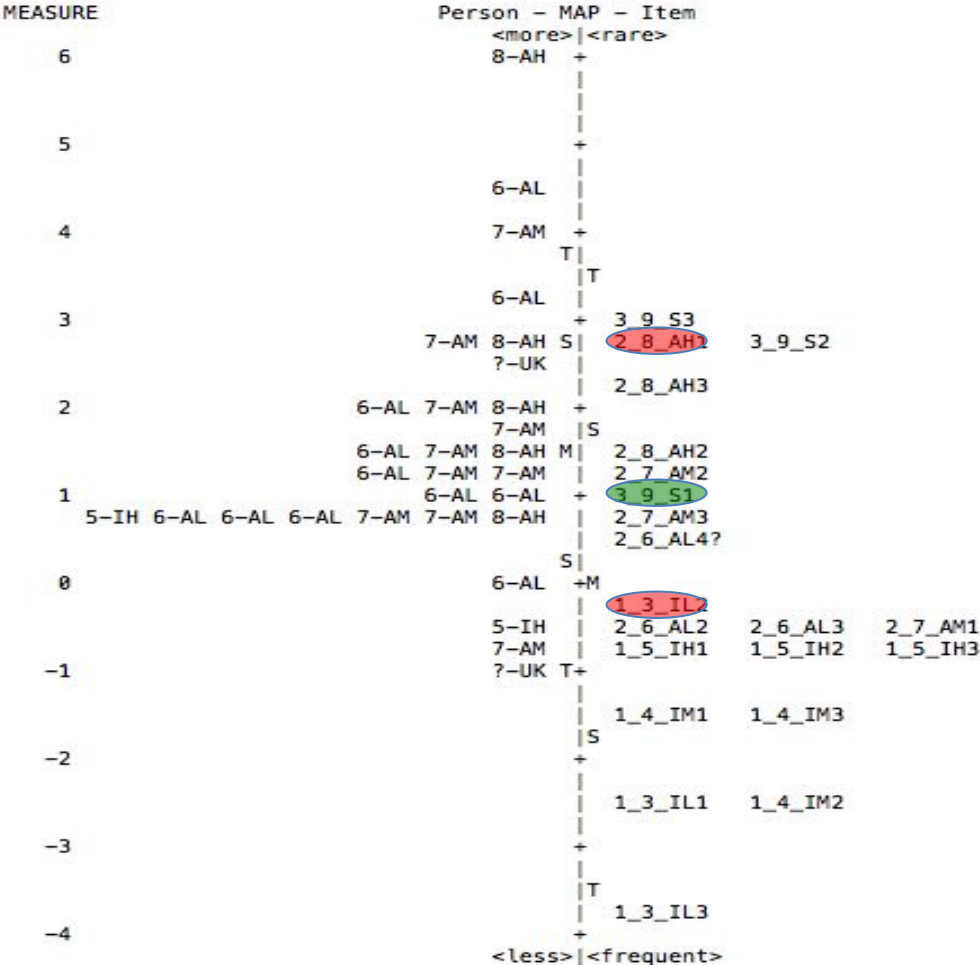
Cronbach Alpha = .85
 (approximate due to
 missing data)

Item reliability
 separation = .89



RQ: To what extent are a No-Video and Video can-do self-assessment survey comparable in terms of...

Score Card: No-Video vs. Video Rasch Analysis	No-Video	Operator	Video
Rating Scales?	 	=	 
Instrument Reliability?		<	 





No-Video Items

Intended Sublevel vs. Item Logit

Legend

Intended Item Difficulties

- 3→Superior
- 2→Advanced
- 1→Intermediate

-  Easier than intended
-  Harder than intended

Easier than Expected

Problem With Can-Do Statement Descriptor?

Superior Descriptor

I can support my opinions clearly and precisely.

EXAMPLES

- explain advantages and disadvantages of various courses of action, such as whether to rent or buy a place to live.
- participate in technical discussions in my field.
- participate in a book discussion.

Not at all



With great difficulty



With some difficulty



Easily



Quite easily



Advanced High Descriptor

I can exchange complex information about academic and professional tasks.

EXAMPLES

- exchange complex information about my academic studies, such as why I chose the field, course requirements, projects, internship opportunities, and new advances in my field.
- exchange complex information about my work responsibilities, such as the hiring process, my work schedule, the nature of my tasks, how I interface with other employees, opportunities for advancement, and new directions in my field.
- exchange complex professional or academic information to engage in collaborative work with my counterparts in different regions or countries.

Not at all



With great difficulty



With some difficulty



Easily



Quite easily



Intermediate Low Descriptor

I can ask and answer questions on factual information that is familiar to me.

EXAMPLES

- geography
- history
- art
- music
- math
- science
- language
- literature

Not at all



With great difficulty



With some difficulty



Easily



Quite easily



Harder
than
Expected

Problem With
Can-Do Statement
Descriptor?

No-Video Items

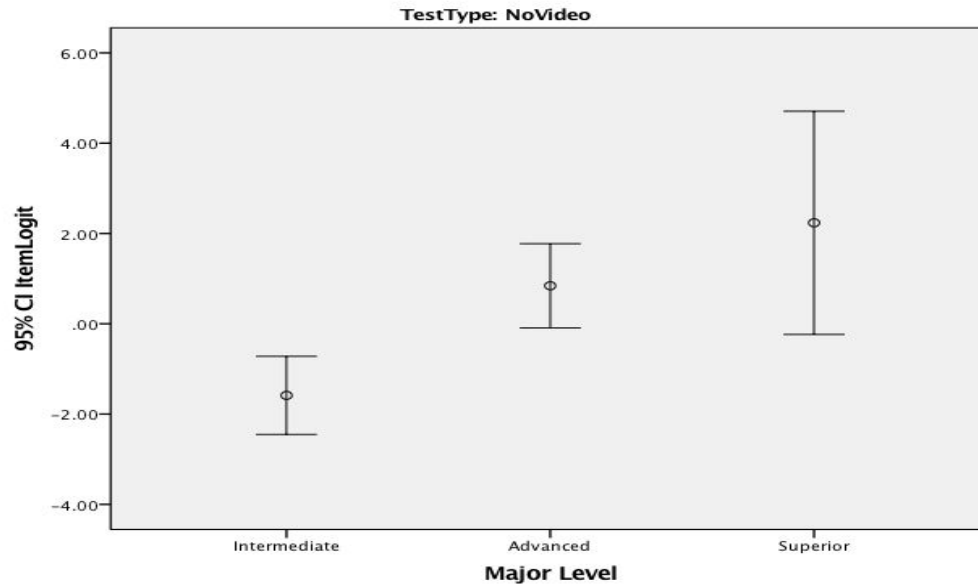
Intended Sublevel vs. Item Logit

$F(2, 18) = 16.55, p < .001$

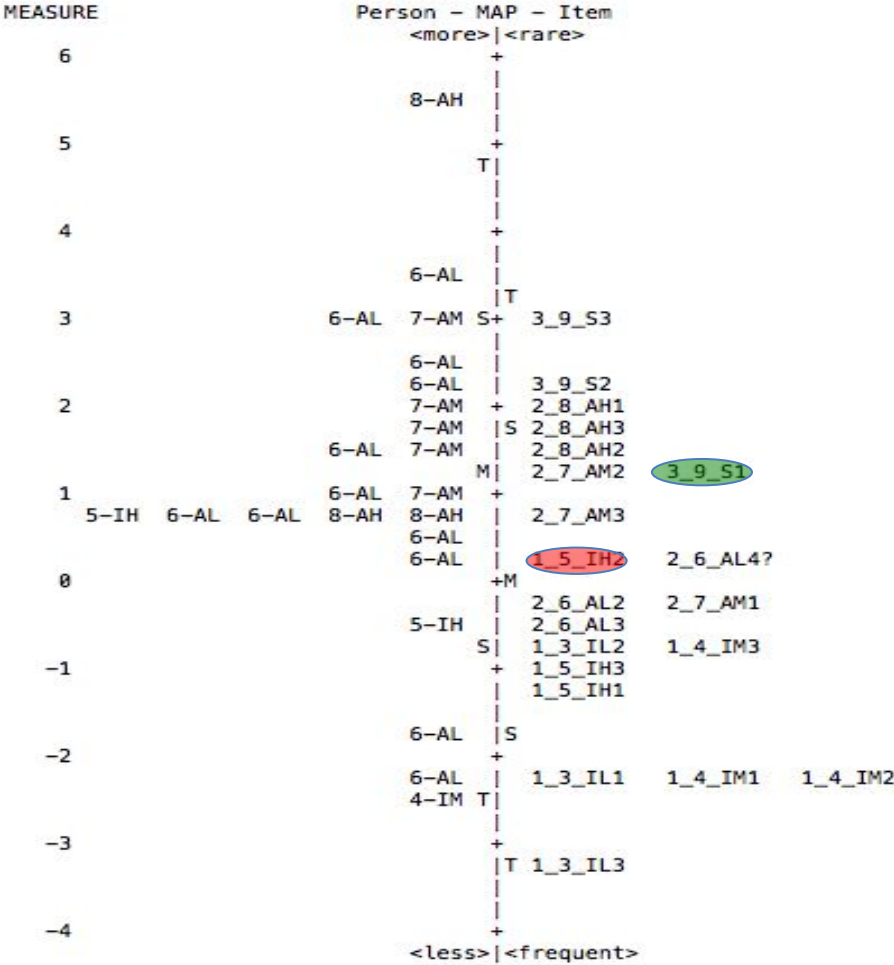
Post Hoc Tests

* Intermediate vs. Advanced
Mean difference -2.42, $p < .001$

Advanced vs. Superior
Mean difference -1.39, $p = .086$



Level	Item Count	Measure	Standard Error	Standard Deviation	Model Reliability
Interm	9	-1.59	0.38	1.06	0.77
Adv	9	0.84	0.41	1.15	0.87
Sup	3	2.24	0.58	0.81	0.83
Total	21	0.00	0.40	1.80	0.94





Video Items

Intended Major Level vs. Item Logit

Legend

Intended Item Difficulties

- 3→Superior
- 2→Advanced
- 1→Intermediate

-  Easier than intended
-  Harder than intended

Easier than Expected

Problem With Can-Do Statement Descriptor?

Superior Descriptor



I can support my opinions clearly and precisely.

EXAMPLES

- explain advantages and disadvantages of various courses of action, such as whether to rent or buy a place to live.
- participate in technical discussions in my field.
- participate in a book discussion.

Not at all



With great difficulty



With some difficulty



Easily



Quite easily



Harder
than
Expected

Problem With
Can-Do Statement
Descriptor?



I can use my language to do a task that requires multiple steps.

EXAMPLES

- give the basic rules of a game or sport and answer questions about them.
- ask for, follow, and give instructions for preparing food.
- ask for and follow directions to get from one place to another.
- tell someone how to access information online.
- explain basic rules, policies, or laws that affect us and answer questions about them.

Not at all

With great difficulty

With some difficulty

Easily

Quite easily



Video Items

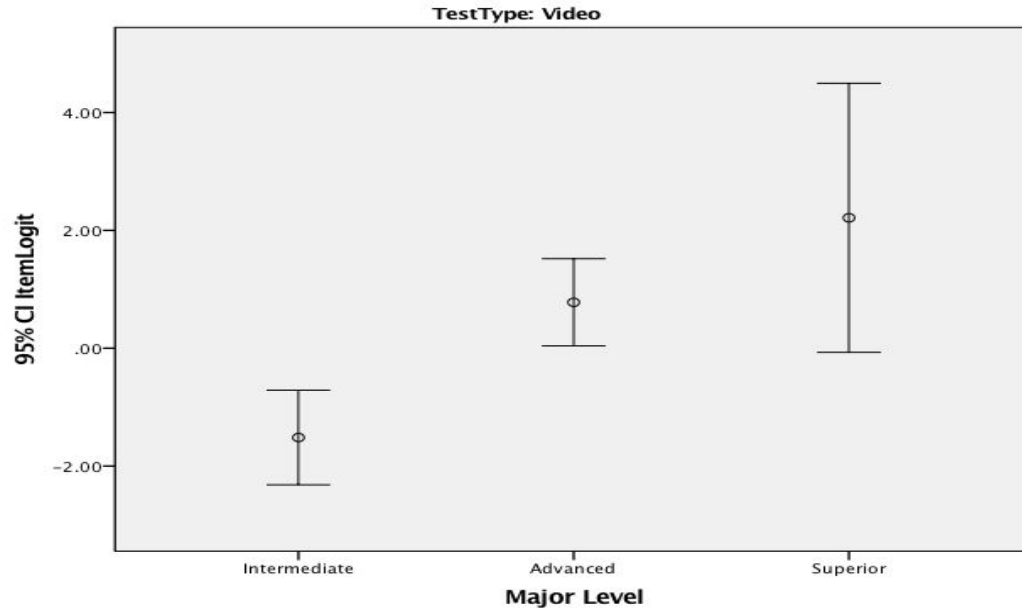
Intended Major Level vs. Item Logit

$F(2, 18) = 20.62, p < .001$

LSD











Intermediate vs. Advanced
Mean difference -2.29, $p < .001$

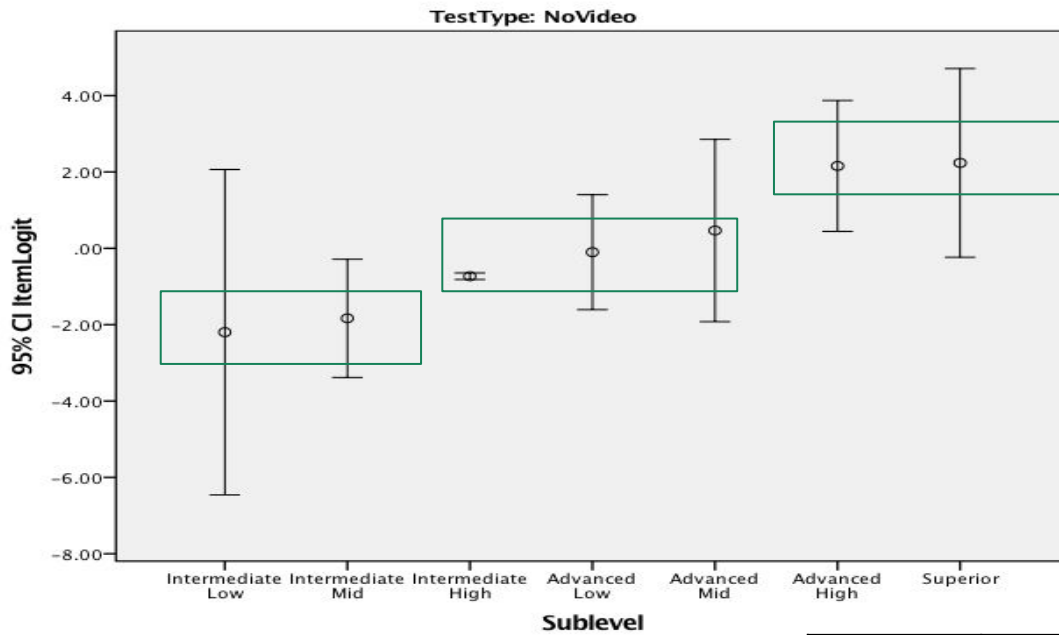
Advanced vs. Superior
Mean difference -1.43, $p = .044$



Level	Item Count	Measure	Standard Error	Standard Deviation	Model Reliability
1	9	-1.52	0.35	0.98	0.72
2	9	0.78	0.32	0.91	0.69
3	3	2.21	0.53	0.75	0.68
	21	0	0.37	1.67	0.91

RQ: To what extent are a No-Video and Video can-do self-assessment survey comparable in terms of...

Score Card: No-Video vs. Video Rasch Analysis	No-Video	Operator	Video
Rating Scales?	 	=	 
Instrument Reliability		<	 
Intended vs. Actual Level Difficulty? (Intended ACTFL level and		<	 
Ability? (OPI level and person logit)			

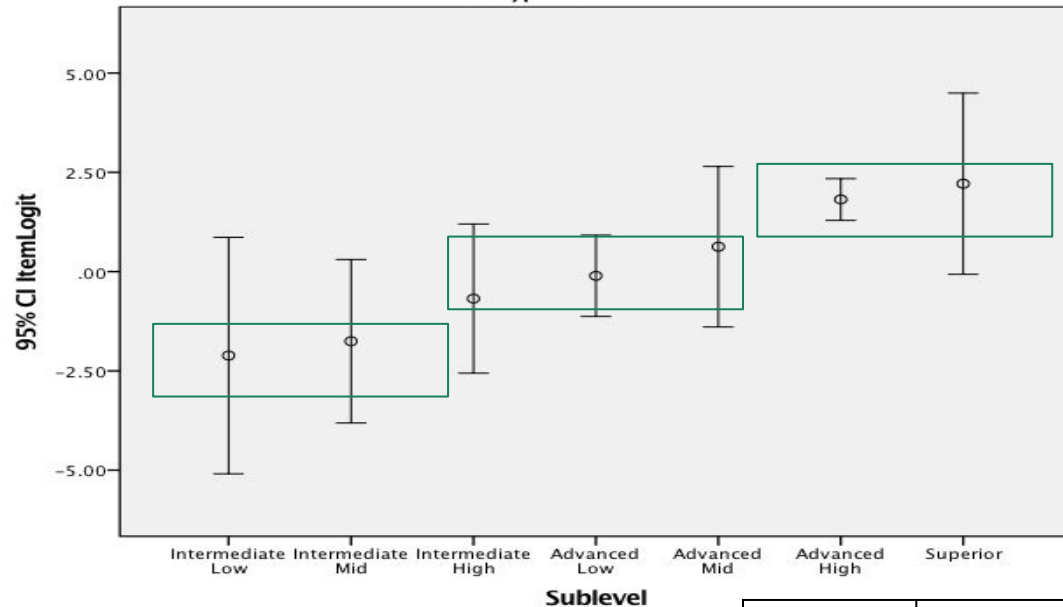


No-Video Items Intended Sublevel vs. Item Logit

$F(6, 14) = 10.69, p < .001$

Level	Item Count	Measure	Standard Error	Standard Deviation	Model Reliability
IL	3	-2.20	0.99	1.40	0.84
IM	3	-1.83	0.36	0.51	0.08
IH	3	-0.73	0.02	0.03	0.00
AL	3	-0.10	0.35	0.50	0.52
AM	3	0.47	0.55	0.78	0.68
AH	3	2.16	0.40	0.56	0.39
S	3	2.24	0.58	0.81	0.83
Total	21	0.00	0.40	1.80	0.94

TestType: Video











Video Items

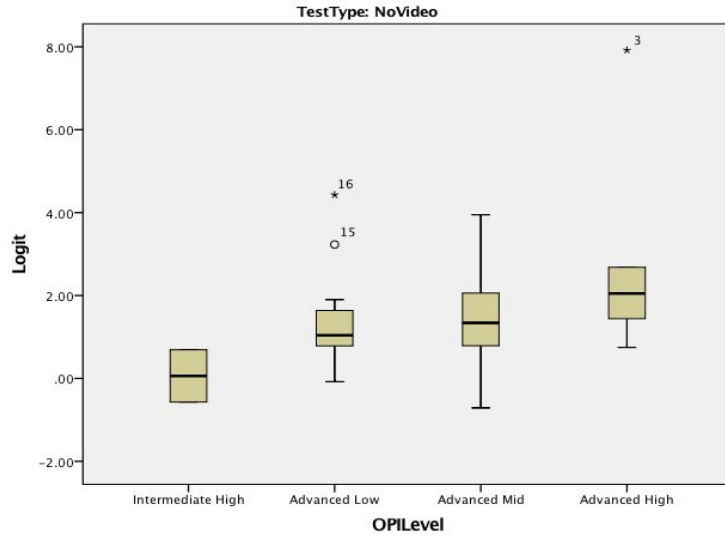
Intended Sublevel vs. Item Logit

$F(6, 14) = 13.16, p < .001$

Level	Item Count	Measure	Standard Error	Standard Deviation	Model Reliability
IL	3	-2.11	0.98	-2.28	0.75
IM	3	-1.75	0.68	-2.17	0.36
IH	3	-0.68	0.61	-1.05	0.26
AL	3	-0.11	0.33	-0.26	0
AM	3	0.62	0.66	0.8	0.37
AH	3	1.82	0.17	1.85	0
S	3	2.21	0.75	2.36	0.68
Total	21	0	1.67	-0.26	0.91

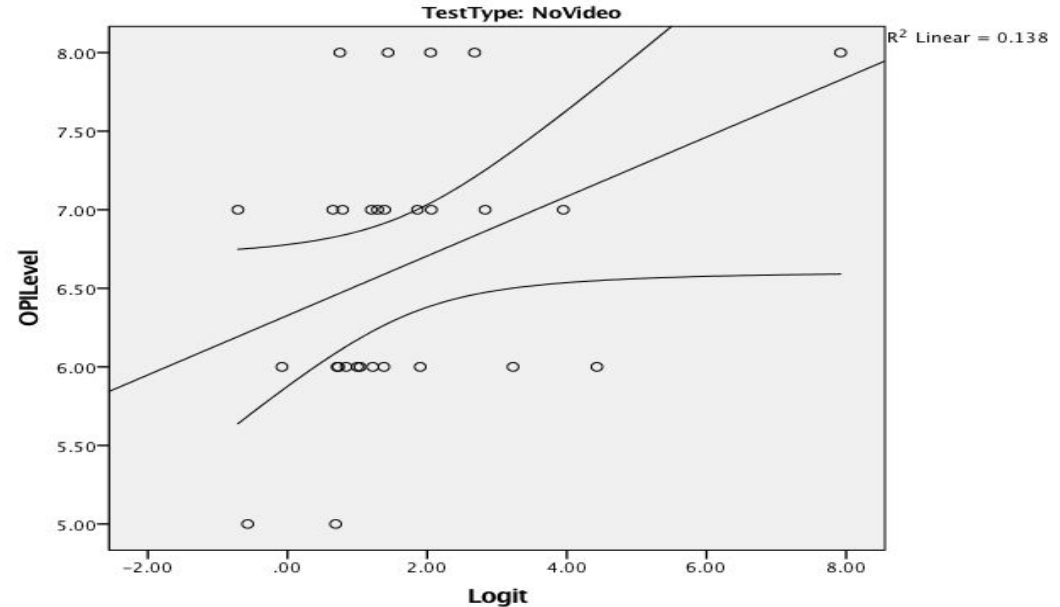
RQ: To what extent are a No-Video and Video can-do self-assessment survey comparable in terms of...

Score Card: No-Video vs. Video Rasch Analysis	No-Video	Operator	Video
Rating Scales?		=	
Instrument Reliability?		<	
Intended vs. Actual Level Difficulty? (Intended ACTFL level and item logit)		<	
Intended Sublevel vs. Item Logit		=	



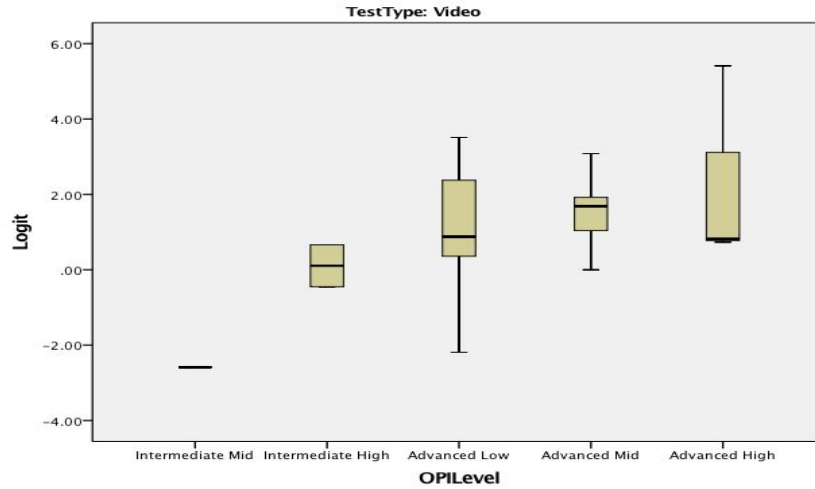
Spearman's Rho = .38,
p = .046

No Video—People OPI Score by Person Logit

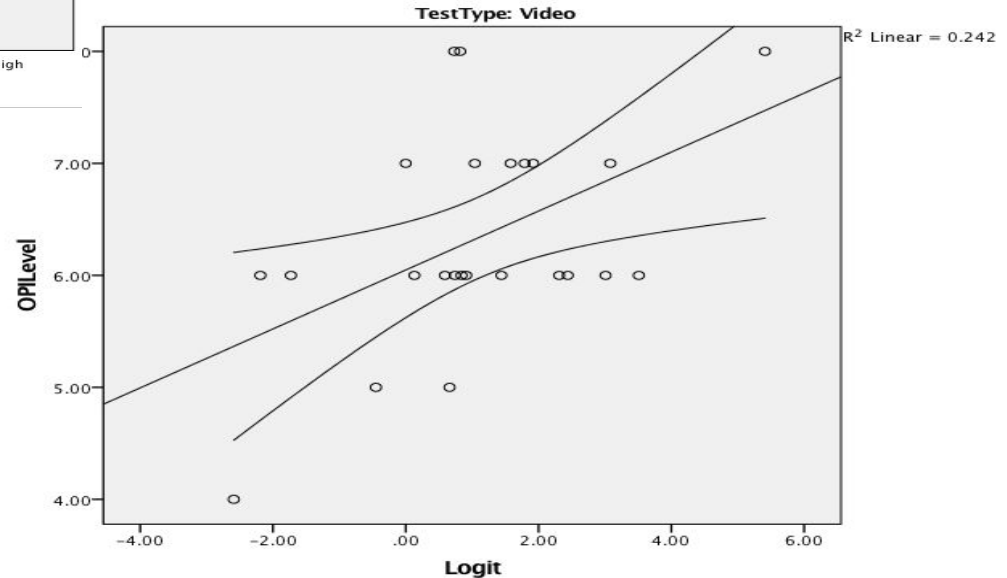


Video—People











OPI Score by Person Logit



Spearman's Rho = .40,
p = .05



RQ: To what extent are a No-Video and Video can-do self-assessment survey comparable in terms of...

Score Card: No-Video vs. Video Rasch Analysis	No-Video	Operator	Video
Rating Scale Diagnosis		=	
Instrument Reliability		<	
Items—Intended Major Level vs. Item Logit		<	
Items—Intended Sublevel vs. Item Logit		=	
Persons— OPI Level vs. Person Logit		=	

Conclusion

- Videos
 - Increased reliability slightly but no guarantee that participants watched them fully
 - Took longer to respond to
 - Slightly lower response rate
 - Language specific—you need videos in each language you want to use it in
 - Getting the videos can be difficult
- Do the ACTFL descriptors need revision?
- What are the effects of having the prompts verbatim from the descriptors?
 - Game the system?
 - Impact of topic?

Study 3

Speaking & Writing (ESL)

Verbatim Can-Do vs. Tailored?

Maria Summers, Troy Cox, & Dan Dewey

Procedure

- Administer Self Assessment Instrument (Writing and Speaking)
- Administer Placement Test Battery
- Use Rasch measurement to analyze reliability of Instruments (Speaking and Writing)
 - Category Diagnosis
 - Rasch Person Separation Measure
 - Alignment of Intended Item Difficulty with actual Item Difficulty
- Correlate Self Assessment Measures with Placement Test Results



Family



Work



Education

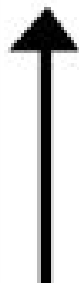


Technology



Food

Superior



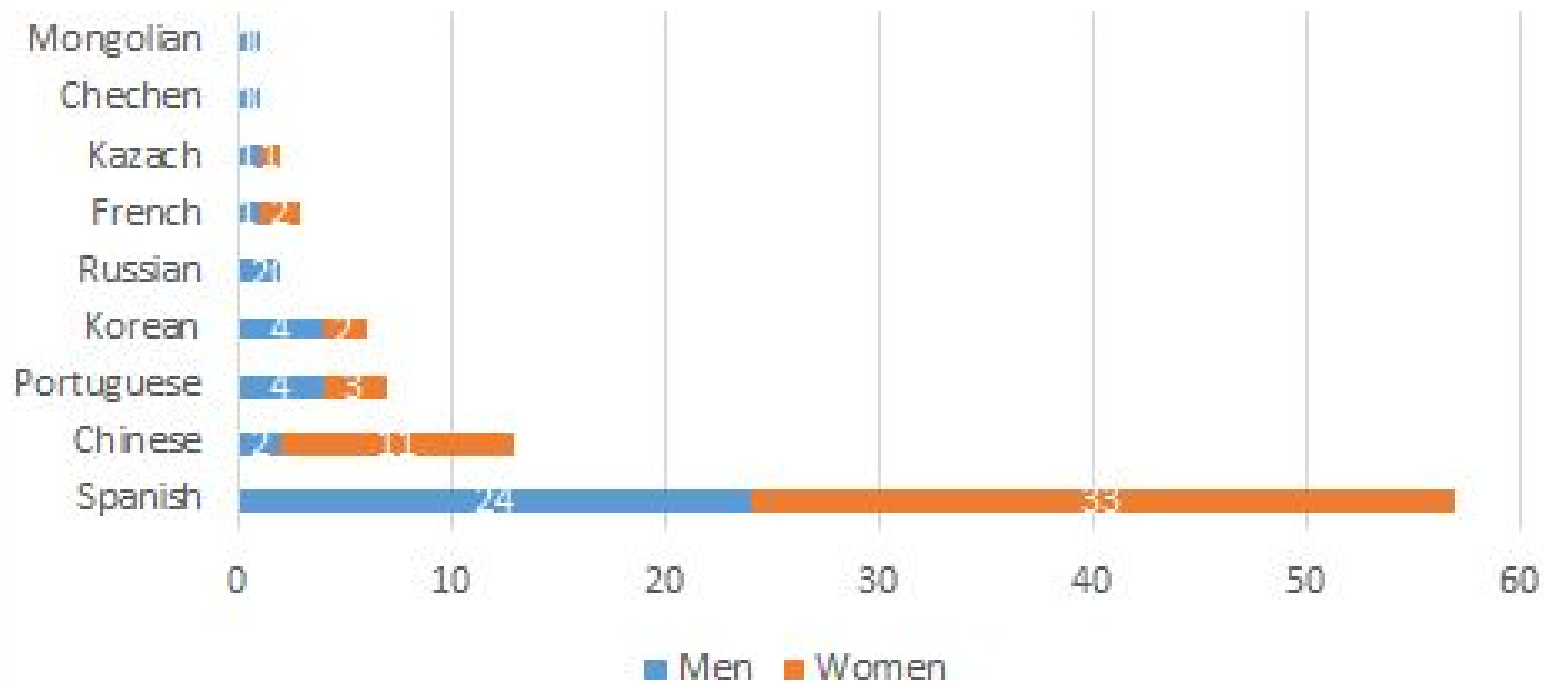
Novice

How confident are you that you could do the following tasks about family without time to prepare or reference tools (such as a dictionary)?

	Strongly Agree 	Somewhat Agree 	Neither Agree or Disagree 	Somewhat Disagree 	Strongly Disagree 
I can name the members of my family.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I can describe what my family members look like.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I can describe my family's hobbies.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I can have a conversation with someone about what my family members do for employment and discover (learn) that same information from the other person.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

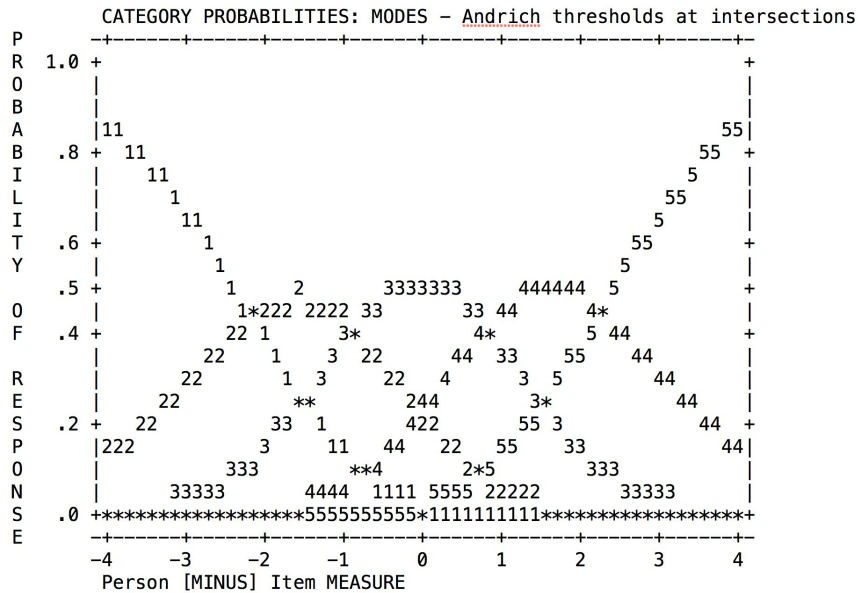


Participants

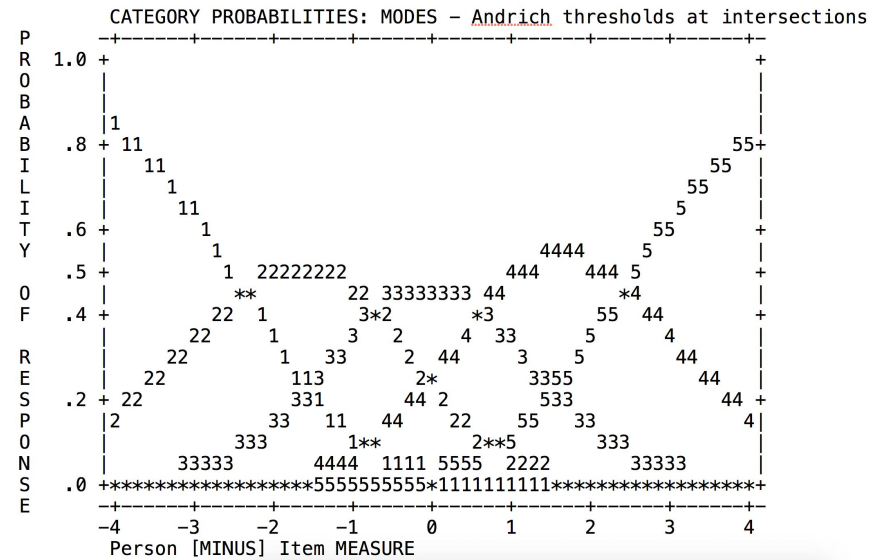


FINDINGS

Rating Scale Diagnostic—How well did the five category scale work?



Speaking Scale



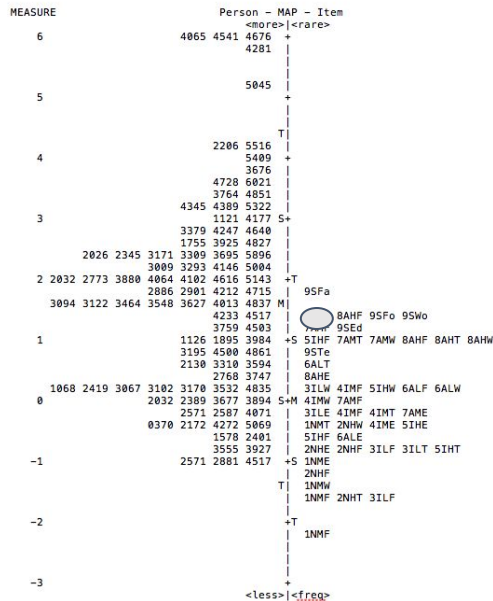
Writing Scale

1-Strongly Disagree
5-Strongly Agree

FINDINGS

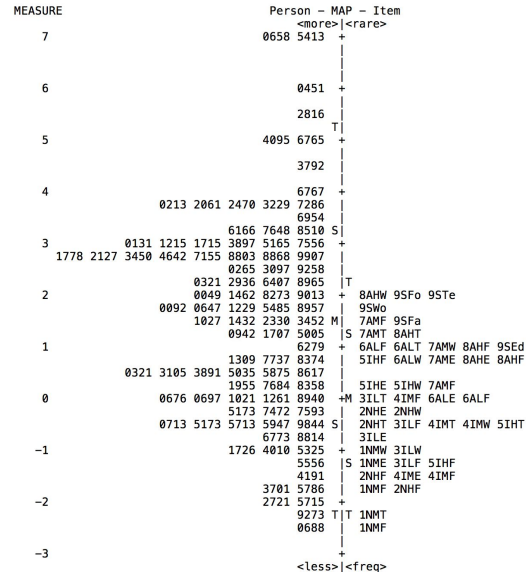
Vertical Scale Map-Speaking

TABLE 1.0 MariaSpeakSelfAssess Z00620W5.TXT Oct 21 2016 12:27
INPUT: 92 Person 45 Item REPORTED: 92 Person 45 Item 5 CATS WINSTEPS 3.92.1



Person Reliability = .91

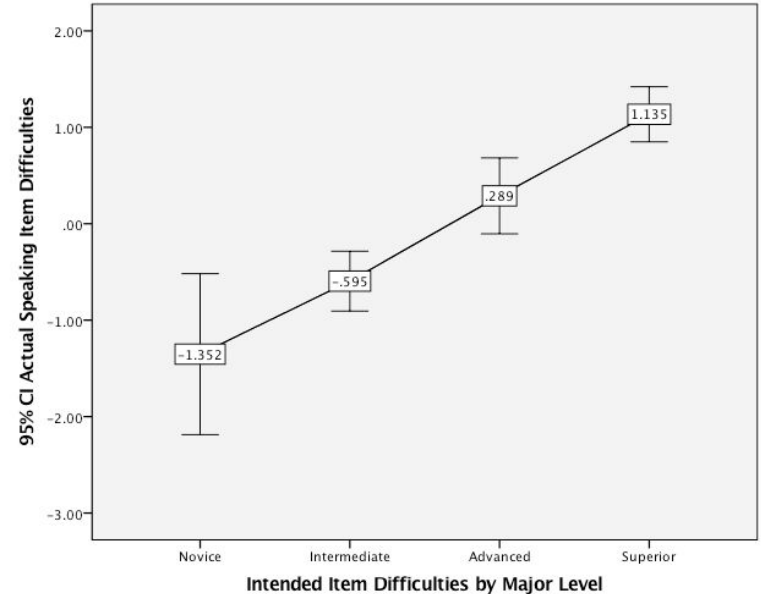
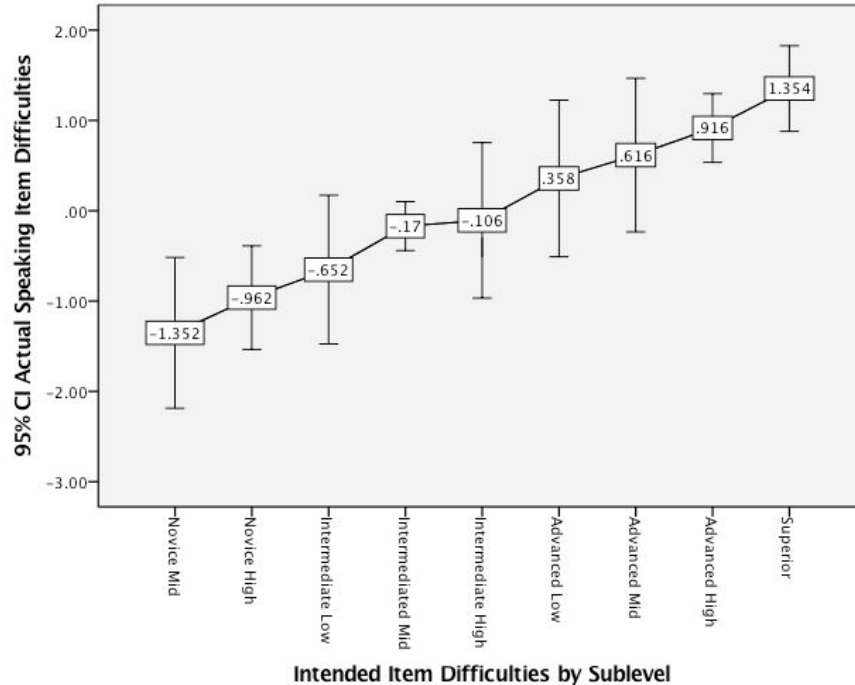
Vertical Scale Map-Writing



Person Reliability = .95

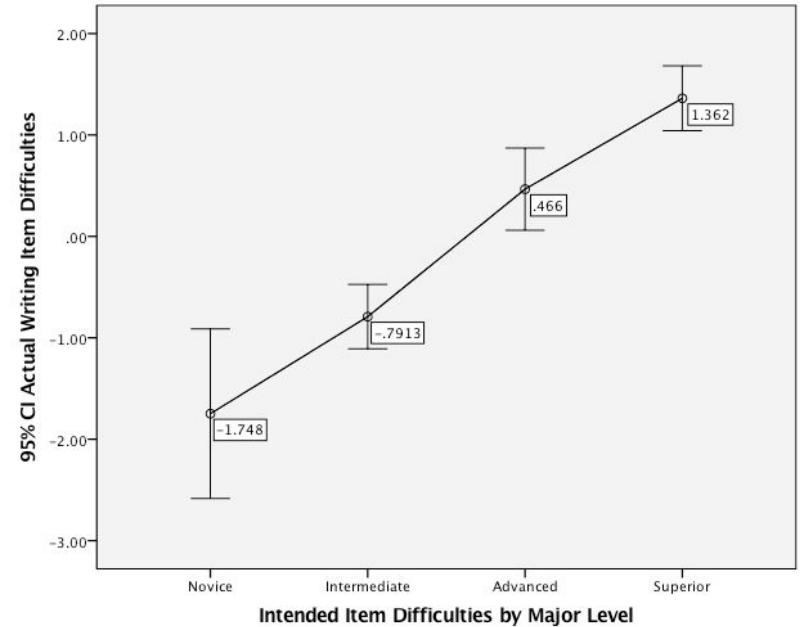
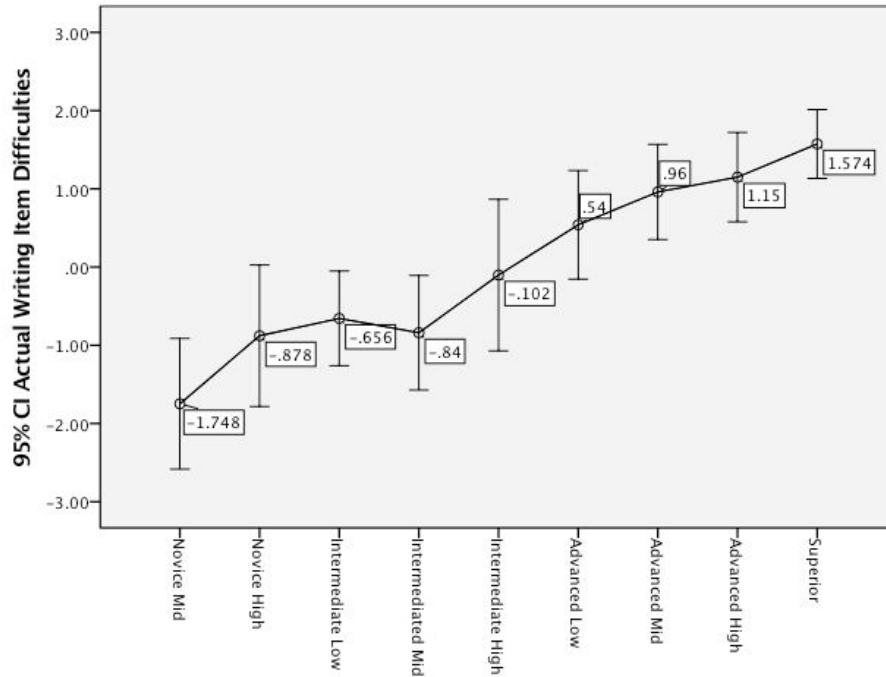
RESULTS—Speaking

95%CI Means of Intended Item Difficulties with Actual Item Difficulties



RESULTS—Writing

95%CI Means of Intended Item Difficulties with Actual Item Difficulties



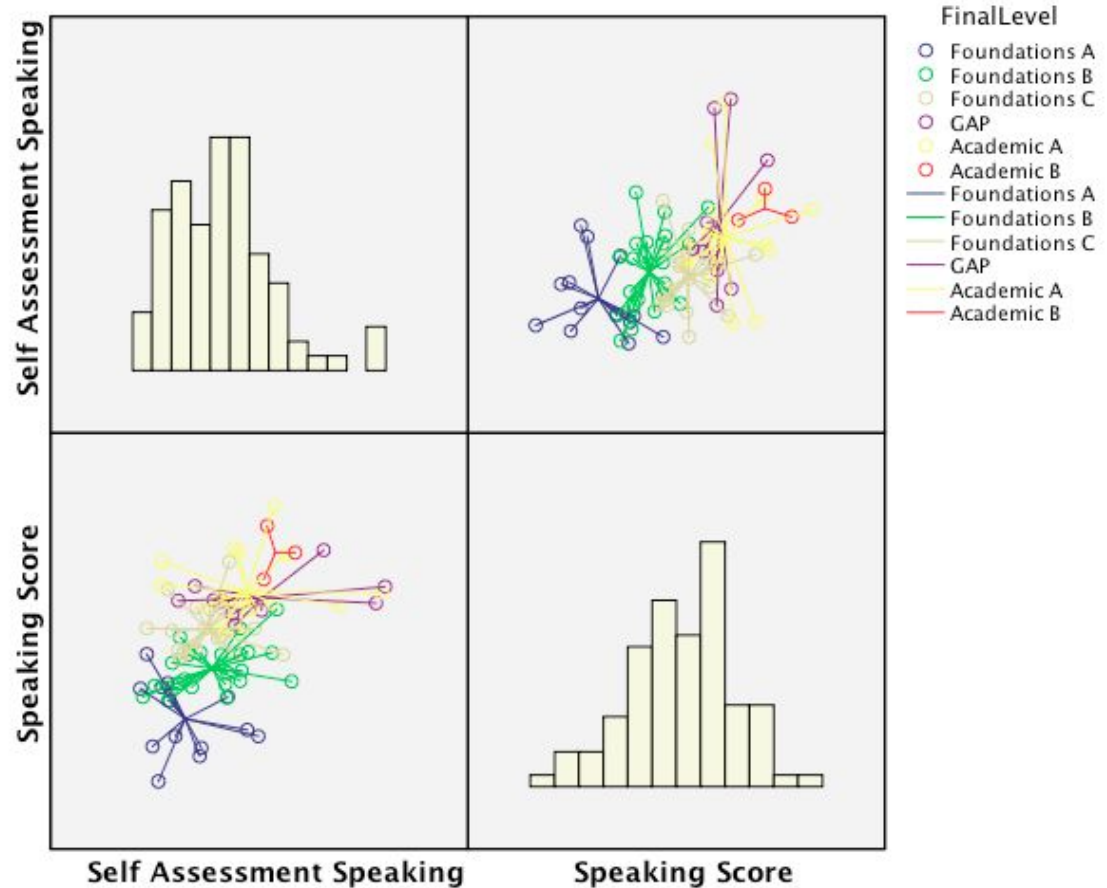
Results Speaking

Regression

$R = .44$

Adj. R Sq = .18

18% of the variance in scores can be predicted by self-perception of speaking. How well did the Speaking Self-Assessments predict actual speaking scores? ability.



Results Writing

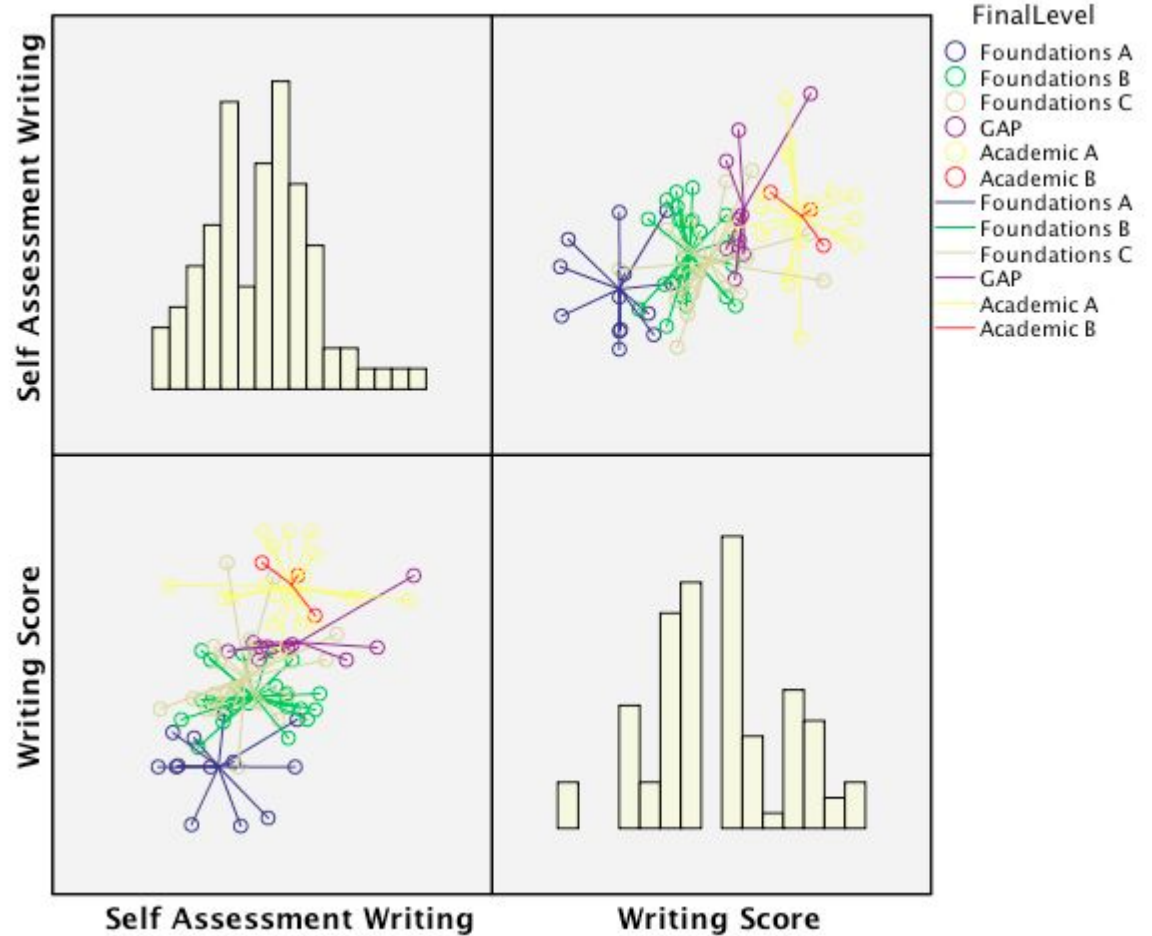
How well did the Writing
Self-Assessment predict
actual writing scores?

Regression

$R = .45$

Adj. $R^2 = .19$

19% of the variance in
scores can be predicted
by self-perception of
writing ability.



CONCLUSION

- The overall instrument is reliable. It can help learners gain an awareness of their perceived ability.
 - Items at the sublevel are not statistically different than the adjacent sublevel. Grouping items by major level criteria, however, does result in items that are statistically different from each other. Perhaps the CAN-DO statements ought reflect the major level, and let the level of performance of the student indicate the sublevel.
- Test-takers with no training in self-assessment are not very accurate in assessing their own language ability.

Study 4

Reading (English —> Russian)

Confidence by Question Language

Jeremy Evans, Troy Cox, Jennifer Bown, & Teresa Bell

Participants

	Group 1	Group 2	Total
n	34	30	64
Gender			
<i>female</i>	10	3	13
<i>male</i>	24	27	51
Mean age	22.8	21.7	22.3

Instrument-Question 1 (Russian)

Question 1 of 20

В г. Старый Оскол Белгородской области открылся Всероссийский театральный фестиваль “Золотая маска”. В ближайшие 5 дней, на сцене местного театра представят свои новые работы московская “Табакерка”, Театр марионеток из Санкт-Петербурга, Воронежский камерный театр.

“Московский ангажемент”, а также пройдет спектакль звезды последних четырех столичных театральных сезонов Григория Гришковца.

Next Question

Какое событие приближается?

A. Популярный актер выступит на сцене в последний раз .

B. Гастролирующие театральные труппы дадут представления в местных театрах.

C. В Старом Осколе пройдет Фестиваль марионеток.

D. В последний раз будет показан мюзикл “Северо-восток”.

E. Я не знаю

My response:

How confident are you in your answer choice?

very unconfident

unconfident

somewhat unconfident

somewhat confident

confident

very confident

50

Indicate your level of anxiety while answering this question.

very low

low

somewhat low

somewhat high

high

very high

50

Instrument-Question 1 (English)

Russian Research Module

Question 1 of 20

Next Question

В г. Старый Оскол Белгородской области открылся Всероссийский театральный фестиваль “Золотая маска”. В ближайшие 5 дней, на сцене местного театра представят свои новые работы московская “Табакерка”, Театр марионеток из Санкт-Петербурга, Воронежский камерный театр.

“Московский ангажемент”, а также пройдет спектакль звезды последних четырех столичных театральных сезонов Григория Гришковца.

What event is coming up?

A. The musical “North East” will end its long run.
B. A marionette festival will be in Old Oskol.
C. A popular actor will give his final performance.
D. Visiting theater groups will perform locally.
E. I don’t know

My response:

How confident are you in your answer choice?

very unconfident unconfident somewhat unconfident somewhat confident confident very confident

50

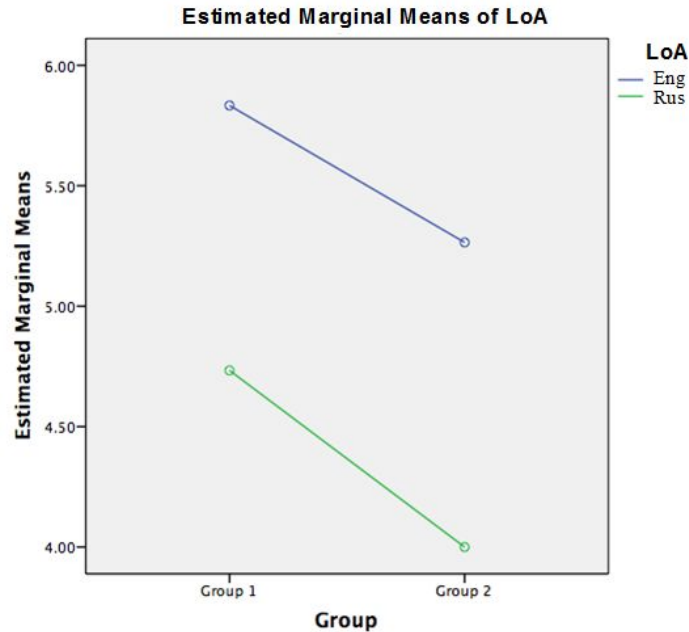
Indicate your level of anxiety while answering this question.

very low low somewhat low somewhat high high very high

50

Question Language	Statistic	Group A	Group B	Total
English	N	30	34	64
	Mean	5.83	5.26	5.55
	SD	2.41	2.16	.27
	95%CI	[4.95, 6.71]	[4.52, 6]	[4.99, 6.12]
Russian	N	30	34	64
	Mean	4.73	4.00	4.37
	SD	1.84	2.04	.24
	95%CI	[4.05, 5.41]	[3.30, 4.70]	[3.88, 4.86]

Mixed Method Repeated Measures ANOVA



- Dependent Variable: Test Score
- Between Subjects Variable: Group (A & B)
- Within Subjects Variable: Language (English & Russian)
- [$F(1, 62) = 21.47, p < .001, \text{partial } \eta^2 = .26$]

What effect does **question language** have on reading proficiency exam scores?

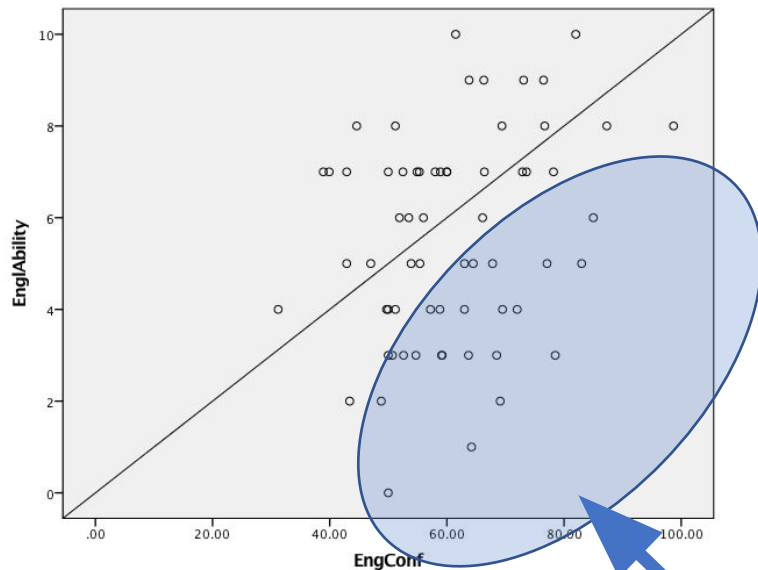
- English questions resulted in scores that were just under 12% higher than the Russian questions.

What's the relationship between confidence and how they scored?

Examinees were more accurate in self-assessment when the Q's were in

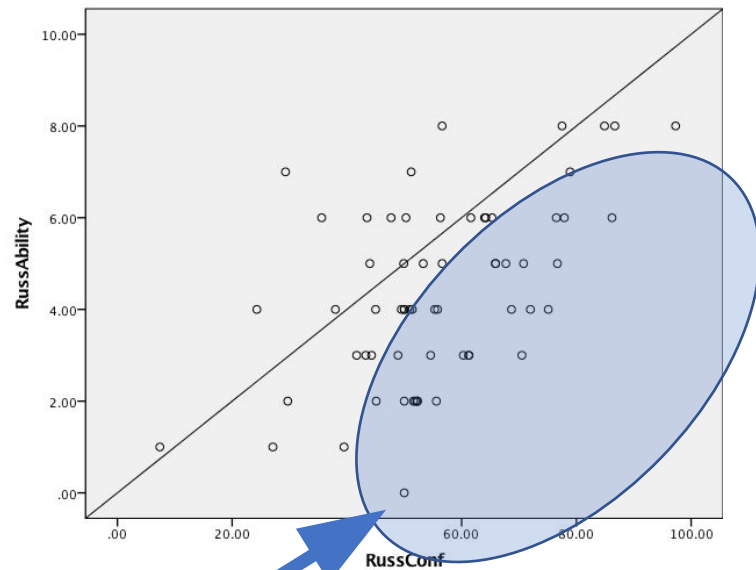
Russian
QL English

Pearson's $r = .275$



QL Russian

Pearson's $r = .533$



Overconfident

Study 5-In Progress Reading (ESL with different L1's) Confidence by Question

Jodi Peterson and Troy Cox

Instrument

Question 1 of 30 Submit Answer

A newspaper ad:

THE SPAGHETTI WAREHOUSE * EXPERIENCED FOOD SERVERS & HOSTESSES

Spaghetti Warehouse is now hiring. Full and part-time. Flexible with school schedules. Great environment, Excellent Pay & Benefits. Apply at: 1226 E. Houston St between 2-4 M-TH

The advertiser

- A. organizes outdoor activities.
- B. wants to rent out a warehouse.
- C. gives students extra training.
- D. offers jobs in a restaurant.

My response:

Please use the scrollbar to view all text as needed.

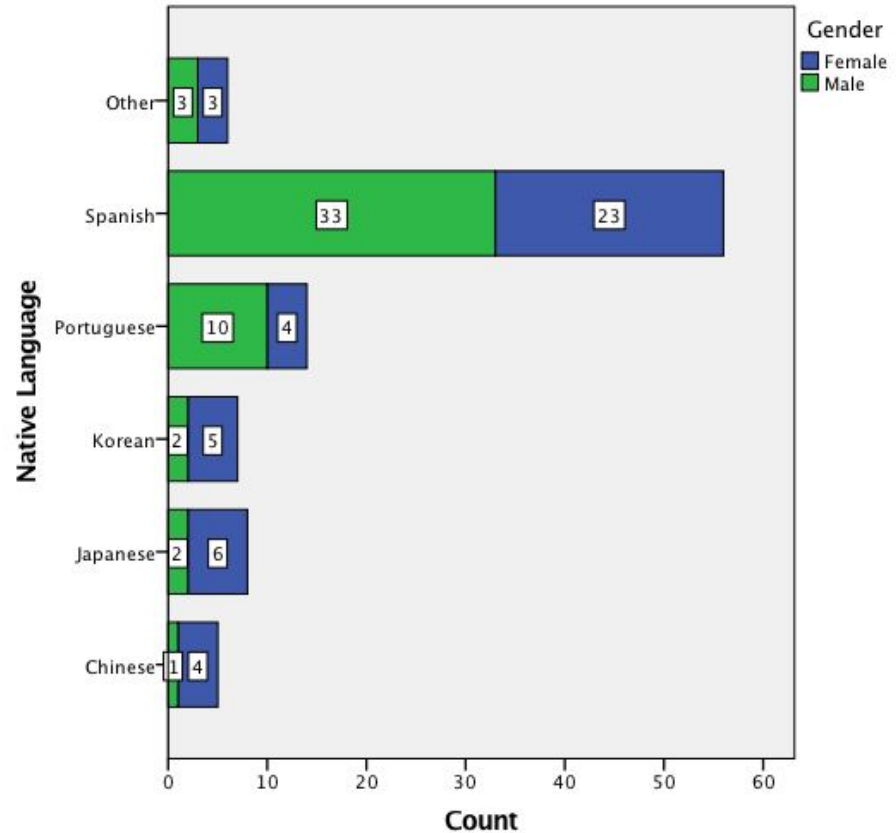
How confident are you in your answer choice?

very unconfident unconfident somewhat unconfident somewhat confident confident very confident

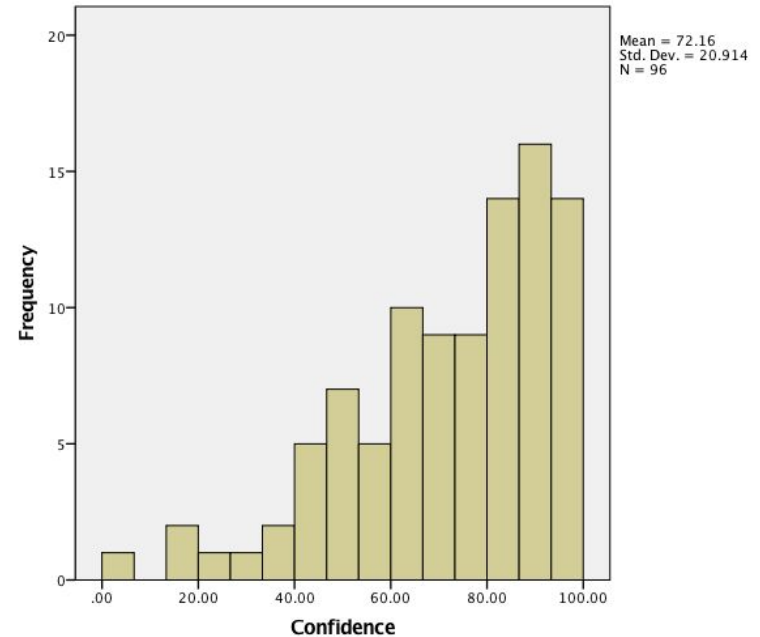
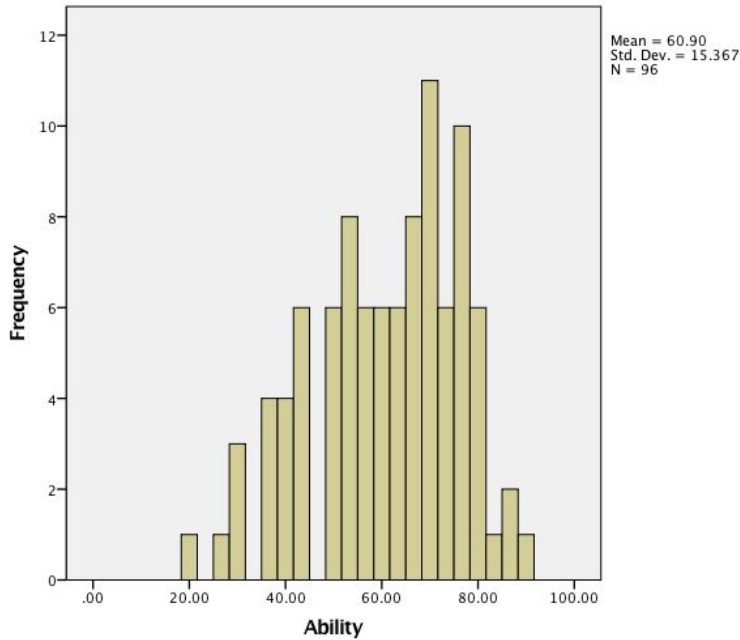
50

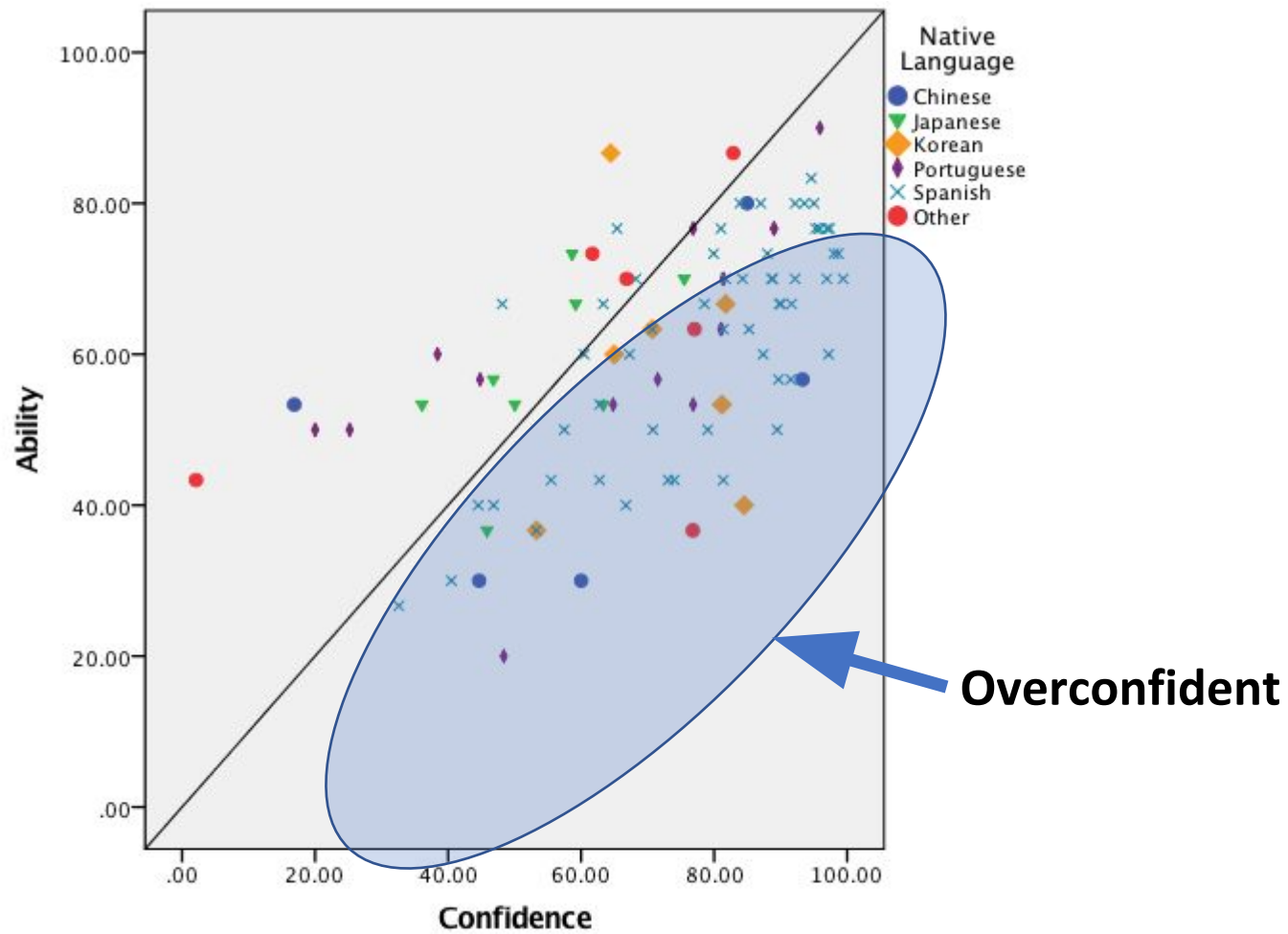
Participants

- New students (n=96) admitted to the IEP with age of the participants ranging from 17 to 63 years old (M = 26.4, SD = 9.3)



Histogram of Ability & Confidence



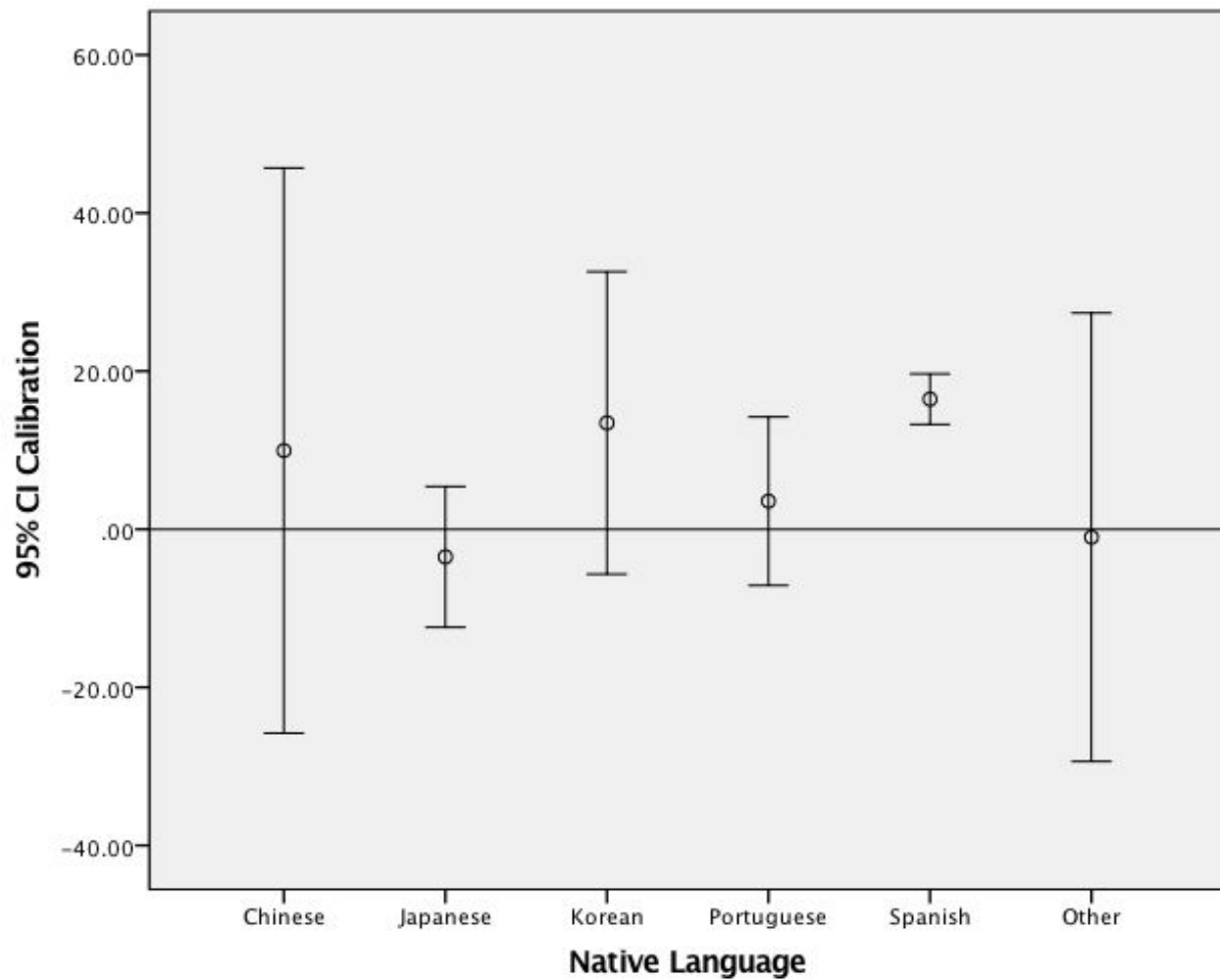


Calibration = Confidence - Ability

A Value of 0 = Perfect calibration

Positive Values = Overconfident

Negative Values = Underconfident



Discussion

- With Rasch modeling when an item and a person are at the same logit value, the probability of a correct answer/possessing the attribute is 50/50.
- How do the logits translate to proficiency ratings?
 - Study 1—> 64% rated themselves at Superior or Higher BEFORE
94% rated themselves at Superior or Higher AFTER
 - Study 2 —> 25% rated themselves as Superior with Video
23% rated themselves at Superior without Video
 - Study 3 —> 61% rated themselves as Superior in Speaking
48% rated themselves at Superior in Writing
- Even with question confidence, learners in Study 4 & 5 were overconfident.

Conclusions



I'm starting with the man in the mirror
I'm asking him to change his ways
And no message could have been any clearer
If you wanna make the world a better place
Take a look at yourself, and then make a change
-MJ