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


## A Table of Five Studies

<table>
<thead>
<tr>
<th></th>
<th>Native Language</th>
<th>Second Language</th>
<th>Instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaking</td>
<td>English</td>
<td>Russian</td>
<td>Self-Retrospective Statements based on Can-Do Subheadings</td>
</tr>
<tr>
<td>Speaking &amp; Writing</td>
<td>Spanish, Chinese, Korean, Portuguese, Russian, French &amp; Others</td>
<td>English</td>
<td>Survey questions based on the Can-Do Statements</td>
</tr>
<tr>
<td>Reading</td>
<td>English</td>
<td>Russian</td>
<td>Confidence slider after each question. Passages in Russian with questions in both English and Russian</td>
</tr>
<tr>
<td></td>
<td>Spanish, Portuguese, Japanese, Korean, Chinese, &amp; Others</td>
<td>English</td>
<td>Confidence slider after each question. Passages and questions in English</td>
</tr>
</tbody>
</table>
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

confidence "high" "low" "know-nothing"

peak of "Mt. Stupid"

valley of despair

plateau of sustainability

slope of enlightenment

wisdom

"guru"

http://www.understandinginnovation.wordpress.com
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

<table>
<thead>
<tr>
<th>ACTFL Standard</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distinguished</strong></td>
<td>I can communicate reflectively on a wide range of global issues and highly abstract concepts in a culturally sophisticated manner.</td>
</tr>
<tr>
<td><strong>Superior</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Advanced High</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Advanced Mid</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Advanced Low</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Intermediate High</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Intermediate Mid</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Intermediate Low</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Novice High</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Novice Mid</strong></td>
<td>I can communicate on very familiar topics using a variety of words and phrases that I have practiced and memorized.</td>
</tr>
<tr>
<td><strong>Novice Low</strong></td>
<td>I can communicate on some very familiar topics using single words and phrases that I have practiced and memorized.</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th></th>
<th>Could not do this even with extensive preparation</th>
<th>Unsure as to whether I could or could not do this</th>
<th>Could do this with extensive preparation</th>
<th>Could do this with minimal preparation</th>
<th>Could do this without any preparation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Internship</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-Internship</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### I could discuss complex information in debates or meetings.

<table>
<thead>
<tr>
<th></th>
<th>Could not do this even with extensive preparation</th>
<th>Unsure as to whether I could or could not do this</th>
<th>Could do this with extensive preparation</th>
<th>Could do this with minimal preparation</th>
<th>Could do this without any preparation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Internship</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-Internship</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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?
<table>
<thead>
<tr>
<th>Category</th>
<th>Absolute Frequency</th>
<th>Relative Frequency</th>
<th>Average Measure</th>
<th>Outfit</th>
<th>Threshold</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Can’t Do</td>
<td>17</td>
<td>1%</td>
<td>-2.98</td>
<td>.98</td>
<td>-4.13</td>
<td>0.26</td>
</tr>
<tr>
<td>2-Unsure</td>
<td>149</td>
<td>6%</td>
<td>-.30</td>
<td>1.61</td>
<td>-4.13</td>
<td>0.26</td>
</tr>
<tr>
<td>3-W/ Ext prep</td>
<td>544</td>
<td>22%</td>
<td>.60</td>
<td>.80</td>
<td>-1.24</td>
<td>0.11</td>
</tr>
<tr>
<td>4-W/ Some prep</td>
<td>1070</td>
<td>43%</td>
<td>2.80</td>
<td>.86</td>
<td>1.13</td>
<td>0.06</td>
</tr>
<tr>
<td>5-No Prep</td>
<td>720</td>
<td>29%</td>
<td>4.91</td>
<td>1.02</td>
<td>4.24</td>
<td>0.07</td>
</tr>
</tbody>
</table>
## Vertical Scale

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

<table>
<thead>
<tr>
<th>#</th>
<th>Students</th>
<th>Items</th>
</tr>
</thead>
</table>
| 3 | THEN Person Ability Estimates  
  Mean = 3.88, SD = 2.83 | NOW Person Ability Estimates  
  Mean = 3.78, SD = 1.83 | Item Difficulty Parameter  
  Mean = 3, SD = 1.7 |
| 3 | 9.91Y | 7.91Y | 3.91F | .96 |
| 6 | 6.91Y | 5.91Y | 5.91C | .95 |
| 5 | 8.91N | 5.91M | 10.91M | 8.91N | 5.91B | .96 |
| 9 | 7.91A | 8.91A | 9.91A | 10.91A | 10.91A | .95 |
| 4 | 9.91D | 9.91E | 9.91F | 5.91L | 5.91L | 5.91T | .95 |
| 3 | 8.91K | 8.91J | 7.91M | 9.91A | 8.91F | 8.91C | 8.91M | 8.91L | .95 |
| 2 | 7.91F | 9.91G | 8.91T | 5.91L | 5.91L | 5.91T | .95 |
| 1 | 9.91W | 7.91A | 7.91B | 9.91C | 9.91D | 9.91E | 9.91F | 9.91G | 9.91H | .95 |
| 0 | 7.91W | 6.91B | 9.91C | 9.91D | 9.91E | 9.91F | 9.91G | 9.91H | 9.91I | .91 |

**Then-Now Self Assessment Vertical Scale Map**

<table>
<thead>
<tr>
<th>Students</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Separation Reliability</td>
<td>.96</td>
</tr>
<tr>
<td>Separation Strata</td>
<td>5.57</td>
</tr>
</tbody>
</table>
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?

- Respondents with high “X”
  - Item response indicates highest level of “X”

- Respondents with mid-range “X”
  - Item response indicates higher level of “X”

- Respondents with low “X”
  - Item response indicates lower level of “X”

Direction of increasing “X”

Direction of decreasing “X”
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?

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Spearman’s Rho</th>
<th>P (one-tailed)</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Then and Pre-OPI</td>
<td>37</td>
<td>.27</td>
<td>.06</td>
<td>Small to Medium</td>
</tr>
<tr>
<td>Now and Post-OPI</td>
<td></td>
<td>.21</td>
<td>.11</td>
<td>Small to Medium</td>
</tr>
</tbody>
</table>
What is the relationship between Then-Now score gain OPI rating gain?

<table>
<thead>
<tr>
<th>N</th>
<th>Spearman’s Rho</th>
<th>P (one-tailed)</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>ThenNow Gain and OPI Gain</td>
<td>37</td>
<td>.21</td>
<td>.10</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Not at all</th>
<th>With great difficulty</th>
<th>With some difficulty</th>
<th>Easily</th>
<th>Quite easily</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Not at all</th>
<th>With great difficulty</th>
<th>With some difficulty</th>
<th>Easily</th>
<th>Quite easily</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.
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
Filtered with 1 Hour Max (7 people [6 Video, 1 No-Video] excluded.

Video Group = 22
No Video Group = 31

8 minutes
5 minutes
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...

<table>
<thead>
<tr>
<th>Score Card: No-Video vs. Video Rasch Analysis</th>
<th>No-Video</th>
<th>Operator (=, &lt; or &gt;)</th>
<th>Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating Scales?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instrument Reliability?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intended vs. Actual Level Difficulty?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Intended ACTFL level and item logit)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predictive Validity of Person Ability?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(OPI level and person logit)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

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.

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>OBSERVED</th>
<th>OBSVD SAMPLE</th>
<th>INFIT</th>
<th>OUTFIT</th>
<th>ANDRICH</th>
<th>CATEGORY MEASURE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>18</td>
<td>6</td>
<td>-2.30</td>
<td>-2.47</td>
<td>1.11 1.11</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>96</td>
<td>30</td>
<td>-1.57</td>
<td>-1.55</td>
<td>1.07 1.07</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>122</td>
<td>38</td>
<td>1.58</td>
<td>1.56</td>
<td>.92 1.09</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>82</td>
<td>26</td>
<td>3.64</td>
<td>3.57</td>
<td>.90 1.93</td>
</tr>
<tr>
<td>MISSING</td>
<td>165</td>
<td>34</td>
<td>.97</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

CATEGORY PROBABILITIES: MODES - Structure measures at intersections

Not Great Diff Some Diff Easily Quite Easily
RQ: To what extent are a No-Video and Video can-do self-assessment survey comparable in terms of…

<table>
<thead>
<tr>
<th>Score Card: No-Video vs. Video Rasch Analysis</th>
<th>No-Video</th>
<th>Operator</th>
<th>Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating Scales?</td>
<td>🌟🌟</td>
<td>=</td>
<td>🌟🌟🌟</td>
</tr>
</tbody>
</table>
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...

<table>
<thead>
<tr>
<th>Score Card: No-Video vs. Video Rasch Analysis</th>
<th>No-Video</th>
<th>Operator</th>
<th>Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating Scales?</td>
<td>🌟🌟</td>
<td>=</td>
<td>🌟🌟🌟</td>
</tr>
<tr>
<td>Instrument Reliability?</td>
<td>🌟</td>
<td>&lt;</td>
<td>🌟🌟🌟</td>
</tr>
</tbody>
</table>
### No-Video Items

**Intended Sublevel vs. Item Logit**

**Legend**
- Intended Item Difficulties
  - 3→Superior
  - 2→Advanced
  - 1→Intermediate

- Green: Easier than intended
- Red: Harder than intended
Easier than Expected

Problem With Can-Do Statement Descriptor?
Harder than Expected

Problem With Can-Do Statement Descriptor?

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.

<table>
<thead>
<tr>
<th>Not at all</th>
<th>With great difficulty</th>
<th>With some difficulty</th>
<th>Easily</th>
<th>Quite easily</th>
</tr>
</thead>
</table>

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

<table>
<thead>
<tr>
<th>Not at all</th>
<th>With great difficulty</th>
<th>With some difficulty</th>
<th>Easily</th>
<th>Quite easily</th>
</tr>
</thead>
</table>
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, \( p = .086 \)
Video Items
Intended Major Level vs. Item Logit

Legend
Intended Item Difficulties

3 → Superior
2 → Advanced
1 → Intermediate

-4
-3
-2
-1
0
1
2
3
4
5
6

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.

<table>
<thead>
<tr>
<th>Not at all</th>
<th>With great difficulty</th>
<th>With some difficulty</th>
<th>Easily</th>
<th>Quite easily</th>
</tr>
</thead>
</table>
Harder than Expected

Problem With Can-Do Statement Descriptor?

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

<table>
<thead>
<tr>
<th>Score Card: No-Video vs. Video Rasch Analysis</th>
<th>No-Video</th>
<th>Operator</th>
<th>Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating Scales?</td>
<td>🌟🌟</td>
<td>🌟🌟</td>
<td>🌟🌟🌟🌟</td>
</tr>
<tr>
<td>Instrument Reliability</td>
<td>🌟</td>
<td>&lt;</td>
<td>🌟🌟🌟🌟</td>
</tr>
<tr>
<td>Intended vs. Actual Level Difficulty? (Intended ACTFL level and Item Logit)</td>
<td>🌟</td>
<td>&lt;</td>
<td>🌟🌟🌟🌟</td>
</tr>
<tr>
<td>Predictive Validity of Person Ability? (OPI level and person logit)</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
F (6, 14) = 10.69, p < .001

<table>
<thead>
<tr>
<th>Level</th>
<th>Item Count</th>
<th>Measure</th>
<th>Standard Error</th>
<th>Standard Deviation</th>
<th>Model Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL</td>
<td>3</td>
<td>-2.20</td>
<td>0.99</td>
<td>1.40</td>
<td>0.84</td>
</tr>
<tr>
<td>IM</td>
<td>3</td>
<td>-1.83</td>
<td>0.36</td>
<td>0.51</td>
<td>0.08</td>
</tr>
<tr>
<td>IH</td>
<td>3</td>
<td>-0.73</td>
<td>0.02</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>AL</td>
<td>3</td>
<td>-0.10</td>
<td>0.35</td>
<td>0.50</td>
<td>0.52</td>
</tr>
<tr>
<td>AM</td>
<td>3</td>
<td>0.47</td>
<td>0.55</td>
<td>0.78</td>
<td>0.68</td>
</tr>
<tr>
<td>AH</td>
<td>3</td>
<td>2.16</td>
<td>0.40</td>
<td>0.56</td>
<td>0.39</td>
</tr>
<tr>
<td>S</td>
<td>3</td>
<td>2.24</td>
<td>0.58</td>
<td>0.81</td>
<td>0.83</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>0.00</td>
<td>0.40</td>
<td>1.80</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Video Items
Intended Sublevel vs. Item Logit

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

<table>
<thead>
<tr>
<th>Level</th>
<th>Item Count</th>
<th>Measure</th>
<th>Standard Error</th>
<th>Standard Deviation</th>
<th>Model Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL</td>
<td>3</td>
<td>-2.11</td>
<td>0.98</td>
<td>-2.28</td>
<td>0.75</td>
</tr>
<tr>
<td>IM</td>
<td>3</td>
<td>-1.75</td>
<td>0.68</td>
<td>-2.17</td>
<td>0.36</td>
</tr>
<tr>
<td>IH</td>
<td>3</td>
<td>-0.68</td>
<td>0.61</td>
<td>-1.05</td>
<td>0.26</td>
</tr>
<tr>
<td>AL</td>
<td>3</td>
<td>-0.11</td>
<td>0.33</td>
<td>-0.26</td>
<td>0</td>
</tr>
<tr>
<td>AM</td>
<td>3</td>
<td>0.62</td>
<td>0.66</td>
<td>0.8</td>
<td>0.37</td>
</tr>
<tr>
<td>AH</td>
<td>3</td>
<td>1.82</td>
<td>0.17</td>
<td>1.85</td>
<td>0</td>
</tr>
<tr>
<td>S</td>
<td>3</td>
<td>2.21</td>
<td>0.75</td>
<td>2.36</td>
<td>0.68</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>0</td>
<td>1.67</td>
<td>-0.26</td>
<td>0.91</td>
</tr>
</tbody>
</table>
RQ: To what extent are a No-Video and Video can-do self-assessment survey comparable in terms of...

<table>
<thead>
<tr>
<th>Score Card: No-Video vs. Video Rasch Analysis</th>
<th>No-Video</th>
<th>Operator</th>
<th>Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating Scales?</td>
<td>✅✅</td>
<td>=</td>
<td>✅✅</td>
</tr>
<tr>
<td>Instrument Reliability?</td>
<td>✅</td>
<td>&lt;</td>
<td>✅✅</td>
</tr>
<tr>
<td>Intended vs. Actual Level Difficulty? (Intended ACTFL level and item logit)</td>
<td>✅</td>
<td>&lt;</td>
<td>✅✅</td>
</tr>
<tr>
<td>Intended Sublevel vs. Item Logit</td>
<td>✅✅</td>
<td>=</td>
<td>✅😂</td>
</tr>
</tbody>
</table>
No Video—People
OPI Score by Person Logit

Spearman’s Rho = .38,
p = .046
Spearman’s Rho = .40, p = .05

Video—People
OPI Score by Person Logit
RQ: To what extent are a No-Video and Video can-do self-assessment survey comparable in terms of...

<table>
<thead>
<tr>
<th>Score Card: No-Video vs. Video Rasch Analysis</th>
<th>No-Video</th>
<th>Operator</th>
<th>Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating Scale Diagnosis</td>
<td>🌟🌟</td>
<td>=</td>
<td>🌟🌟🌟</td>
</tr>
<tr>
<td>Instrument Reliability</td>
<td>🌟</td>
<td>&lt;</td>
<td>🌟🌟🌟</td>
</tr>
<tr>
<td>Items—Intended Major Level vs. Item Logit</td>
<td>🌟</td>
<td>&lt;</td>
<td>🌟🌟🌟</td>
</tr>
<tr>
<td>Items—Intended Sublevel vs. Item Logit</td>
<td>🌟🌟</td>
<td>=</td>
<td>🌟🌟🌟</td>
</tr>
<tr>
<td>Persons—OPI Level vs. Person Logit</td>
<td>😞</td>
<td>=</td>
<td>😞</td>
</tr>
</tbody>
</table>
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
<table>
<thead>
<tr>
<th>Family</th>
<th>Work</th>
<th>Education</th>
<th>Technology</th>
<th>Food</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superior</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Novice</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How confident are you that you could do the following tasks about family without time to prepare or reference tools (such as a dictionary)?

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Somewhat Agree</th>
<th>Neither Agree or Disagree</th>
<th>Somewhat Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Neutral" /></td>
<td><img src="image" alt="Frown" /></td>
<td><img src="image" alt="Sad" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task</th>
<th>Strongly Agree</th>
<th>Somewhat Agree</th>
<th>Neither Agree or Disagree</th>
<th>Somewhat Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I can name the members of my family.</td>
<td><img src="image" alt="Smiley" /></td>
<td><img src="image" alt="Neutral" /></td>
<td><img src="image" alt="Frown" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
</tr>
<tr>
<td>I can describe what my family members look like.</td>
<td><img src="image" alt="Neutral" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
</tr>
<tr>
<td>I can describe my family’s hobbies.</td>
<td><img src="image" alt="Neutral" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
</tr>
<tr>
<td>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.</td>
<td><img src="image" alt="Neutral" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
<td><img src="image" alt="Sad" /></td>
</tr>
</tbody>
</table>
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

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 Sq = .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

<table>
<thead>
<tr>
<th></th>
<th>Group 1</th>
<th>Group 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>n</strong></td>
<td>34</td>
<td>30</td>
<td>64</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>female</td>
<td>10</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>male</td>
<td>24</td>
<td>27</td>
<td>51</td>
</tr>
<tr>
<td><strong>Mean age</strong></td>
<td>22.8</td>
<td>21.7</td>
<td>22.3</td>
</tr>
</tbody>
</table>
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 кукол из Санкт-Петербурга, Воронежский камерный театр.
"Московский антагонизм", а также пройдет спектакль звезды последних четырех столичных театральных сезонов Григория Гринько.

Какое событие приближается?
A. Популярный актер выступит на сцене в последний раз.
B. Гастролирующие театральные труппы дают представления в местных театрах.
C. В Старом Осколе пройдет Фестиваль кукол.
D. В последний раз будет показан мюзикл “Северо-восток”.
E. Я не знаю

How confident are you in your answer choice?
very unconfident  unconfident  somewhat unconfident  somewhat confident  confident  very confident

Indicate your level of anxiety while answering this question.
very low  low  somewhat low  somewhat high  high  very high
What event is coming up?
A. The musical “North East” will end its long run.
B. A marionette festival will be in Old Oscol.
C. A popular actor will give his final performance.
D. Visiting theater groups will perform locally.
E. I don’t know

My response: [ ]
<table>
<thead>
<tr>
<th>Question Language</th>
<th>Statistic</th>
<th>Group A</th>
<th>Group B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English</strong></td>
<td>N</td>
<td>30</td>
<td>34</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>5.83</td>
<td>5.26</td>
<td>5.55</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>2.41</td>
<td>2.16</td>
<td>.27</td>
</tr>
<tr>
<td></td>
<td>95%CI</td>
<td>[4.95, 6.71]</td>
<td>[4.52, 6]</td>
<td>[4.99, 6.12]</td>
</tr>
<tr>
<td><strong>Russian</strong></td>
<td>N</td>
<td>30</td>
<td>34</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>4.73</td>
<td>4.00</td>
<td>4.37</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.84</td>
<td>2.04</td>
<td>.24</td>
</tr>
<tr>
<td></td>
<td>95%CI</td>
<td>[4.05, 5.41]</td>
<td>[3.30, 4.70]</td>
<td>[3.88, 4.86]</td>
</tr>
</tbody>
</table>
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
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

- Ability: Mean = 60.99, Std. Dev. = 15.367, N = 96
- Confidence: Mean = 72.16, Std. Dev. = 20.914, N = 96
Overconfident
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