

# Vocabulary Size as a Screener for Reading Proficiency

Erwin Tschirner, U of Leipzig

Jane Hacking, U of Utah

Fernando Rubio, U of Utah

**L2ReC**

SECOND LANGUAGE  
TEACHING & RESEARCH CENTER

UNIVERSITÄT LEIPZIG

# Overview

**Reading Proficiency and Vocabulary Size**

**Research Questions**

**Methods and Results**

**Discussion**



# Background

## **Strong relationship between reading proficiency and receptive vocabulary size**

- $r$ : between .5 and .85 (Milton 2013)

## **Estimates of vocabulary size required for general reading proficiency (wide range of non-academic texts) (Nation 2006)**

- 4,000 word families (newspaper and novels: 95% text coverage)
- 8,000 word families (newspaper: 98% text coverage)
- 9,000 word families (novels: 98% text coverage)

## **Estimates of vocabulary size required for academic reading proficiency**

- 10,000 lemmas (Hazenberg & Hulstijn 1996; Treffers-Daller & Milton 2013)

## CEFR and Vocabulary Size

		English (N=232)	English (N=192)
CEFR	ILR	Milton 2010	Huhta u.a. 2011
C1	3	5.000	5.000
B2	2+		
B1	2	3.000	3.000
A2	1+	2.000	2.000
A1	1	1.500	975



# Research Questions

- 1. How well does vocabulary size measured as the receptive knowledge of various bands of the most frequent 5,000 words in German, Russian, and Spanish predict reading proficiency as defined by the ILR?**
- 2. What ILR reading proficiency levels are predicted by what vocabulary sizes in these languages?**
- 3. What are the differences, if any, between German, Russian, and Spanish with respect to the relationship between vocabulary size and reading proficiency level?**



# Participants

<b>Spanish</b>	<b>Russian</b>	<b>German</b>
<b>52</b>	<b>48</b>	<b>97</b>



# Instruments

## ACTFL Reading Proficiency Test (RPT)

- ACTFL, ILR, and CEFR Ratings
- Russian: ILR; Spanish: ACTFL; German: ACTFL or CEFR
- All non-ILR results recoded according to the ILR algorithm
- Level Check: 10 Passages and 30 Items (50 min)
- Full Test: ILR 1-3+: 25 Passages and 75 Items (2 hours)
- Semi-Adaptive Full Test: 15 Passages and 45 Items (75 min)



# Instruments

## Vocabulary Levels Test (VLT)

- **Basis: Nation 1990**
- **Five Bands: 1-1000; 1001-2000; 2001-3000; 3001-4000; 4001-5000**
- **Each Band: 60 words: 30 nouns, 18 verbs, 12 adjectives (random)**
- **Level: Highest band with 80% correct**



## Vocabulary Levels Test

Choose a word from the left for each meaning on the right.  
Write the number of the word next to its meaning.

1	business		
2	clock		part of a house
3	horse		animal with four legs
4	pencil		something used for writing
5	shoe		
6	wall		

## How many words do you know?

Test your receptive and productive vocabulary in different languages.

What percentage of the 5000 most frequent words have you mastered?

- **German**  
receptive  
productive

- **English**  
receptive  
productive

- **French**  
receptive  
productive

- **Arabic**  
receptive  
productive

- **Japanese**  
receptive  
productive

- **Russian**  
receptive  
productive

## ■ About the Vocabulary Tests

Questions or suggestions:  
[info@itt-leipzig.de](mailto:info@itt-leipzig.de)

## Vocabulary Tests

- **Italian**  
receptive  
productive

- **Portuguese**  
receptive  
productive

- **Spanish**  
receptive  
productive

- **Chinese**  
receptive  
productive

- **American English**  
receptive  
productive

## In cooperation with:

Herder-Institute Leipzig

University of Leipzig -  
Language Centre

# Results

## VLT: Internal Consistency

	<b><i>N</i></b>	<b>Alpha</b>	<b>80%</b>
<b>German</b>	<b>97</b>	<b>.938</b>	<b>.774</b>
<b>Russian</b>	<b>48</b>	<b>.951</b>	<b>.959</b>
<b>Spanish</b>	<b>52</b>	<b>.951</b>	<b>.956</b>

Cronbach's Alpha Between Bands (Above .8 = good; above .9 = excellent)

Pearsons's Correlation Between Total Score and Level ( $p < .01$ )



## Correlations Between RPT and Vocabulary Size

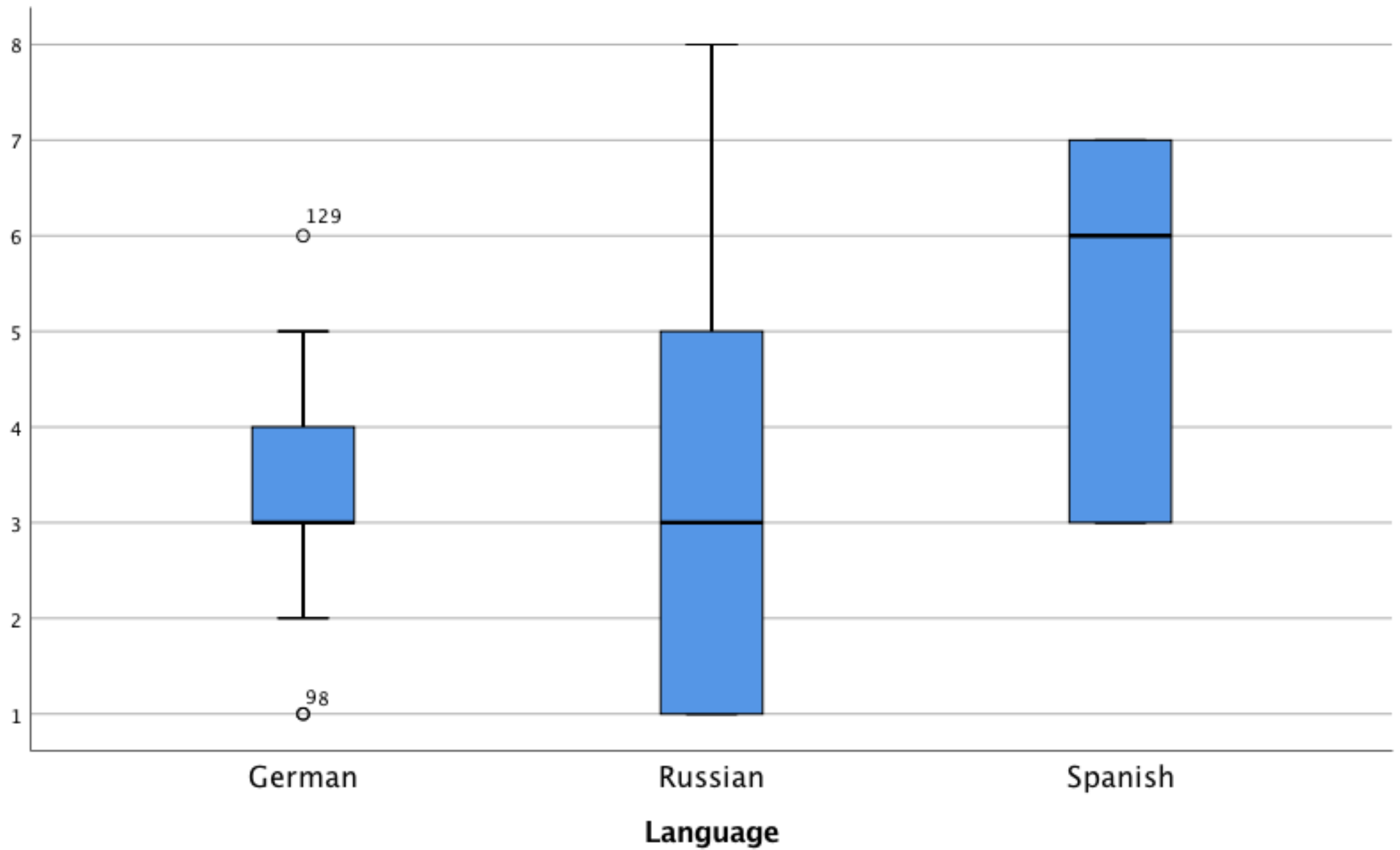
	<i>N</i>	Total	75%	<b>80%</b>	85%
German	97	.448	.546	<b>.570</b>	.483
Russian	48	.908	.872	<b>.869</b>	.822
Spanish	52	.793	.769	<b>.812</b>	.793

Spearman's Rho:  $p < .01$  (two-tailed)



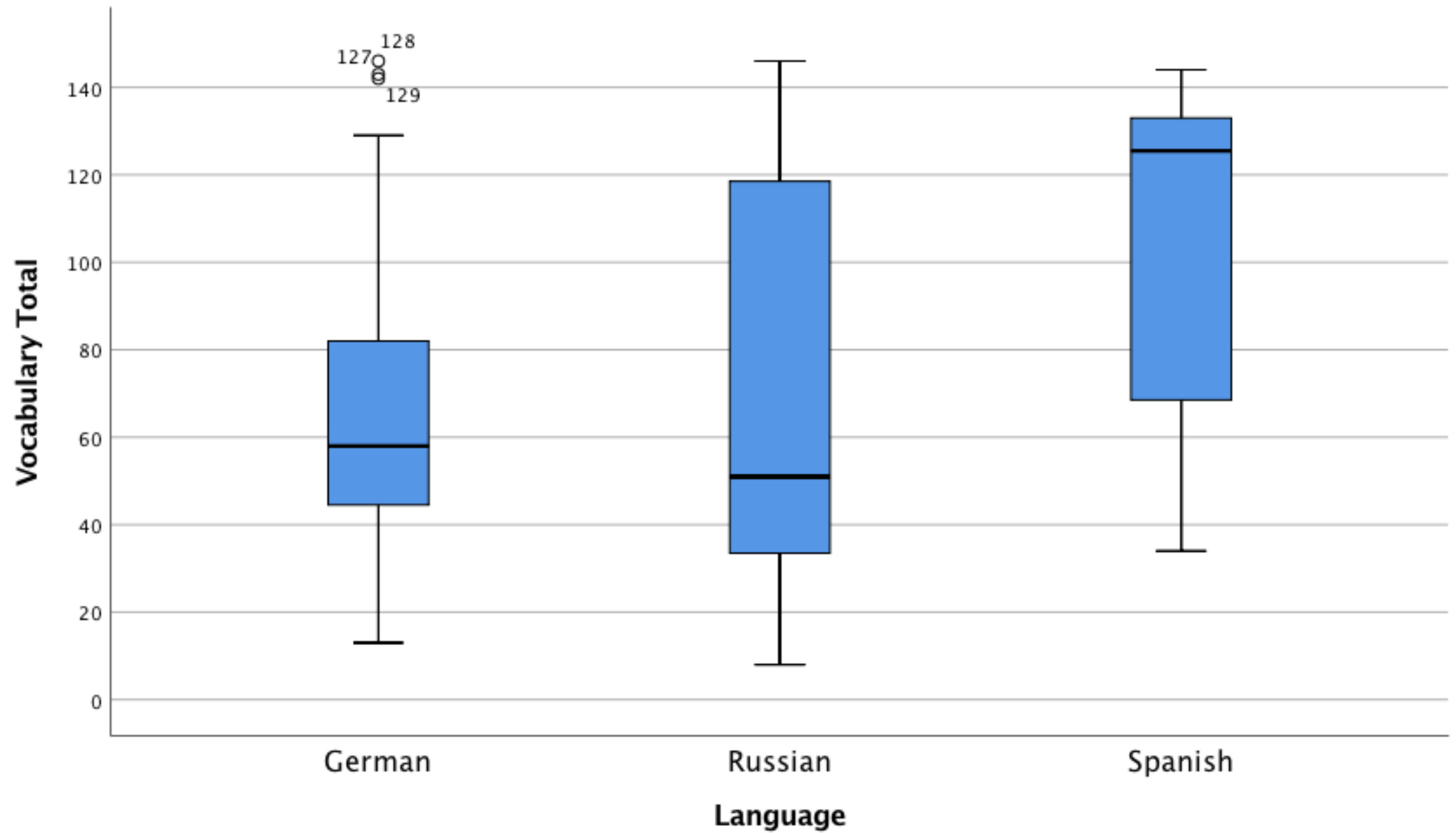
# ILR Levels

ILR Numeric (1=0, 2=0+, 3=1, 4=1+, 5=2, 6=2+, 7=3, 8=3+)



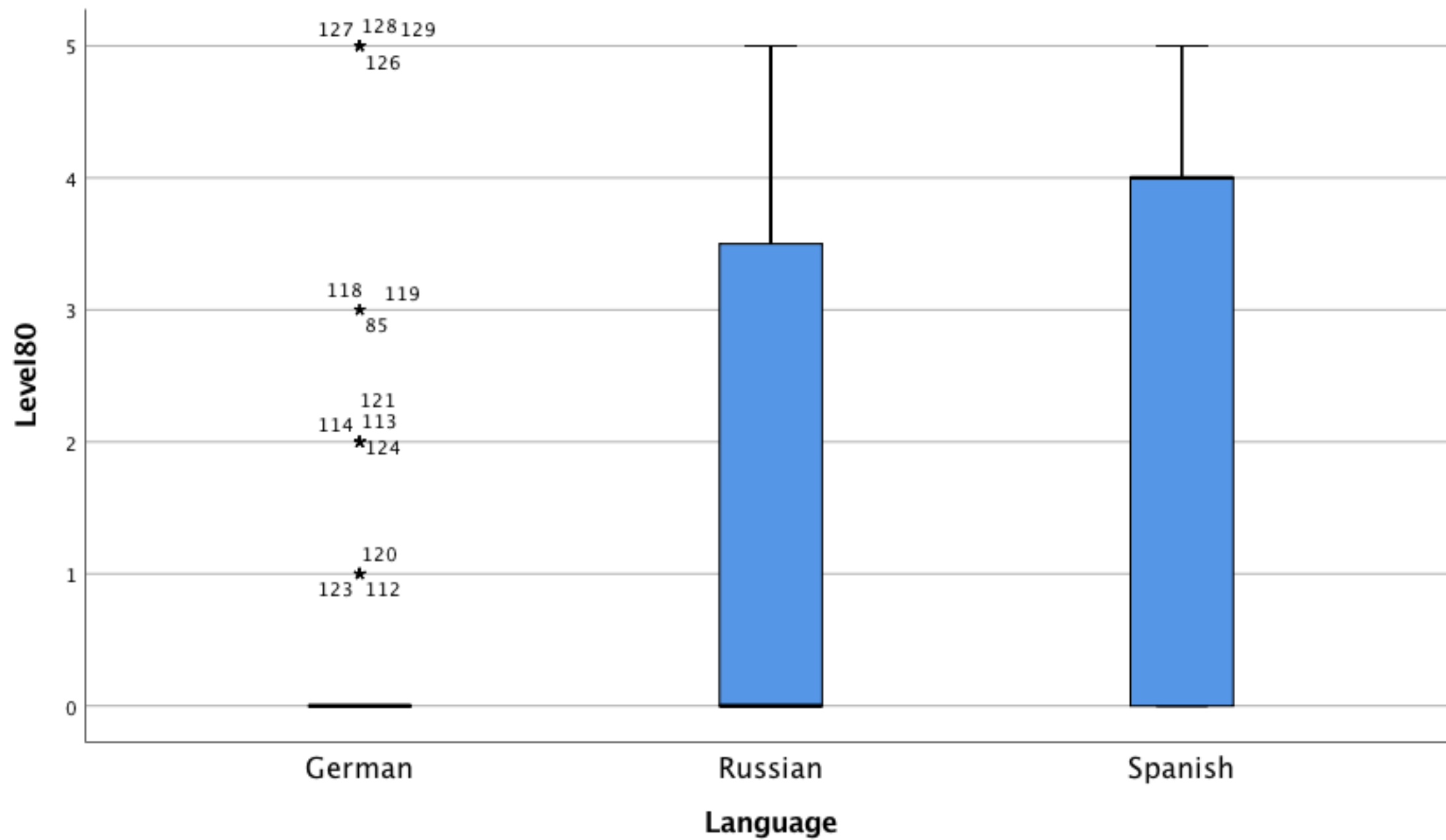


### Vocabulary Total (Out of 150)





### Vocabulary Size (80%)



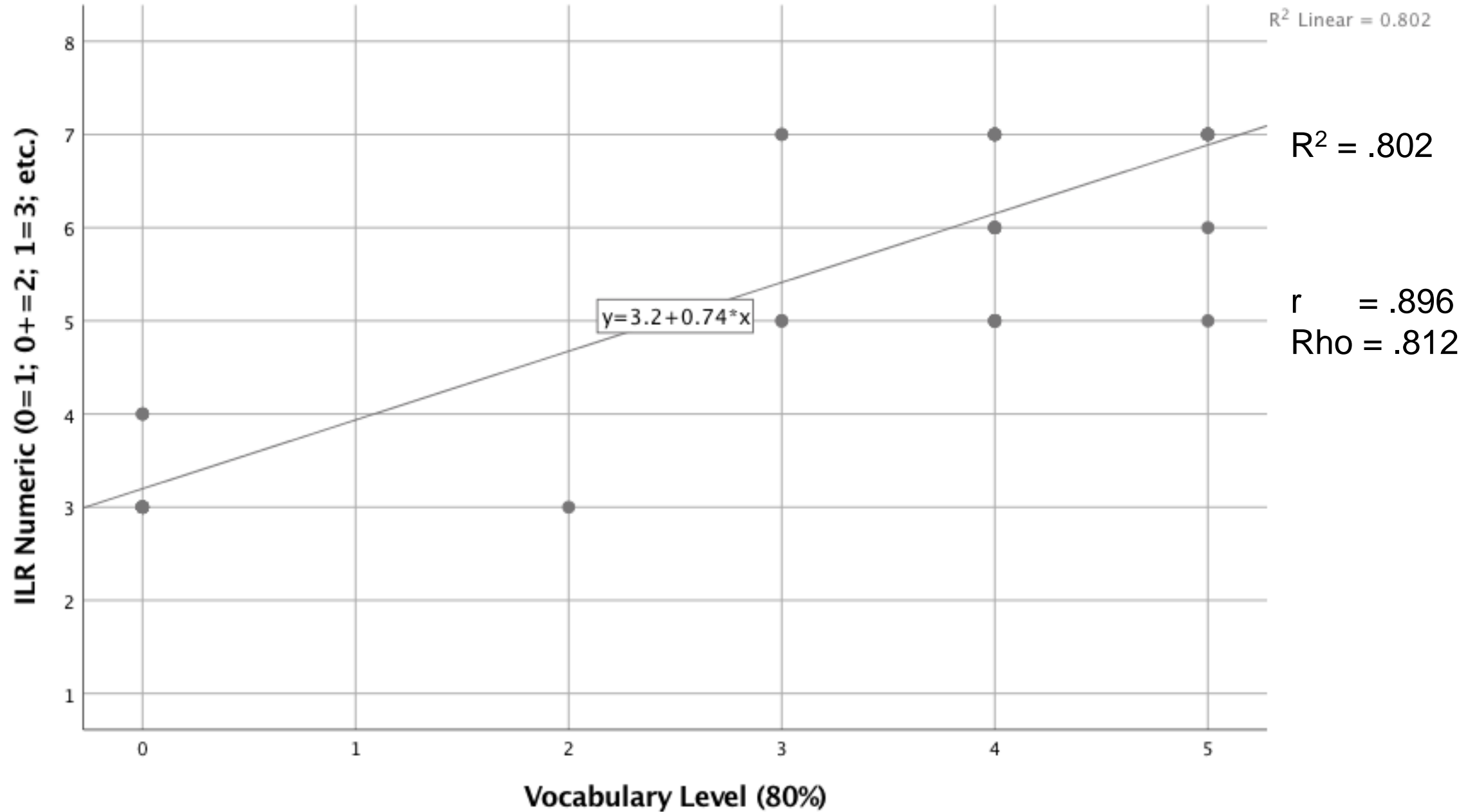


# Regression Analysis Spanish

# Crosstabulation: Spanish Vocabulary \* ILR

	<b>0</b>	<b>0+</b>	<b>1</b>	<b>1+</b>	<b>2</b>	<b>2+</b>	<b>3</b>	<b>3+</b>	<b>Total</b>
<b>0</b>			<b>13</b>	<b>3</b>					<b>16</b>
<b>1</b>									
<b>2</b>			<b>1</b>						<b>1</b>
<b>3</b>					<b>2</b>		<b>2</b>		<b>4</b>
<b>4</b>					<b>5</b>	<b>5</b>	<b>9</b>		<b>19</b>
<b>5</b>					<b>1</b>	<b>1</b>	<b>10</b>		<b>12</b>
<b>Total</b>			<b>14</b>	<b>3</b>	<b>8</b>	<b>6</b>	<b>21</b>		<b>52</b>

# Spanish



# Regression Analysis Predicting Spanish Reading Proficiency

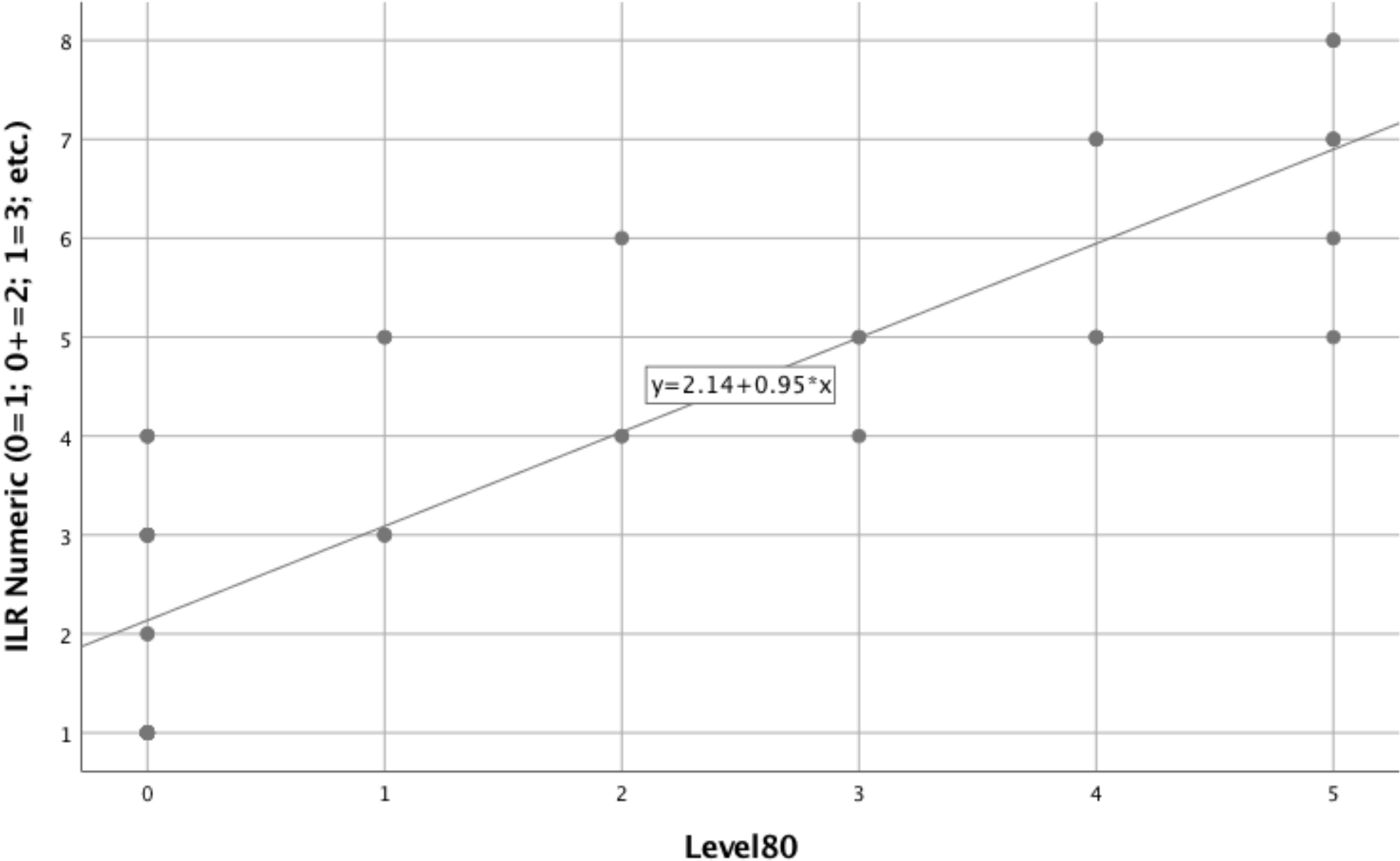
<b>Vocab Level</b>	<b>1000</b>	<b>2000</b>	<b>3000</b>	<b>4000</b>	<b>5000</b>
<b>Numeric Mean</b>	<b>(3.94)</b>	<b>(4.68)</b>	<b>5.42</b>	<b>6.16</b>	<b>6.90</b>
<b>ILR Level</b>	<b>1+</b>	<b>&lt;2</b>	<b>&gt;2</b>	<b>2+</b>	<b>3</b>

# Regression Analysis Russian

# Crosstabulation: Russian Vocabulary \* ILR

	<b>0</b>	<b>0+</b>	<b>1</b>	<b>1+</b>	<b>2</b>	<b>2+</b>	<b>3</b>	<b>3+</b>	<b>Total</b>
<b>0</b>	<b>13</b>	<b>2</b>	<b>7</b>	<b>3</b>					<b>25</b>
<b>1</b>			<b>3</b>		<b>2</b>				<b>5</b>
<b>2</b>				<b>2</b>		<b>1</b>			<b>3</b>
<b>3</b>				<b>1</b>	<b>2</b>				<b>3</b>
<b>4</b>					<b>3</b>		<b>2</b>		<b>5</b>
<b>5</b>					<b>1</b>	<b>1</b>	<b>3</b>	<b>2</b>	<b>7</b>
<b>Total</b>	<b>13</b>	<b>2</b>	<b>10</b>	<b>6</b>	<b>8</b>	<b>2</b>	<b>5</b>	<b>2</b>	<b>48</b>

# Russian



$R^2 \text{ Linear} = 0.745$

$R^2 = .745$

$r = .863$

$\text{Rho} = .869$

# Regression Analysis Predicting Russian Reading Proficiency

<b>Vocab Level</b>	<b>1000</b>	<b>2000</b>	<b>3000</b>	<b>4000</b>	<b>5000</b>
<b>Numeric Mean</b>	<b>3.09</b>	<b>4.04</b>	<b>4.99</b>	<b>5.94</b>	<b>6.89</b>
<b>ILR Level</b>	<b>1</b>	<b>1+</b>	<b>2</b>	<b>2+</b>	<b>3</b>

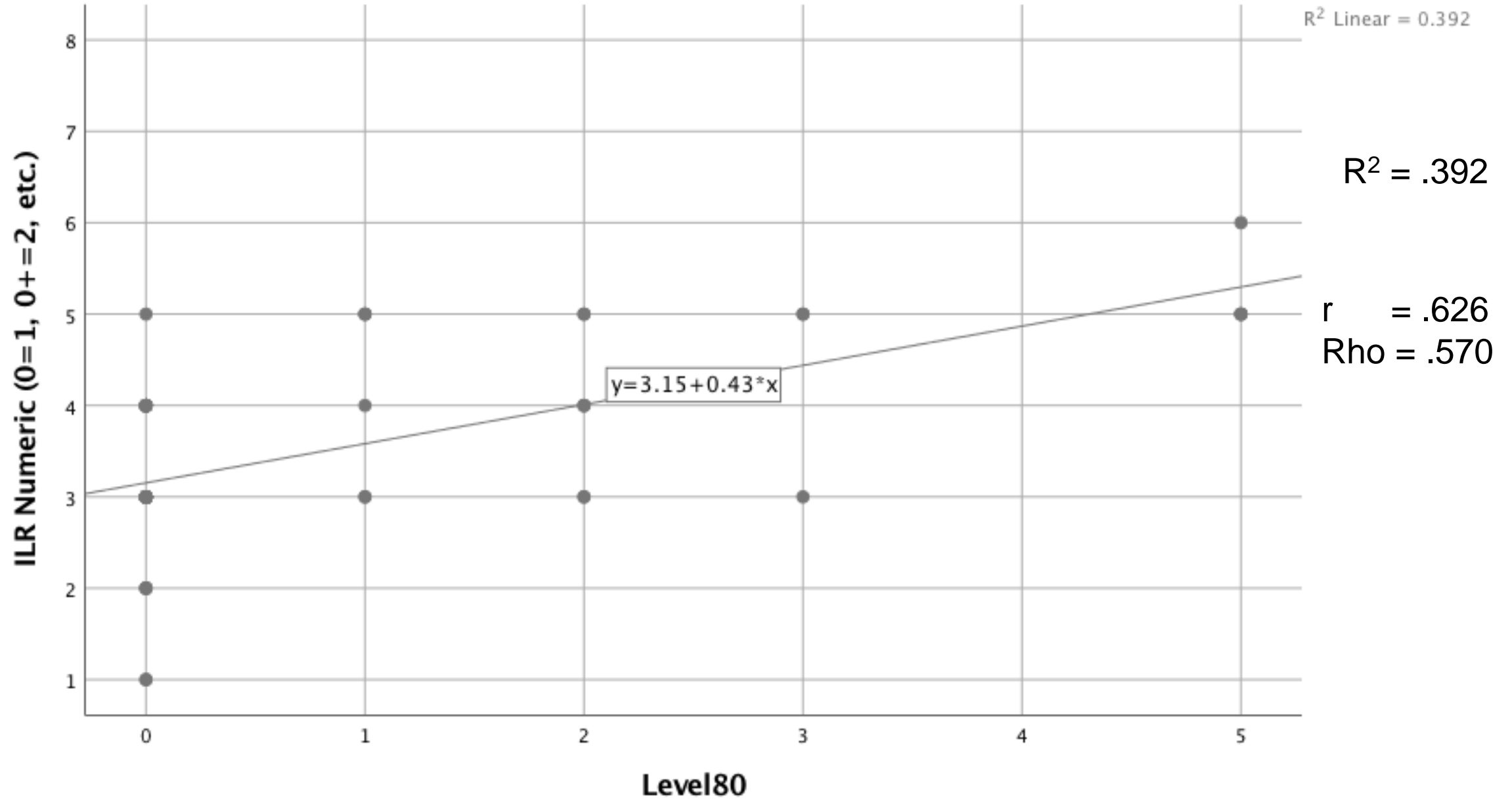


# Regression Analysis German

# Crosstabulation: German Vocabulary \* ILR

	<b>0</b>	<b>0+</b>	<b>1</b>	<b>1+</b>	<b>2</b>	<b>2+</b>	<b>3</b>	<b>3+</b>	<b>Total</b>
<b>0</b>	<b>2</b>	<b>4</b>	<b>54</b>	<b>15</b>	<b>1</b>				<b>76</b>
<b>1</b>			<b>2</b>	<b>1</b>	<b>3</b>				<b>6</b>
<b>2</b>			<b>2</b>	<b>3</b>	<b>2</b>				<b>7</b>
<b>3</b>			<b>1</b>		<b>2</b>				<b>3</b>
<b>4</b>									<b>0</b>
<b>5</b>					<b>4</b>	<b>1</b>			<b>5</b>
<b>Total</b>	<b>2</b>	<b>4</b>	<b>59</b>	<b>19</b>	<b>12</b>	<b>1</b>			<b>97</b>

# German



# Discussion



# Vocabulary Size and Reading Proficiency

<b>CEFR</b>	<b>ILR</b>	<b>Present Study</b>	<b>Milton 2010</b>	<b>Huhta u.a. 2011</b>
<b>C1</b>	<b>3</b>	<b>5000</b>	<b>5000</b>	<b>5000</b>
<b>B2</b>	<b>2+</b>	<b>4000</b>		
<b>B1</b>	<b>2</b>	<b>3000</b>	<b>3000</b>	<b>3000</b>
<b>A2</b>	<b>1+</b>	<b>2000</b>	<b>2000</b>	<b>2000</b>
<b>A1</b>	<b>1</b>	<b>1000</b>	<b>1500</b>	<b>975</b>

# False Positives and False Negatives

**False Positive:** Test predicts level when in fact the candidate has not reached it yet.

**False Negative:** Test predicts level has not been reached when in fact the candidate has reached it.

# Crosstabulation: Spanish and Russian

	<b>0</b>	<b>0+</b>	<b>1</b>	<b>1+</b>	<b>2</b>	<b>2+</b>	<b>3</b>	<b>3+</b>	<b>Total</b>
<b>0</b>	<b>13</b>	<b>2</b>	<b>20</b>	<b>6</b>					<b>41</b>
<b>1</b>			<b>3</b>		<b>2</b>				<b>5</b>
<b>2</b>			<b>1</b>	<b>2</b>		<b>1</b>			<b>4</b>
<b>3</b>				<b>1</b>	<b>4</b>		<b>2</b>		<b>7</b>
<b>4</b>					<b>8</b>	<b>5</b>	<b>11</b>		<b>24</b>
<b>5</b>					<b>2</b>	<b>2</b>	<b>13</b>	<b>2</b>	<b>19</b>
<b>Total</b>	<b>13</b>	<b>2</b>	<b>24</b>	<b>9</b>	<b>16</b>	<b>8</b>	<b>26</b>	<b>2</b>	<b>100</b>

# Predictions

Vocabulary Size	ILR Reading
1000	1
2000	1+
3000	2
4000	2+
5000	3



# Predictions: Spanish and Russian

	<b>0</b>	<b>0+</b>	<b>1</b>	<b>1+</b>	<b>2</b>	<b>2+</b>	<b>3</b>	<b>3+</b>	<b>Total</b>
<b>0</b>	13	2	20	6					41
<b>1</b>			3	2	2				5
<b>2</b>			1	2		1			4
<b>3</b>				1	4		2		7
<b>4</b>					8	5	11		24
<b>5</b>					2	2	13	2	19
<b>Total</b>	13	2	24	9	16	8	26	2	100

# False Positives: Spanish and Russian

	0	0+	1	1+	2	2+	3	3+	Total
0	13	2	20	6					41
1		0%	3	2	2				5
2			25%	2	1				4
3				14%	4	2			7
4					33%	5	11		24
5					21%		13	2	19
Total	13	2	24	9	16	8	26	2	100

# False Positives Beyond Adjacent Cells

	0	0+	1	1+	2	2+	3	3+	Total
0	13	2	20	6					41
1	0%		3	2	2				5
2		0%	1	2	1				4
3			0%	1	4	2			7
4				0%	8	5	11		24
5					11%	2	13	2	19
Total	13	2	24	9	16	8	26	2	100

# False Negatives: Spanish and Russian

	0	0+	1	1+	2	2+	3	3+	Total
0	13	2	83%	67%					41
1			3		13%				5
2			1	2		13%			4
3				1	4		50%		7
4					8	5			24
5					2	2	13	2	19
Total	13	2	24	9	16	8	26	2	100

# False Negatives Beyond Adjacent Cells

	0	0+	1	1+	2	2+	3	3+	Total
0	13	2	20	67%					41
1			3		13%				5
2			1	2		13%			4
3				1	4		8%		7
4					8	5	11		24
5					2	2	13	2	19
Total	13	2	24	9	16	8	26	2	100

# Reading Test Ranges

**Reduce False Positives: Start at one sublevel below prediction.**

**Reduce False Negatives: End at one or two sublevels above prediction.**

# Reading Test Ranges Based on Vocabulary Size

	<b>0</b>	<b>0+</b>	<b>1</b>	<b>1+</b>	<b>2</b>	<b>2+</b>	<b>3</b>	<b>3+</b>	<b>Total</b>
<b>0</b>	<b>13</b>	<b>2</b>	<b>20</b>	<b>6</b>					<b>41</b>
<b>1</b>			<b>3</b>		<b>2</b>				<b>5</b>
<b>2</b>			<b>1</b>	<b>2</b>		<b>1</b>			<b>4</b>
<b>3</b>				<b>1</b>	<b>4</b>		<b>2</b>		<b>7</b>
<b>4</b>					<b>8</b>	<b>5</b>	<b>11</b>		<b>24</b>
<b>5</b>					<b>2</b>	<b>2</b>	<b>13</b>	<b>2</b>	<b>19</b>
<b>Total</b>	<b>13</b>	<b>2</b>	<b>24</b>	<b>9</b>	<b>16</b>	<b>8</b>	<b>26</b>	<b>2</b>	<b>100</b>

# Caveats

**Limited number of languages and subjects**

**Predictions may differ depending on**

- Target language
- Language distance (including common heritage)

**VLTs may be best for low stakes situations (placement; screener test; supporting evidence)**