Assessing written English skills for business communication using time-constrained tasks

Alistair Van Moere

Ryan Downey

Masanori Suzuki

Mallory Klungtvedt

Pearson

Knowledge Technologies

Development of a Workplace Writing Test

1. Background
2. Test & task design
3. Validity questions
4. Results
5. Conclusions

LTRC, Ann Arbor, MI   June 24, 2011

PEARSON

Widely-used Assessments of Written Skills

Test	# of Tasks	# of Items	Total Test Time	Scale	Scores Reported
TOEIC – Writing Test	3	8	60 mins	0-200	Overall Writing Score
IELTS – General Training	2	2	60 mins	0-9	Overall Writing Score
BULATS	2	2	45 mins	0-100	Overall Writing Score

Needs Gap

- Few authentic measures of writing <u>efficiency</u>

- Lack of task variety

- 2-3 weeks to receive scores

- Only an overall score reported

- Inflexible structure: only BULATS offers a writing only test

PEARSON

Needs Analysis

Interviewed 10 companies from 5 countries

Online questionnaire, 157 respondents

- Multi-national companies
- Business Process Outsourcing (BPOs) companies

- HR managers
- Recruitment managers
- Training managers

PEARSON

Needs Analysis Results

Area	Results
Structure	Modular: Skills can be assessed either together or separately
Content	General English + Business English
Length	35 – 45 minutes + Quick score return
Tasks	Authentic, real-time tasks, e.g. writing an email
Feedback	Writing skills: Voice, tone, organization, succinctness

SPOKEN MODULE   **WRITTEN MODULE**

PEARSON
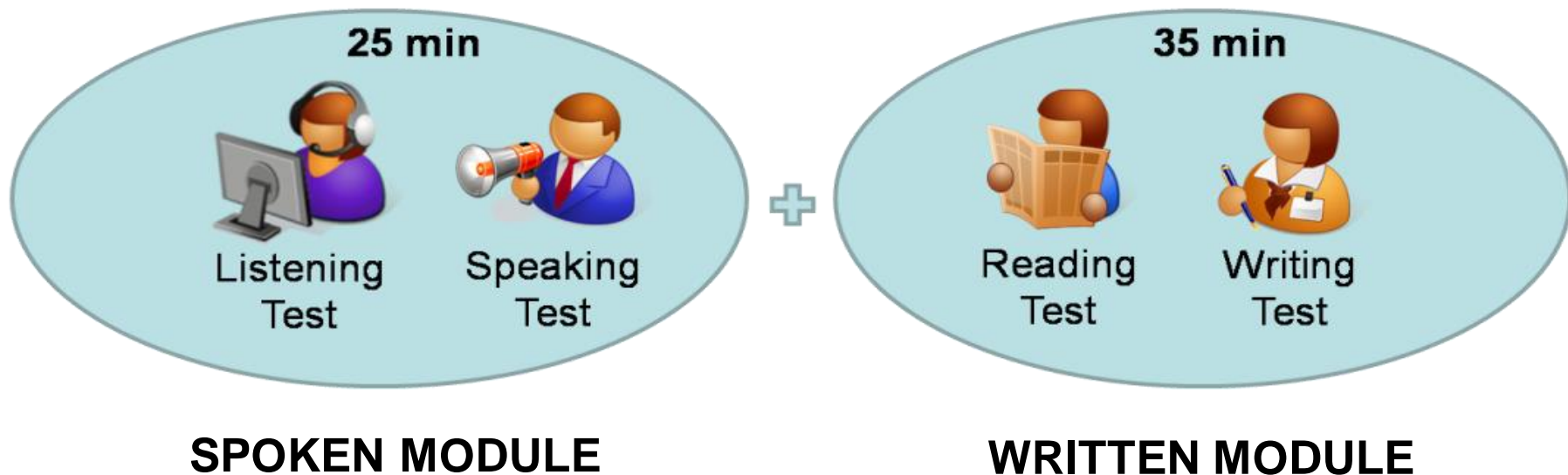
Testing goals: Written Module

- Flexible testing: target desired skills
- Speed and convenience
- Quick score turnaround
- Workplace-relevant tasks
- <u>Efficiency</u> and <u>appropriateness</u> of written skills

PEARSON

Written Module

Time/ mins	Task	Num of Items	Description
1	Typing	1	Type as many words in the presented text as possible
8	Sentence Completion	18	Read a sentence and enter a word that is missing
6	Dictation	14	Listen to a sentence and type it
8	Passage Reconstruction	4	Read a passage for 30 sec, then reproduce the passage in 90 sec
18	Email Writing	2	Compose an email to a client, colleague, supervisor in 9 min

45 mins, 5 tasks, 39 items

PEARSON

PEARSON

Versant Pro - Writing (Sample)

Time Left
0:04

Question Number
2 of 2

Part D: Passage Reconstruction

You read

> Thank you so much for being so understanding about our delay of shipment. It has been quite difficult to get materials from our suppliers due to the recent weather conditions. It is an unusual circumstance. In any case, we should be able to ship the products to you tomorrow. In the meantime, if you have any questions, please feel free to contact me.

PEARSON

Versant Pro - Writing (Sample)

Time Left
1:19

Question Number
2 of 2

MENU

Part D: Passage Reconstruction

You type

Next

LTRC, Ann Arbor, MI   June 24, 2011

PEARSON

Part E: E-Mail Writing

You read

You type   Word count: 0

The company you work for recently hired several new employees. Your sales team has been asked for ideas about the training they should receive. Write an email to your manager, Mr. Jenkins, suggesting three areas new-employee training should focus on.

Your suggestions must come from the following three themes:

- company history
- product knowledge
- communicating with customers

You should include all three themes. Provide supporting ideas for each of your suggestions.

Next

SCORE REPORT — Versant Pro - Writing

Test Identification Number: 12345678
Test Completion Date: January 1, 2010
Test Completion Time: 1:23 PM (UTC)

OVERALL SCORE: 52

SKILL AREA	SCORE	20	30	40	50	60	70	80
Overall	52							
Grammar	58							
Vocabulary	53							
Organization	45							
Voice & Tone	47							
Reading Comprehension	57							

DESCRIPTION	
Overall	The Overall Score of the test represents the ability to understand English input and write accurate, appropriate texts at a functional pace for everyday and workplace purposes. Scores are based on a weighted combination of five sub-scores. Scores are reported in the range from 20 to 80.
Candidate's Capabilities	Candidate understands texts using a variety of words and structures, and given enough time can produce written texts for general purposes. Writing contains errors or inappropriate word choice, but the message is clear to a sympathetic reader.

© 2010 Pearson Education, Inc. or its affiliate(s). All rights reserved. Ordinate and Versant are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliate(s). Other names may be the trademarks of their respective owners.

For more information, visit us online at www.VersantTest.com

PEARSON

Overall Score (20-80)

- Grammar
- Vocabulary
- Organization
- Voice & Tone
- Reading Comprehension

Additional Information
- Typing Speed
- Typing Accuracy

Development of a Workplace Writing Test

1. Background
2. Test & Task Design
   a) Item specifications
   b) Item development
   c) Field testing
   d) Rating scale design
3. Validity Questions
4. Results
5. Discussion

PEARSON

Item Specifications

Part E: E-Mail Writing

You read

You type   Word count: 0

The company you work for recently hired several new employees. Your sales team has been asked for ideas about the training they should receive. Write an email to your manager, Mr. Jenkins, suggesting three areas new-employee training should focus on.

Your suggestions must come from the following three themes:

- company history
- product knowledge
- communicating with customers

You should include all three themes. Provide supporting ideas for each of your suggestions.

Email Writing task with 3 themes:

- Cognitively relevant

- No specific business/domain knowledge required

- Free of cultural/geographic bias

- Elicits opportunities to demonstrate tone, voice, organization

- Control for creativity

- Constrain topic of responses for prompt-specific automated scoring models

Next

Item Development

Source material	• Texts modeled on actual workplace emails • Situations inspired from workplace communication

Word list	• General English: Switchboard Corpus
	    ~8,000 most frequent words
	• Business English: 4 corpus-based business word lists
	    ~3,500 most frequent words

Expert review	• Internal reviews by test developers • External reviews by subject matter experts

PEARSON

Rating Scales

Passage Reconstruction

Trait (0-6)	Description
Narrative Clarity & Accuracy	Ability to reproduce writing in clear, understandable, accurate English. Ability to convey fully the situation, characters, actions and ending of a text.

Email Writing

Trait (0-6)	Description
Organization	Extent to which the content of the email is logically organized.
Email Conventions	Extent to which the text conforms to conventions of formal email writing.
Grammatical Range & Accuracy	Range and control of sentence structure, grammar, and accuracy of spelling, punctuation, and capitalization.
Task Completion	Extent to which the written performance fulfills the requirements of the task, and addresses and elaborates on the three themes.
Voice & Tone	Extent to which the writing is appropriate for / demonstrates an awareness of the topic, purpose, and audience.
Vocabulary/Word Choice	Range of vocabulary and the accuracy and appropriacy of its use for the topic, purpose, and audience.

Field Testing

Top 10 Countries
India
Korea
Philippines
China
Japan
Argentina
Iran
Armenia
Mexico
Germany

Category	Num Tests	Mean Age
Non-Natives	1600+	27.9
Natives	700+	35.6

Other countries include:

France, Spain, Italy, Costa Rica, Russia, Iraq, Taiwan, Czech, Columbia, Yemen, Iran, Malaysia, Vietnam, Thailand, Venezuela, Nepal, etc…..

51 countries

58 L1s

Period: August 2009 – November 2009

Development of a Workplace Writing Test

1. Background
2. Test & task design
3. **Validity questions**
4. Results
5. Conclusions

PEARSON

Validity Questions

1. Do the tasks elicit performances which can be scored reliably?
   1. Rater reliability
   2. Generalizability?

2. Does the rating scale operate effectively?
   1. Do the traits tap distinct abilities?
   2. Are the bands separable?

3. What is the performance of machine scoring?
   1. Reliability
   2. Correlation with human judgments

Rater Reliability

Passage Reconstruction

Trait	Average of rater pairings
Narrative Clarity & Accuracy	.94
Exact Agreement	47 %

(21,200 ratings, 9 raters)

Email Writing

Trait	Average of rater pairings
Email Conventions	.85
Task Comp	.84
Organization	.85
Word Choice	.88
Grammar	.87
Voice & Tone	.81
Overall Score	.94
Exact Agreement	47 %

PEARSON

Generalizability Coefficients

		Number of Raters			
		1	2	3	4
Number of items	1	0.71	0.75	0.77	0.77
	2	0.83	0.86	0.87	0.87
	3	0.88	0.90	0.91	0.91
	4	0.90	0.92	0.93	0.93

		Average of 2 ratings
Number of items	1	0.76
	2	0.86
	3	0.90
	4	0.93

Passage Reconstruction
(n=2,118 * 4 prompts * 2 ratings)

Validity Questions

1. Do the tasks elicit performances which can be scored reliably?
   1. Rater reliability
   2. Generalizability?

2. **Does the rating scale operate effectively?**
   1. **Do the traits tap distinct abilities?**
   2. **Are the bands separable?**

3. What is the performance of machine scoring?
   1. Reliability
   2. Correlation with human judgments

PEARSON

Email Writing

```
-------------------------------------------------------------------------------------------
|Measr|+calls        |+graders        |+ trait          | Gram |Voc |Conv | Tone | Org |Task|
-------------------------------------------------------------------------------------------
+   5 + *.           +                +                 + (6) + (6) + (4) + (6) + (6) + (6) +
|     | .            |                |                 |     |     |     |     |     |     |
|     | *.           |                |                 | --- | --- |     |     |     |     |
|     | *            |                |                 |     |     |     |     | --- |     |
+   4 +              +                +                 +     +     +     + --- +     +     +
|     | .            |                |                 |     |     |     |     |     | --- |
|     | .            |                |                 |  5  |  5  |     |     |     |     |
|     | .            |                |                 |     |     | --- |     |  5  |     |
+   3 + *.           +                +                 +     +     +     + 5   +     +     +
|     | **.          |                |                 | --- | --- |     |     |     |  5  |
|     | **.          |                |                 |     |     |     |     | --- |     |
|     | *.           |                |                 |     |     | 3   | --- |     | --- |
+   2 + *            +                +                 + 4   + 4   +     +     +     +     +
|     | ***.         |  rater1 rater3 |                 |     |     |     | 4   | 4   |     |
|     | ***.         |  rater4        |                 |     |     |     |     |     | 4   |
|     | ****         |  rater5        |                 | --- |     | --- |     |     |     |
+   1 + ***.         |  rater2        +                 +     + --- +     + --- + --- +     +
|     | ***.         |                |                 |     |     |     |     |     | --- |
|     | *****.       |                | conv            |     |     |     |     |     |     |
|     | ***.         |                | task            | 3   |     | 2   | 3   | 3   |     |
*   0 * ******.      *                * vocab           *     * 3   *     *     *     * 3   *
|     | ******       |                | gram    org tone|     |     |     |     |     |     |
|     | *******.     |                |                 |     |     |     | --- | --- |     |
|     | *****        |                |                 | --- |     |     |     |     | --- |
+  -1 + *****.       +                +                 +     + --- + --- +     +     +     +
|     | ***.         |                |                 |     |     |     |     |     |     |
|     | ***.         |                |                 |     |     |     | 2   |     | 2   |
|     | ***.         |                |                 |     |     |     |     | 2   |     |
+  -2 + **.          +                +                 + 2   + 2   +     +     +     +     +
|     | *.           |                |                 |     |     | 1   |     |     | --- |
|     | ***.         |                |                 |     |     | --- |     |     |     |
|     | *.           |                |                 |     | --- |     |     | --- | 1   |
+  -3 + .            +                +                 +     +     +     +     +     +     +
|     | *.           |                |                 | --- |     |     | 1   |     |     |
|     | .            |                |                 |     | 1   | --- |     |     |     |
|     | .            |                |                 |     |     |     |     | 1   | --- |
+  -4 + .            +                +                 + 1   +     +     +     +     +     +
|     | .            |                |                 |     |     |     | --- |     |     |
|     | .            |                |                 |     | --- |     |     |     |     |
|     | .            |                |                 | --- |     |     |     | --- |     |
+  -5 +              +                +                 +     +     +     +     +     +     +
|     |              |                |                 |     |     |     |     |     |     |
|     |              |                |                 |     |     |     |     |     |     |
|     | .            |                |                 |     |     |     |     |     |     |
+  -6 + *********.   +                +                 + (0) + (0) + (0) + (0) + (0) + (0) +
-------------------------------------------------------------------------------------------
|Measr| * = 5        |+graders        |+                | S.1 | S.2 | S.3 | S.4 | S.5 | S.6 |
-------------------------------------------------------------------------------------------
```

Inter-correlation matrix

	Conven-tions	Task Completion	Organi-zation	Vocab-ulary	Gram-mar
Email Conventions					
Task Completion	0.79				
Organization	0.83	0.90			
Vocabulary	0.83	0.89	0.88		
Grammar	0.82	0.85	0.87	0.92	
Voice & Tone	0.82	0.87	0.89	0.91	0.88

1. Do the tasks elicit performances which can be scored reliably?
    1. Rater reliability
    2. Generalizability?

2. Does the rating scale operate effectively?
    1. Do the traits tap distinct abilities?
    2. Are the bands separable?

3. What is the performance of machine scoring?
    1. Reliability
    2. Correlation with human judgments

Subscore reliability

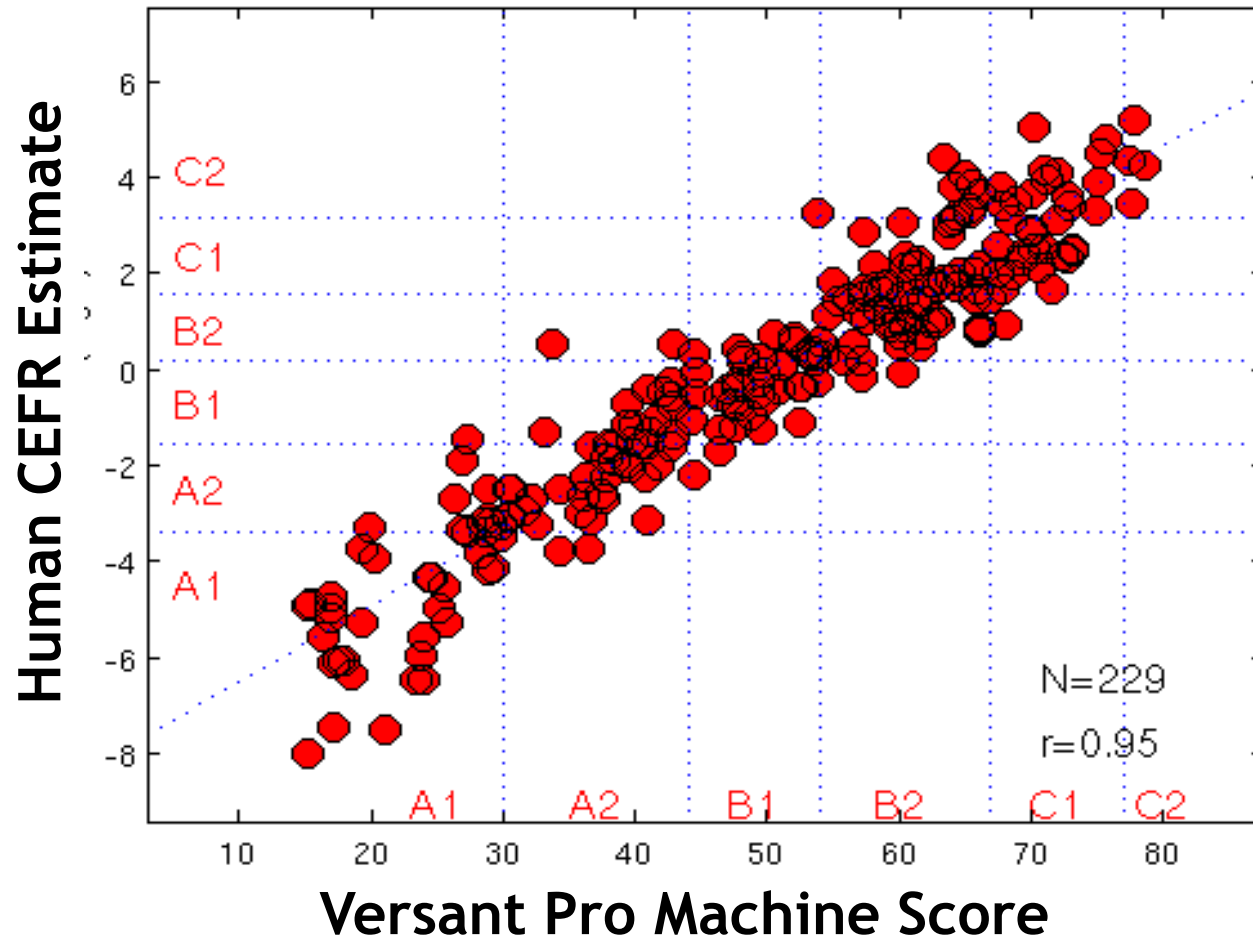|	Split-half Reliability	Machine-to-Human Correlation
Grammar	.98	.99
Vocabulary	.91	.98
Organization	.87	.90
Voice & Tone	.90	.91
Comprehension	.93	.96
Overall	.98	.98

Email items - Machine score vs Human Score

PEARSON
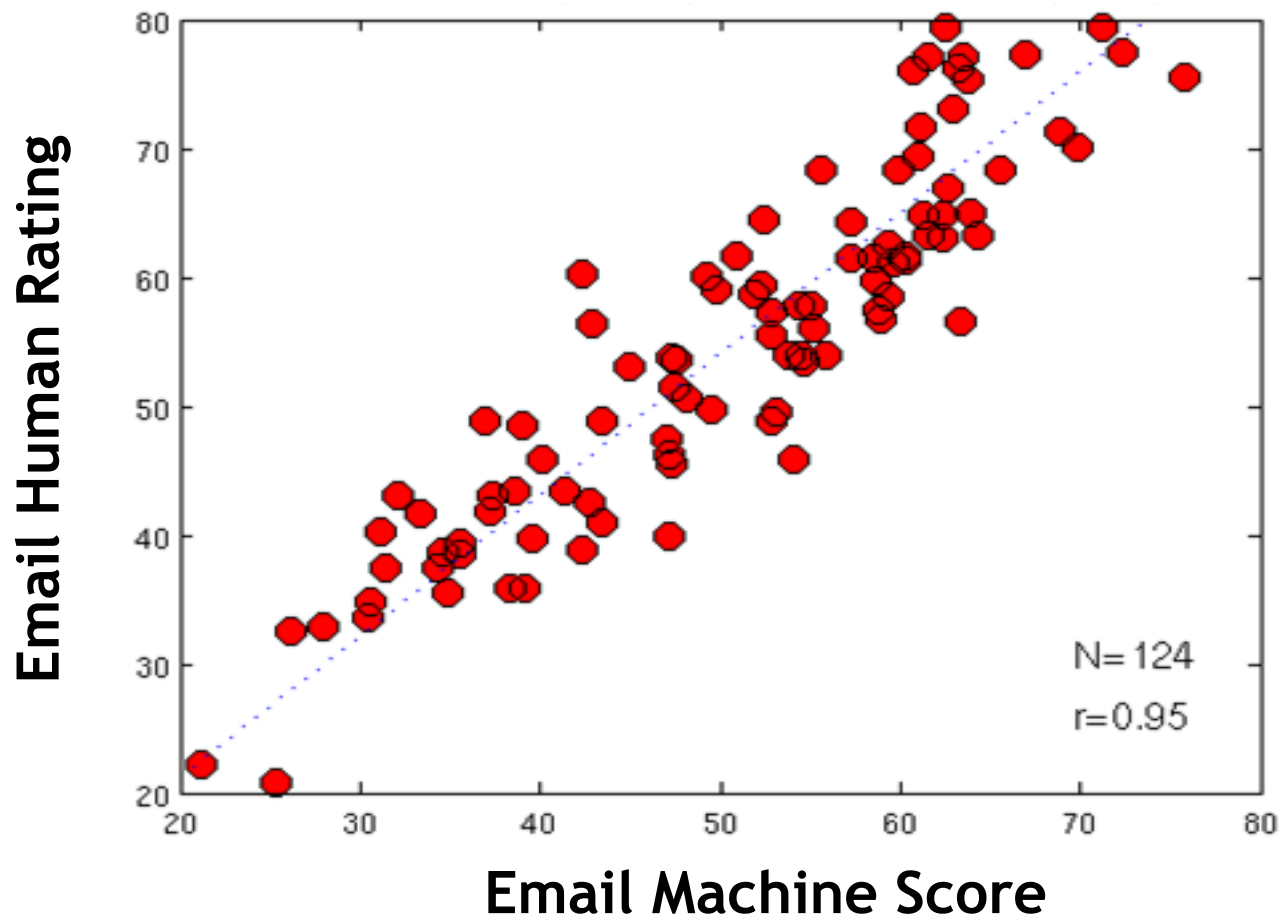
Versant Pro - Machine score vs Human Score
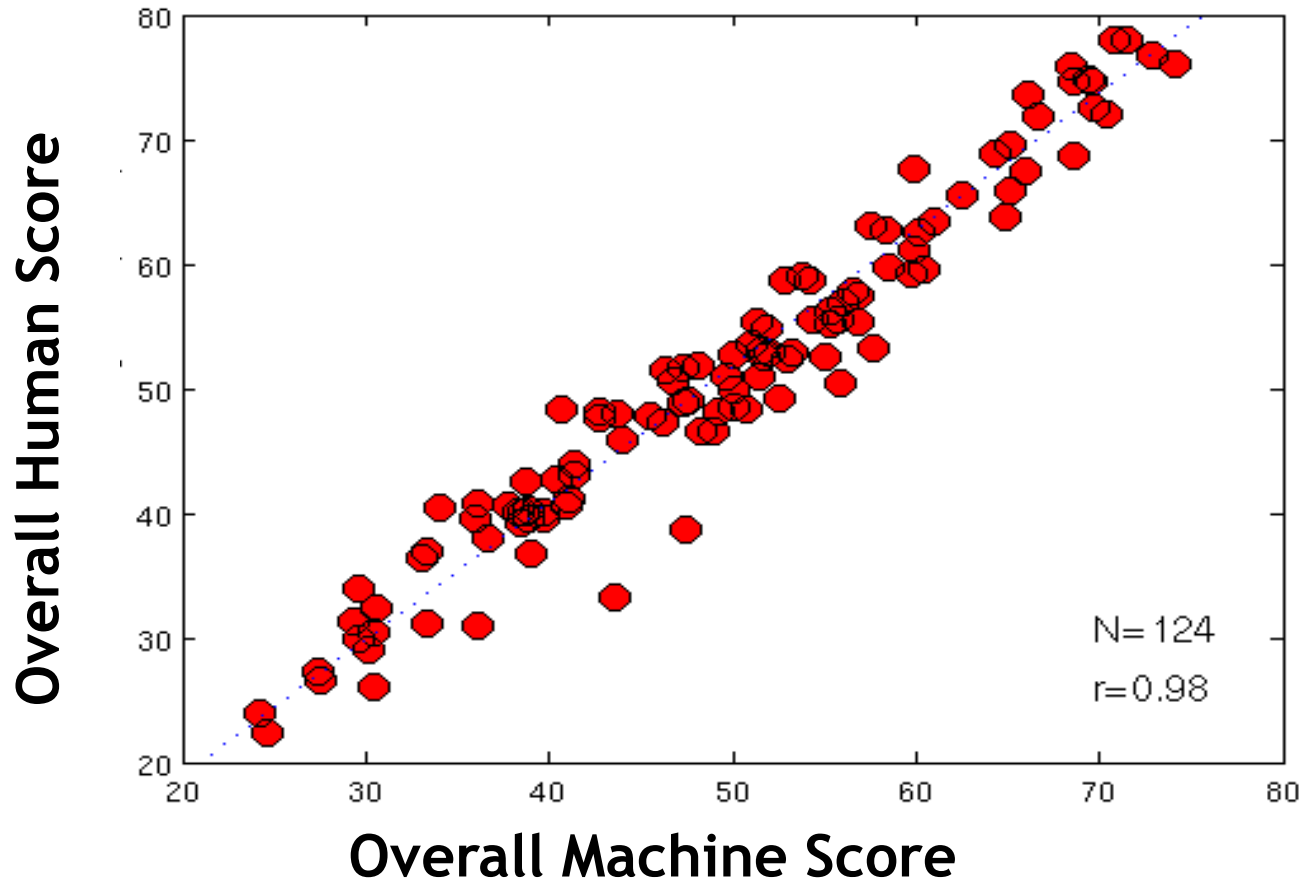
Machine score vs CEFR judgments

Limitations/Further work

- We do not have a complete validation argument
  - Predictive validity
  - Concurrent validity
  - Dimensionality (factor analysis, SEM)
- Score use in specific contexts
- Constructs not assessed, under-represented
  - Different types of writing (e.g., summaries)
  - More reading-specific items

Conclusion

Automatically-scored test of workplace written skills:

- Modular, flexible
- Short (45-mins)
- 5-min score turnaround
- Job relevant
- Task variety

- **Common shortfall in task design for written tests is <u>planning</u> time and <u>execution</u> time**

  - **We believe that shorter, more numerous, real-time tasks are construct-relevant, efficient and reliable.**

Thank you