



Rethinking the Rating Process: Solution to the Threshold Performance Dilemma

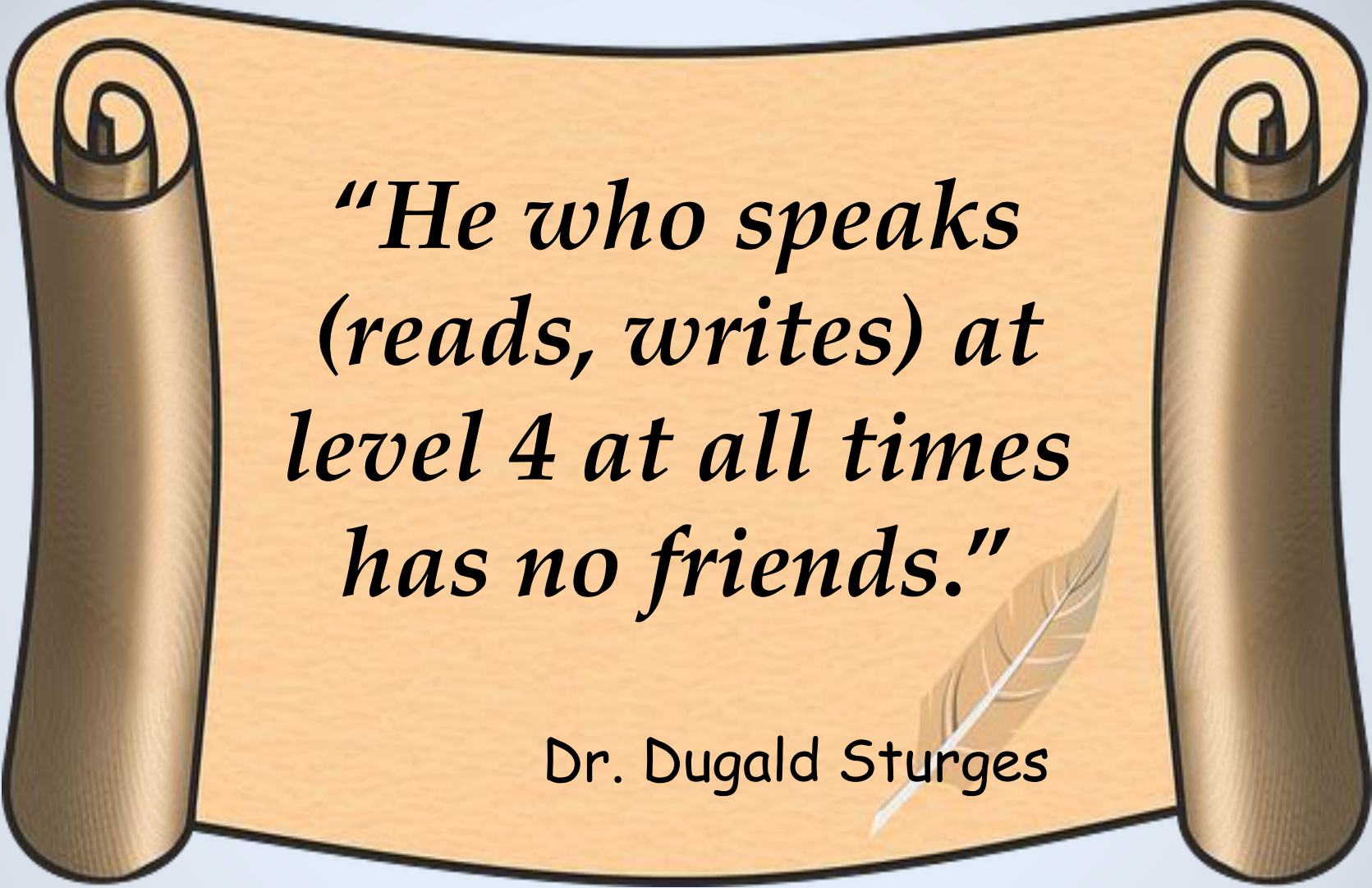
Research into Level 4 Reading Proficiency IAW STANAG 6001

Jana Vasilj-Begovic, *FL Stds O – MPGG, Canada, and Associate BILC Secretary*

Mary Jo DiBiase-Lubrano, *Associate director Yale University Center for Language Study (USA)*

BILC Annual Conference

Slovenia, May 2021

A scroll with a feather. The scroll is unrolled in the center, showing a light brown, textured surface. The ends of the scroll are rolled up. A single feather is positioned in the lower right quadrant of the scroll. The text is written in a black, serif font.

*“He who speaks
(reads, writes) at
level 4 at all times
has no friends.”*

Dr. Dugald Sturges



- ❖ Introduction and Background
- ❖ Level 4 Reading Skill Construct
- ❖ Prototype Test Design
- ❖ Statement of Problem & Research Questions
- ❖ Retrodictive Modelling Approach (RMA)
- ❖ Piloting and results
- ❖ Conclusions

Introduction & Background

- ❖ WG on Level 4 Proficiency established by BILC at its 2010 Istanbul Conference
- ❖ Prototype reading test developed as one of the products
- ❖ Language Needs Analysis (LNA), 2015: **real** Level 3 proficiency needed for most job tasks

Literature Review

- ❖ Edwards (1996) highly individualized or culture-specific forms of discourse, abstract metaphors, and symbolism. The author leaves historical, cultural or other references and assumptions unexplained.
- ❖ Child (1998) author's unique point of view, and the method of argumentation may be complex and innovative at higher levels of proficiency
- ❖ Lowe, (1998) texts are abstract & culturally dense with embedding syntax used with virtuosity
- ❖ Clifford (2013) and Aschuler (2002) reading as a cognitive process involving also general intellectual reasoning



Content – Task - Accuracy

Demonstrates strong competence in reading all styles and forms of the written language used for professional purposes, including texts from unfamiliar general and professional-specialist areas. Can readily follow unpredictable turns of thought on any subject matter addressed to the general reader. Shows both global and detailed understanding of texts including highly abstract concepts.

Contexts include newspapers, magazines, and professional literature written for the well-educated reader and may contain topics from such areas as economics, culture, science, and technology, as well as from the reader's own field.

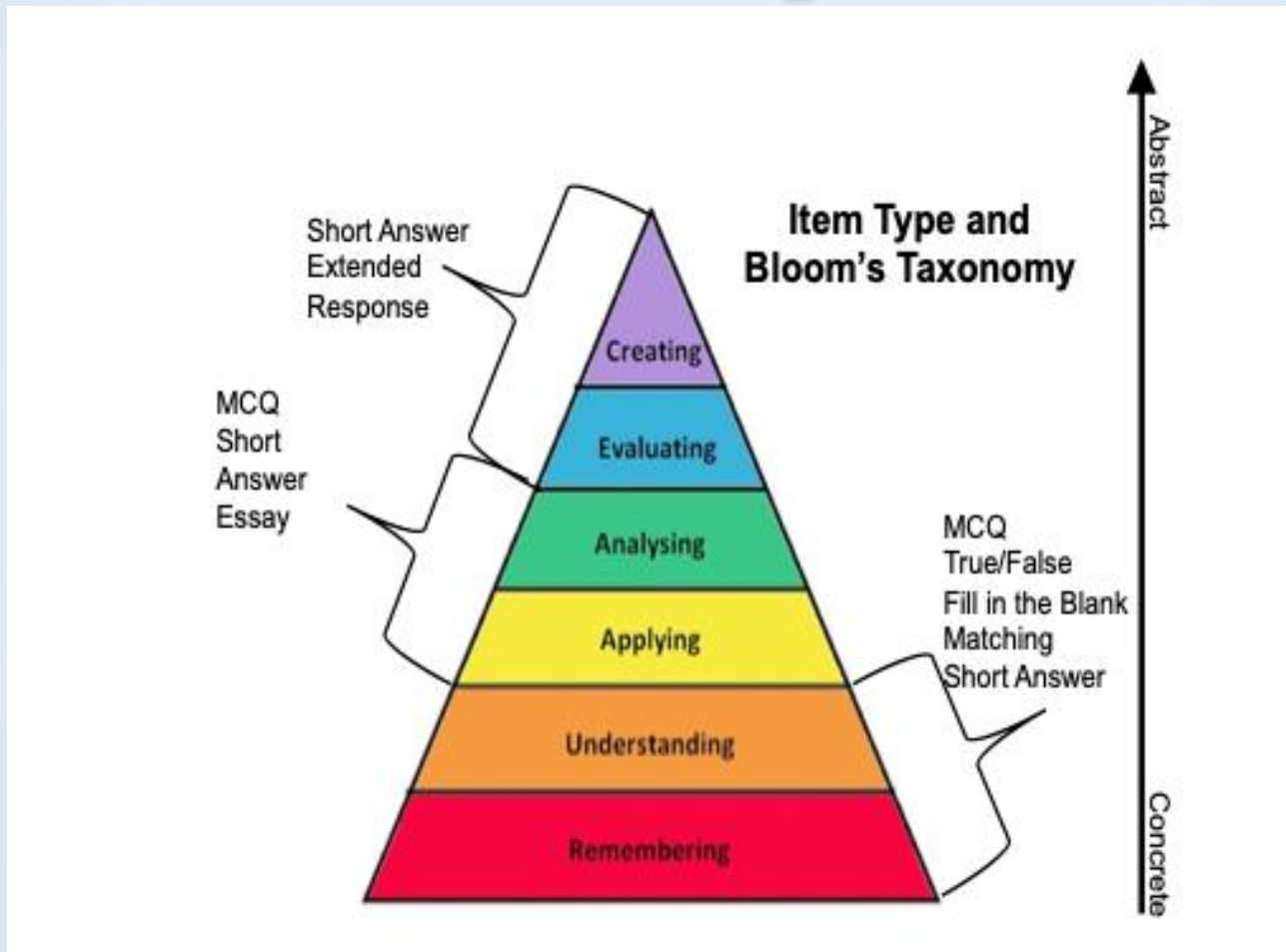
Can understand almost all cultural references and can relate a specific text to other written materials within the culture. Demonstrates a firm grasp of stylistic nuances, irony, and humour.

“Reading comprehension is “the active, automatic, far-transfer process of using one’s internalized language and culture expectancy system to efficiently comprehend an authentic text for the purpose for which it was written.” Dr. R. Clifford

Prototype Test Design

- ❑ Reading is tested via speaking/writing
- ❑ Format: two texts of ca. 1400 words each
- ❑ Six questions per text
- ❑ Administration & Rating:
 - speaking modality – face to face or telephonic (ideally 2 testers/raters)
 - writing modality – 2 raters
 - responses rated as Successful / Partial / Unsuccessful

Constructed Response Items



(Scrock & Coscarelli, 2007) •

Statement of the Problem

Challenges in rating constructed responses:

- ❖ Acceptable answers difficult to foresee
- ❖ Uniqueness of texts which can lead to different interpretations
- ❖ Standard setting to determine acceptable categories of candidates

Retrodictive Modeling Approach (RMA)

The principles underpinning the RMA method lie on the assumption that the ‘judges’ :

- ❖ are well-versed with the STANAG proficiency descriptors;
- ❖ understand the inferences made on the test scores,
- ❖ are able to conceptualize what a threshold performance would look like in real world scenarios.

Retrodictive Modeling Approach (RMA)

- ❖ Raters' conceptual understanding of Level 4 proficiency was triangulated with the patterns which emerged from their analytical ratings.
- ❖ Sessions were conducted in plenary (2016 & 2018) and online using email and Qualtrics surveys.
- ❖ The cut score yielded was further validated through multiple rounds of ratings
- ❖ Qualitative evidence in the form of surveys were collected from the raters.

Research Questions

- ❖ RQ1: *Does the threshold performance, identified through the RMA, accurately reflect the construct of STANAG Level 4 reading prototype test and validity of uses and interpretations that can be made on the basis of the test scores?*
- ❖ RQ2: *How do judges' holistic evaluations correlate with their analytical scores, and how reliable are they in predicting the threshold performances?*

Method

- ❖ Calibration and multiple rounds of rating coded samples
- ❖ Ratings were awarded: an initial holistic score based on the STANAG 6001 Level 4 descriptor, and a detailed analytical score, reported as “Successful”, “Unsuccessful” and “Partially successful.”
- ❖ Subsequently, analytical ratings were awarded (per response) which informed the Final holistic rating
- ❖ Re-rating samples, applying the cut score and confirming final holistic rating

RMA Steps

- ❖ Norming sessions to agree on text and item levels (calibration);
- ❖ Norming sessions to agree on acceptable responses;
- ❖ Matching responses to the descriptor;
- ❖ Identifying patterns of performances at different ranges within the same level (threshold, mid, high);
- ❖ Analyzing raters' scores on the performance grids;
- ❖ Trialing the scoring on actual responses.

RMA-based Rating Method

- ❖ Assisted raters in the qualitative conceptualization of Level 4 reading proficiency IAW STANAG 6001.
- ❖ Enabled the WG to develop *a posteriori* rating criteria and establish the Minimally Acceptable Candidate (MAC).

Methodology

Convenience sampling:

- ❖ 38 participants recruited on voluntary basis from those serving in international posts (e.g., International military staff (Brussels), SHAPE, SACT, Defense institutes in Denmark, Sweden, NL, Germany :

Military N= 32

Civilians: N =6

Native readers: N= 3

Non-native readers: N= 35

Results

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Model Measure	Infit S.E.	Outfit MnSq	Estim. ZStd	Correlation MnSq	Exact Discrm	Agree. PtMea					
PtExp	Obs %	Exp %	N	Raters											
147	144	1.02	1.07	-.53	.12	1.22	1.9	1.37	1.9	.63	.50	.63	59.0	53.7	5 rater 5
134	144	.93	.93	-.35	.12	.88	-1.1	.80	-1.1	1.11	.68	.63	68.6	55.3	2 rater 2
112	144	.78	.69	-.03	.12	.92	-.6	.81	-1.0	1.19	.68	.64	70.1	57.0	3 rater 5
105	144	.73	.62	.08	.12	.82	-1.5	.67	-1.8	1.28	.72	.63	76.2	57.2	4 rater 4
61	144	.42	.25	.84	.14	1.26	1.7	1.17	.6	.74	.47	.58	62.8	55.5	1 rater 1
111.8	144	.78	.71	.00	.13	1.02	.1	.96	-.3		.61				Mean (Count: 5)
29.5	.0	.21	.28	.47	.01	.18	1.5	.26	1.4		.10				S.D. (Population)
33.0	.0	.23	.32	.53	.01	.21	1.7	.29	1.5		.12				S.D. (Sample)

FACETS with three variables: examinees, raters and test items

(Dr. Troy Cox)

Discussion

RQ1: Does the threshold performance, identified through the RMA, accurately reflect the construct of STANAG Level 4 reading prototype test and validity of uses and interpretations that can be made on the basis of the test scores?

- Partial ratings do not contribute significantly to evaluating the responses (psychological factor and non compensatory scale)
- Double rating is fundamental
- Most initial ratings were revised after analytical rating

Discussion

RQ2: How do judges' holistic evaluations correlate with their analytical scores and how reliable are they in predicting the threshold performances?

- ❖ Pre and post ratings show there is a high inter and intra-rater reliability between initial holistic ratings and final holistic ratings when the cut score identified through the RMA was applied.
- ❖ Consistency within and among the raters, and confirms raters' conceptualization of STANAG Level 4 reader,
- ❖ Given the high stakes, it is a proven best practice in test development and administration to always have two raters who agree on a rating.

Limitations

- ❖ Small sample size (N=38)
- ❖ More research into consequential validity, scoring validity and concurrent validity studies (reading to write; reading to speak);
- ❖ Minimal combinations of S, U and P ratings, analyzed in conjunction with the consistent ratings of initial and final holistic ratings would then yield a cut score.

Conclusions

- ❖ Importance of having expert raters on subjectively-scored responses;
- ❖ Two raters, if not more in case of discrepancy in judgment;
- ❖ Rating should initially be done individually and then compared to a second rater's evaluation.
- ❖ Analytical scoring contributes to guide the rater as to whether the final rating should be changed from the initial rating.
- ❖ A holistic rating alone does not seem to consistently predict the performance of the sample.

Publication

DiBiase-Lubrano, M.J.& Vasilj- Begovic, J. (2021). Rethinking the rating process: solution to the threshold performance Dilemma. *Journal of Distinguished Language Studies*, 7, 2—40.

<https://msipress.com/journal-for-distinguished-language-studies/>

https://www.amazon.com/dp/1950328856/ref=sr_1_1?dchild=1&keywords=journal+for+distinguished+language+studies&qid=1620410623&sr=8-1

Coupon code: ad40 (40% discount)

Acknowledgments

Peggy Garza (U.S. Marshall Center, Germany);

Keith Wert (U.S. Marshall Center, Germany), BILC secretary and Chair at the time of this project.

Special Thanks to

Drs. Ray Clifford & Troy Cox (BYU, USA) for their support, advice and technical expertise throughout the process.

Working Group Members

Ms Petya Georgieva (National Defense College, BLG);

Ms Nancy Powers & Dr. May Tan (Military Personnel Generation Group, St., CAN);

Mr. Käre Kildevang and Dr. Allen Christiansen (Royal Danish Defense College, DNK);

Dr. Donald Sturges (Bundessprachennamt, GER);

Mr. Gerard Seinhorst (National Defense Language Center, NL);

Ms Annette Nolan & Mr. Keith Farr (Swedish Defense College, SWE), Mr. David Oglesby (U.S. Marshall Center, Germany), and Dr. Martha Herzog (DLIFLC, USA);

References

- Alschuler, C. & Moussa, N. (2002). *Testing reading at ILR Levels 4, 4+ and 5: A Tester Training Model*. Presentation to the Interagency Language Roundtable Committee, Foreign Service Institute, Washington, D.C.
- Child, J.R. (1998) Language skill levels, textual modes and the rating process. *Foreign Language Annals 3*: 381–397.
- Clifford
- DiBiase-Lubrano, M.J. & Vasilji-Begovic-J. (2021). Rethinking the rating process: solution to the threshold dilemma problem. *Journal of Distinguished Language Studies*, 7, 20-40.
- Edwards, A. L. (1996). Reading proficiency assessment and the ILR/American Council on the teaching of foreign languages text typology: A reevaluation. *Modern Language Journal 80*: 350–361.
- Shrock, S. & Coscarelli, W. (2007). *Criterion-referenced test development. Technical and legal guidelines for corporate training (3rd edition)*. Internet: Pfeiffer & Company.