Variations in setting cut scores: How comparable are cut scores across media, methods, time, and instrument length?

Charalambos (Harry) Kollias

17th EALTA Conference – Variations on Common Standards, Online Conference, 4th June, 2021

Standard setting



Definition

1. **Decision-making process**
   - establishing cut scores

2. **Task**
   - pre-set activities to facilitate judges' decision-making process

Challenges: logistics









consequences

consequences

cut score studies …

- not being conducted

  (Tannenbaum, 2013)

- not being replicated

  (Dunlea & Figueras, 2012)

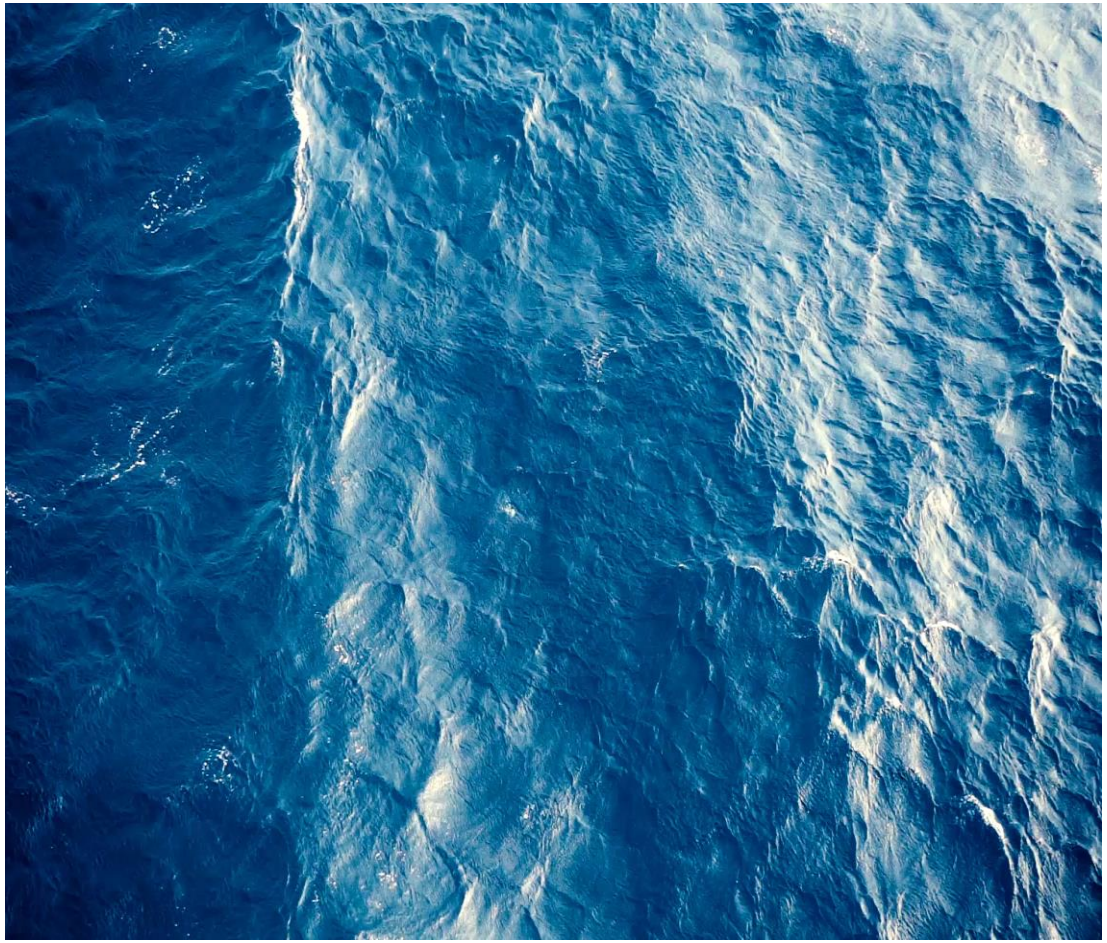
solution

virtual standard setting





The sea of virtual standard setting (VSS)





The chartered VSS waters

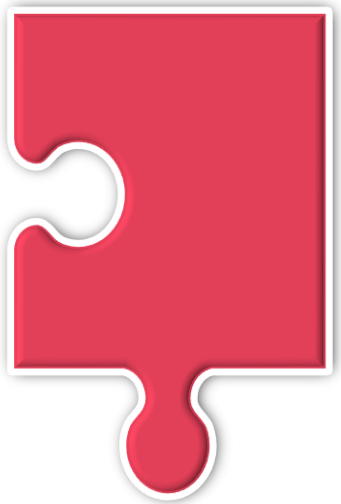
few empirical studies

results comparable with *F2F* (Katz & Tannenbaum; 2014)

feasible in **series of smaller sessions** …

… through different media:

- emails (Harvey & Way, 1999);
- audio-conferencing (Katz & Tannenbaum; 2014);
- combination of audio & video (Katz & Tannenbaum. 2009)].



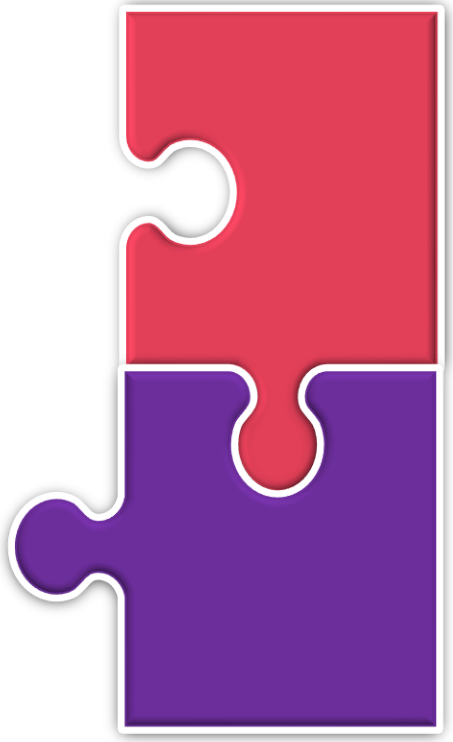
… in asynchronous environments

… in combined asynchronous & synchronous environments

The chartered VSS waters cont.

synchronous full-day workshop feasible;

both audio and video virtual environments equally appropriate for setting reliable & defensible cut scores;



cut scores comparable, replicable, and interchangeable;

virtual environments not hindering communication.

judges' showed preference towards video.

judges' decision making processes not hindered by virtual environments;

(Kollias, 2018)

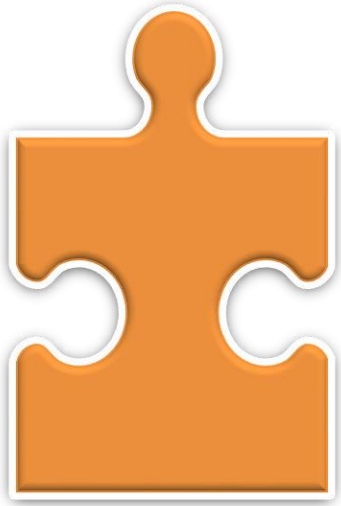
Another piece ...

Comparability of cut scores across

… modes and media (audio, video, and *FTF*)

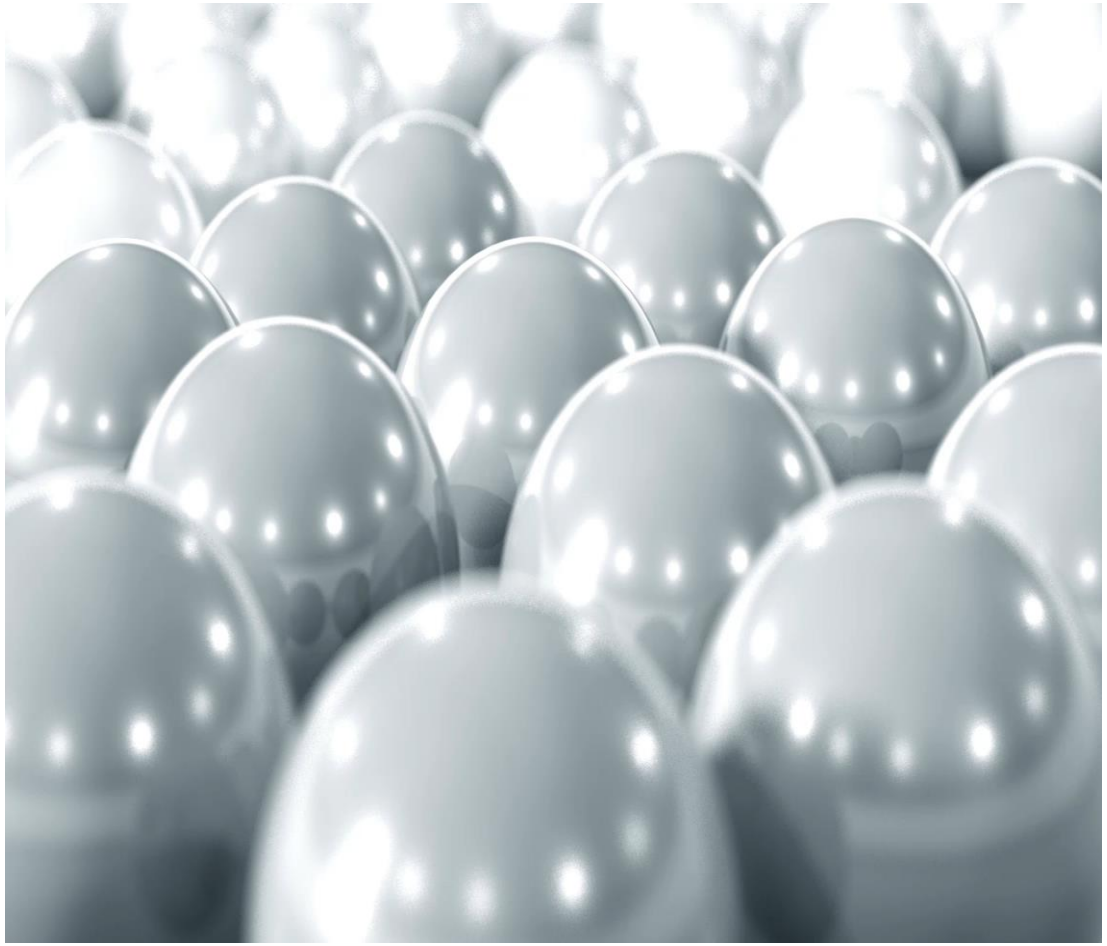
… time (2011 vs. 2014)

… instrument length (item subset vs. full set)



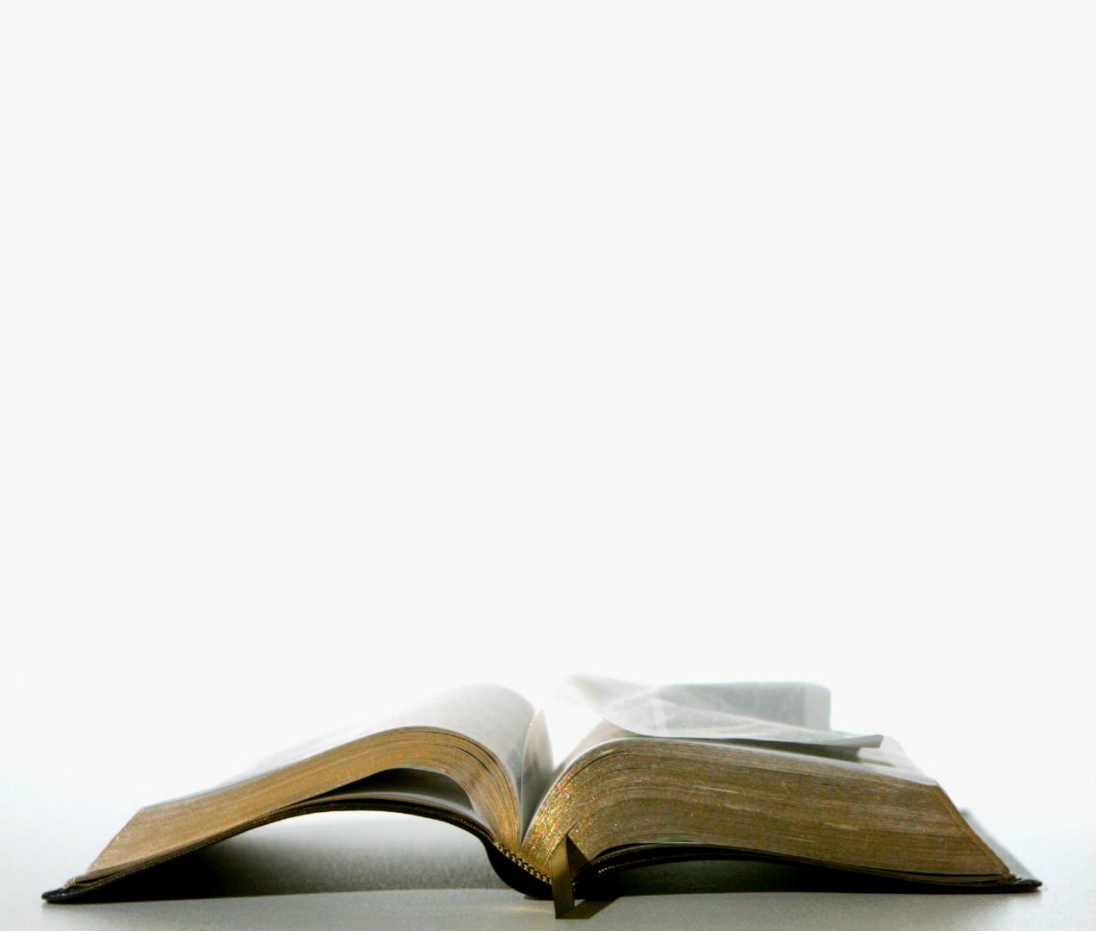
… methods (e.g. Modified Angoff , Yes/No Angoff, Extended Angoff)

Reviewing the literature: item subsets & Angoff methods





Literature review



Subset of items

- approx. 50% of operational items for comparability within 1 SE (Ferdous & Plake, 2007)
- alternative item selection (Ferdous & Smith, 2008)
- randomly selected 50% - differences in second round estimates (Buckendahl *et al*. 2010)
- cut scores means not significantly different (Kannan *et al*., 2015)

Background

2011 – F2F (Kollias, 2012)

Panellists: 15

Group: 1 [G5 (subset), G6 (original)]

Instrument: B1 English language test – Grammar, Vocabulary, & Reading (GVR)

Form: A

Items: 75 (original)

- 25 Grammar items
- 25 Vocabulary items
- 25 Reading items

SS Method: Modified Angoff

Rounds: 2

2014 – Virtual (Kollias, 2017)

Panellists: 45

Groups: 4 (G1, G2, G3, G4)

Instrument: 2 ***** B1 English language tests – Grammar, Vocabulary, & Reading (GVR)

Forms: A & B (equated)

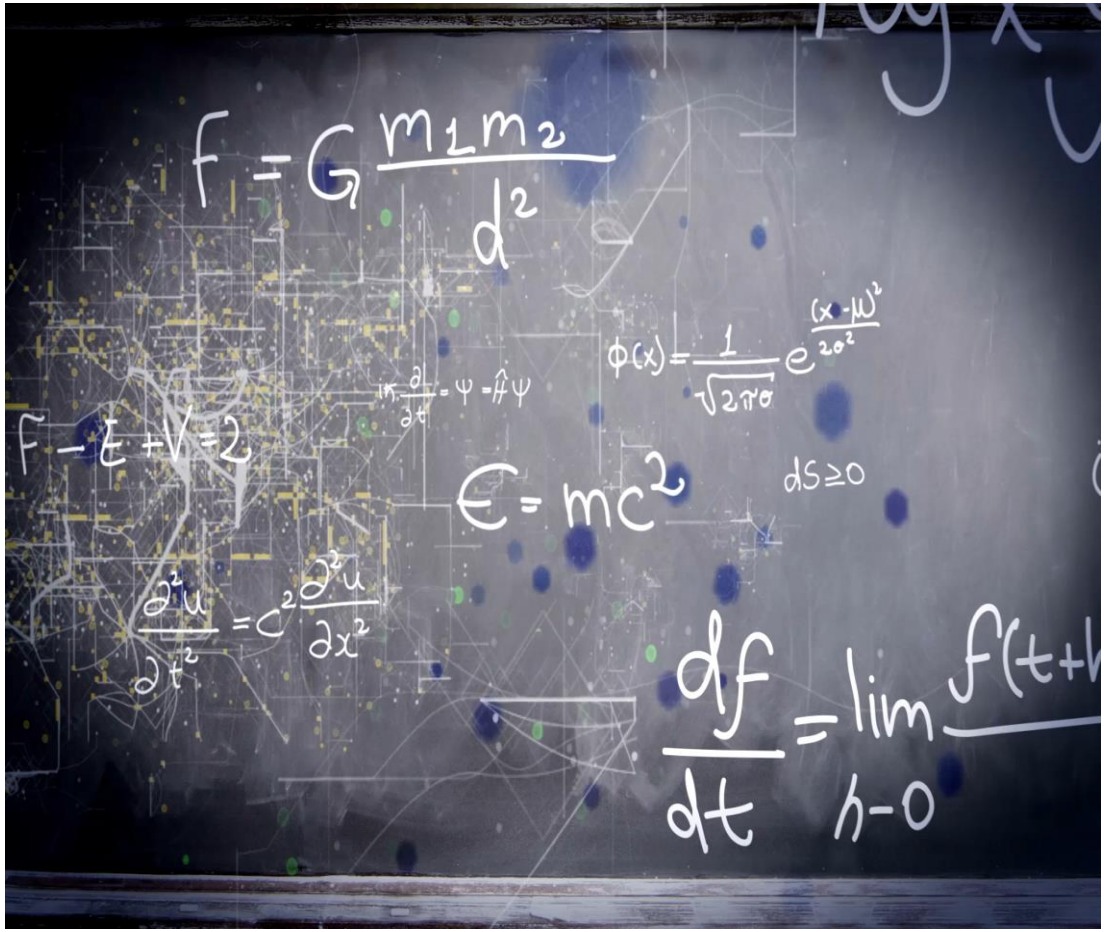
Items: 45 (subset of items)

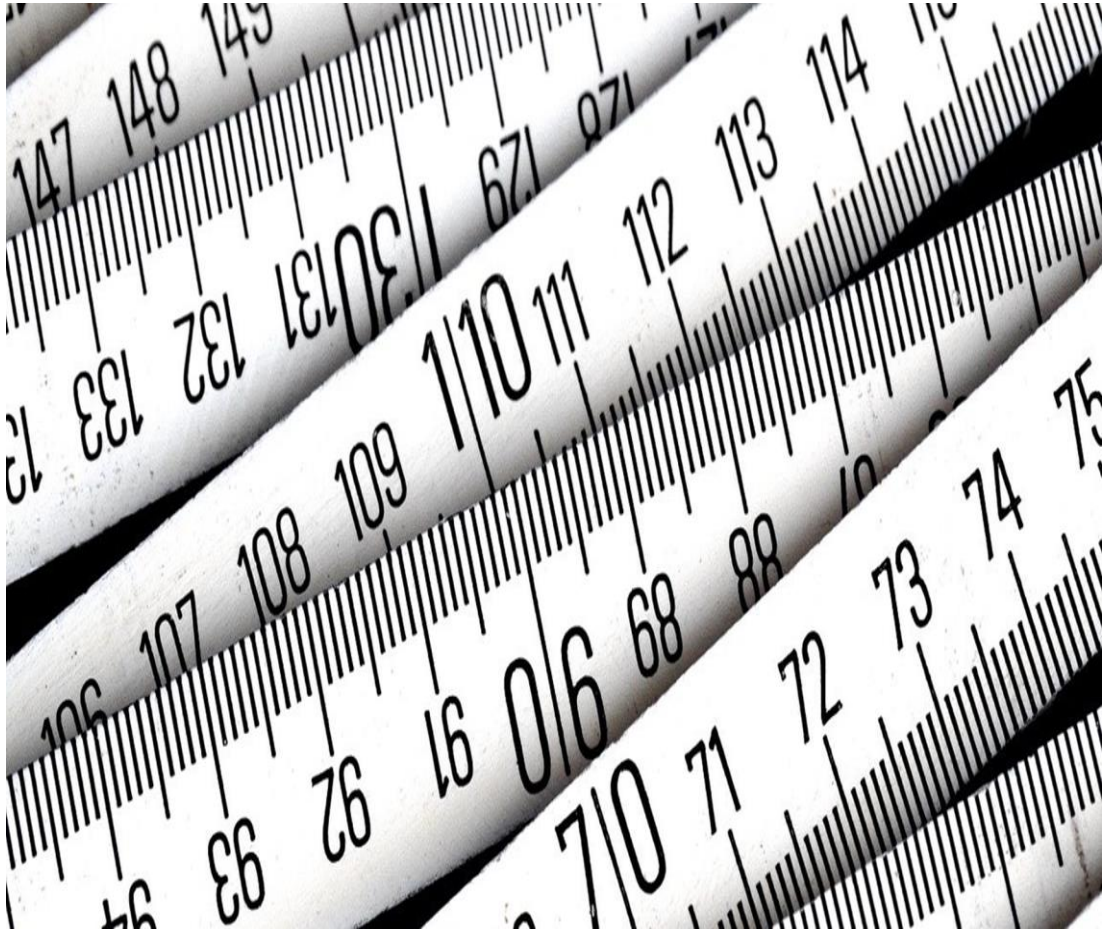
- 15 Grammar items
- 15 Vocabulary items
- 15 Reading items

SS Methods: Yes/No Angoff & Extended Angoff

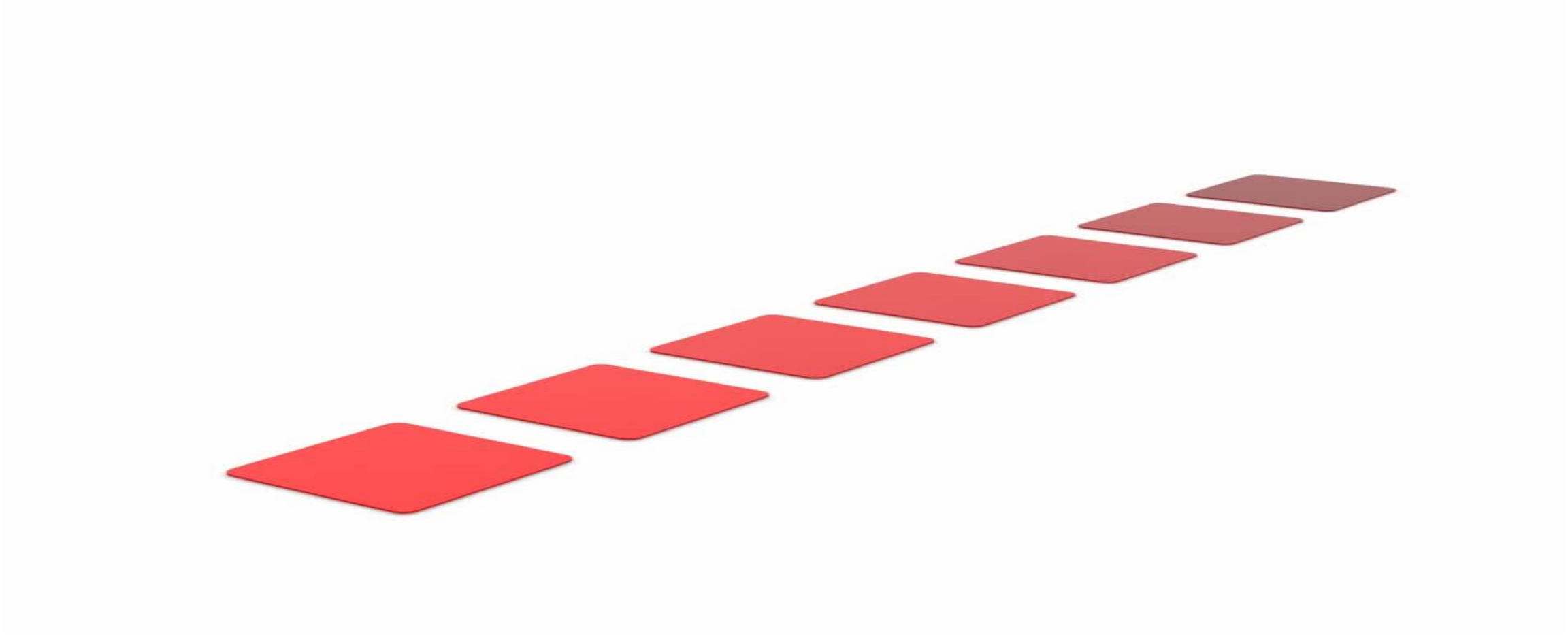
Rounds: 3

Model: Many-facet Rasch Measurement (MFRM)





Cut score measures



Round 1: Test Form A

Measr	+Judges G1 (AUDIO)	+Judges G2 (VIDEO)	+Judges G3 (VIDEO)	+Judges G4 (AUDIO)	+Judges G5 (F2F Subset)	+Judges G6 (F2F Full set)
2	+ More demanding	+ More demanding	+ More demanding	+ More demanding	+ More demanding	+ More demanding
				PA37G4R1		
	PA07G1R1	PA22G2R1			PA56G5R1	PA71G6R1
			PA25G3R1 PA26G3R1	PA42G4R1		
1	+	+	+ PA34G3R1	+ PA39G4R1	+ PA55G5R1	+
		PA19G2R1		PA38G4R1 PA46G4R1		PA70G6R1
	PA01G1R1	PA15G2R1 PA17G2R1	PA23G3R1	PA36G4R1 PA41G4R1 PA43G4R1		
		PA12G2R1	PA28G3R1		PA48G5R1	PA63G6R1 PA72G6R1
	PA02G1R1	PA13G2R1	PA32G3R1 PA35G3R1		PA50G5R1 PA54G5R1 PA57G5R1	PA65G6R1 PA69G6R1
	PA08G1R1 PA09G1R1	PA11G2R1	PA29G3R1	PA40G4R1	PA53G5R1	
	PA03G1R1	PA20G2R1	PA24G3R1 PA31G3R1 PA33G3R1	PA45G4R1	PA61G5R1	PA76G6R1
		PA10G2R1				PA64G6R1 PA68G6R1
* 0	* PA05G1R1	*	*	*	* PA49G5R1 PA58G5R1 PA59G5R1	* PA73G6R1 *
	PA06G1R1		PA27G3R1		PA52G5R1	PA74G6R1
	PA04G1R1	PA16G2R1 PA18G2R1 PA21G2R1		PA44G4R1		PA67G6R1
					PA51G5R1 PA60G5R1	PA62G6R1 PA66G6R1 PA75G6R1
					PA47G5R1	
		PA14G2R1				
-1	+	+	+	+	+	+
	Less demanding	Less demanding	Less demanding	Less demanding	Less demanding	Less demanding
Measr	+Judges G1 (AUDIO)	+Judges G2 (VIDEO)	+Judges G3 (VIDEO)	+Judges G4 (AUDIO)	+Judges G5 (F2F Subset)	+Judges G6 (F2F Full set)

Round 1: Test Form B

Measr	+Judges G1 (VIDEO)	+Judges G2 (AUDIO)	+Judges G3 (AUDIO)	+Judges G4 (VIDEO)
2	+ More demanding	+ More demanding	+ More demanding	+ More demanding
			PB26G3R1	
	PB08G1R1			
			PB25G3R1	
		PB13G2R1		PB46G4R1
1	+ PB01G1R1	+ PB22G2R1	+	+
		PB17G2R1		
		PB10G2R1	PB32G3R1 PB34G3R1	PB41G4R1
				PB43G4R1
	PB07G1R1	PB11G2R1		PB37G4R1 PB42G4R1
			PB31G3R1	PB39G4R1
		PB15G2R1 PB16G2R1 PB19G2R1 PB20G2R1	PB28G3R1 PB35G3R1	
	PB02G1R1 PB06G1R1 PB09G1R1		PB33G3R1	
* 0	* PB05G1R1	* PB12G2R1 PB14G2R1	* PB23G3R1 PB24G3R1 PB27G3R1	* PB40G4R1 PB44G4R1 PB45G4R1 *
	PB03G1R1	PB21G2R1	PB29G3R1	
		PB18G2R1		PB36G4R1 PB38G4R1
	PB04G1R1			
-1	+	+	+	+
	Less demanding	Less demanding	Less demanding	Less demanding
Measr	+Judges G1 (VIDEO)	+Judges G2 (AUDIO)	+Judges G3 (AUDIO)	+Judges G4 (VIDEO)

Round 2: Test Form A

Measr	Judges G1 (AUDIO)	Judges G2 (VIDEO)	Judges G3 (VIDEO)	Judges G4 (AUDIO)	Judges G5 (F2F Subset)	Judges G6(F2F full)
2	More demanding	More demanding	More demanding	More demanding	More demanding	More demanding
			PA26G3R2			
			PA34G3R2			
	PA08G1R2					
		PA15G2R2 PA17G2R2				
		PA22G2R2		PA37G4R2	PA48G5R2	PA63G6R2 PA71G6R2
1	PA07G1R2	PA11G2R2 PA19G2R2	PA27G3R2	PA46G4R2	PA56G5R2	
			PA25G3R2	PA42G4R2		
		PA13G2R2 PA14G2R2 PA20G2R2 PA21G2R2	PA29G3R2 PA32G3R2		PA55G5R2	PA70G6R2
	PA01G1R2			PA38G4R2 PA39G4R2 PA40G4R2		
	PA09G1R2		PA23G3R2	PA41G4R2	PA50G5R2 PA61G5R2	PA65G6R2 PA76G6R2
	PA03G1R2 PA06G1R2		PA24G3R2 PA28G3R2 PA33G3R2		PA53G5R2 PA54G5R2 PA57G5R2	PA69G6R2 PA72G6R2
	PA02G1R2 PA04G1R2 PA05G1R2	PA18G2R2	PA31G3R2 PA35G3R2	PA43G4R2 PA45G4R2	PA58G5R2	
		PA12G2R2		PA44G4R2	PA59G5R2	PA68G6R2 PA73G6R2
0		PA10G2R2			PA49G5R2	PA64G6R2 PA74G6R2
					PA52G5R2	
		PA16G2R2		PA36G4R2	PA47G5R2	PA67G6R2
					PA60G5R2	PA62G6R2
						PA75G6R2
					PA51G5R2	
						PA66G6R2
-1						
	Less demanding	Less demanding	Less demanding	Less demanding	Less demanding	Less demanding
Measr	Judges G1 (AUDIO)	Judges G2 (VIDEO)	Judges G3 (VIDEO)	Judges G4 (AUDIO)	Judges G5 (F2F Subset)	Judges G6(F2F full)

Round 2: Test Form B

Measr	+Judges G1 (VIDEO)	+Judges G2 (AUDIO)	+Judges G3 (AUDIO)	+Judges G4 (VIDEO)
2	+ More demanding	+ More demanding	+ More demanding	+ PB46G4R2 More demanding
	PB08G1R2			
		PB13G2R2		
			PB34G3R2	
1	+	+	+ PB23G3R2 PB25G3R2	+
			PB24G3R2 PB27G3R2	PB41G4R2 PB43G4R2
		PB10G2R2 PB11G2R2 PB14G2R2 PB17G2R2	PB26G3R2	PB37G4R2 PB42G4R2 PB44G4R2
			PB31G3R2	
	PB01G1R2 PB07G1R2			
	PB02G1R2 PB06G1R2	PB12G2R2 PB19G2R2 PB20G2R2 PB21G2R2 PB22G2R2	PB32G3R2 PB33G3R2	
		PB15G2R2	PB29G3R2 PB35G3R2	PB40G4R2 PB45G4R2
	PB03G1R2 PB04G1R2 PB05G1R2 PB09G1R2	PB18G2R2		
* 0	*	*	*	* PB39G4R2 *
		PB16G2R2		PB36G4R2
			PB28G3R2	PB38G4R2
-1	+	+	+	+
	Less demanding	Less demanding	Less demanding	Less demanding
Measr	+Judges G1 (VIDEO)	+Judges G2 (AUDIO)	+Judges G3 (AUDIO)	+Judges G4 (VIDEO)

Round 2: Pairwise interactions

Group	Group	Sig.
G1-A	G1-V	No
G1-A	G2-A	No
G1-A	G3-A	No
G1-A	G4-A	No
G1-A	G5-F2F	No
G1-A	G6-F2F	No
G1-V	G2-V	No
G1-V	G3-V	No
G1-V	G4-V	No
G1-V	G5-F2F	No
G1-V	G6-F2F	No

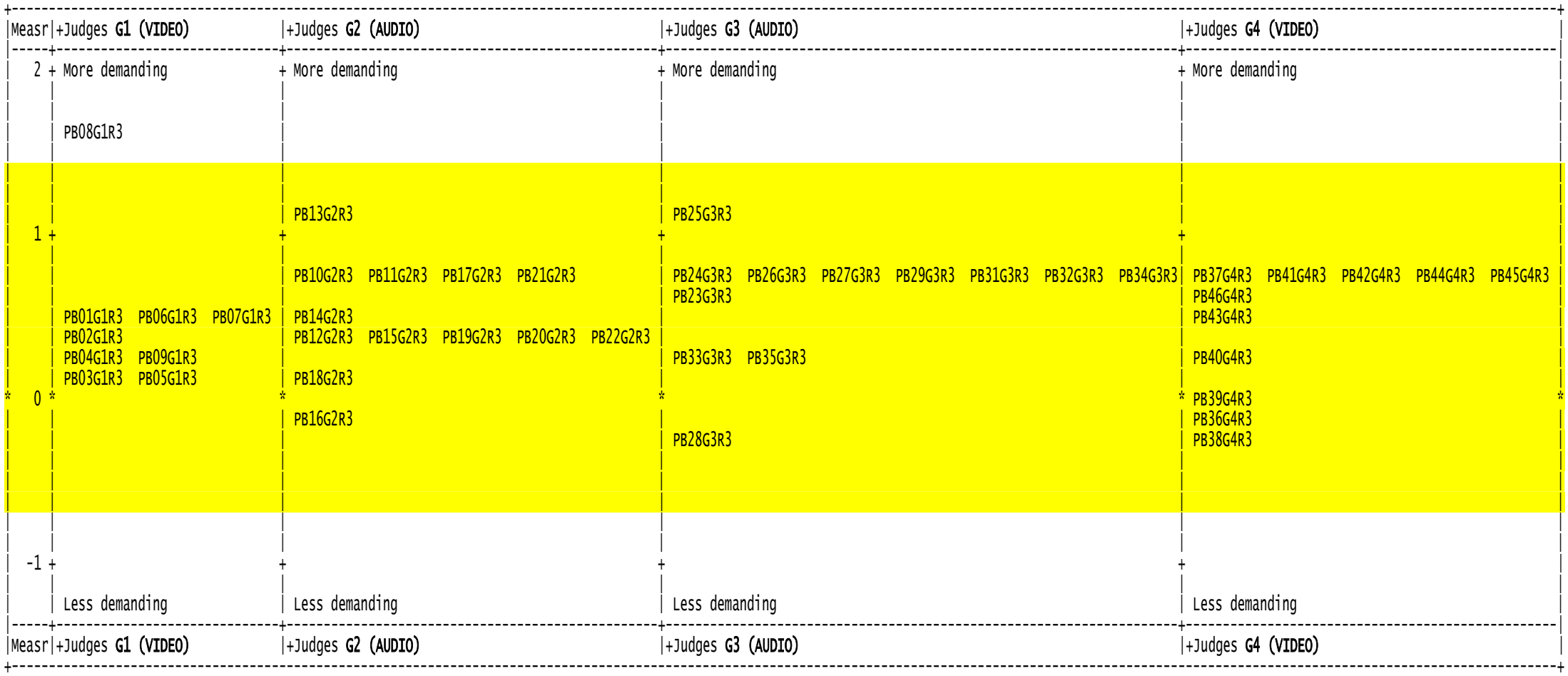
Group	Group	Sig.
G2-A	G2-V	No
G2-A	G3-A	No
G2-A	G4-A	No
G2-A	G5-F2F	No
G2-A	G6-F2F	No
G2-V	G3-V	No
G2-V	G4-V	No
G2-V	G5-F2F	No
G2-V	G6-F2F	No
G3-A	G3-V	No
G3-A	G4-A	No

Group	Group	Sig.
G3-A	G5-F2F	No
G4-A	G6-F2F	No
G3-V	G4-V	No
G3-V	G5-F2F	No
G3-V	G6-F2F	No
G4-A	G4-V	No
G4-A	G5-F2F	No
G4-A	G6-F2F	No
G4-V	G5-F2F	No
G4-V	G6-F2F	No
G5-F2F	G6-F2F	No

Round 3: Test Form A & F2F Round 2

Measr	+Judges G1 (AUDIO)	+Judges G2 (VIDEO)	+Judges G3 (VIDEO)	+Judges G4 (AUDIO)	+Judges G5 (F2F Subset)	+Judges G6(F2F full)
2	More demanding	More demanding	More demanding	More demanding	More demanding	More demanding
	PA08G1R3					
			PA34G3R3			
			PA25G3R3			
			PA26G3R3		PA48G5R2	PA63G6R2 PA71G6R2
1					PA56G5R2	
		PA10G2R3	PA27G3R3 PA32G3R3	PA37G4R3 PA42G4R3		
	PA07G1R3	PA11G2R3 PA13G2R3 PA15G2R3 PA19G2R3	PA23G3R3 PA24G3R3 PA29G3R3 PA31G3R3 PA35G3R3	PA46G4R3	PA55G5R2	PA70G6R2
	PA01G1R3 PA04G1R3 PA06G1R3	PA17G2R3 PA20G2R3	PA28G3R3	PA38G4R3 PA39G4R3 PA40G4R3 PA41G4R3		
	PA09G1R3				PA50G5R2 PA61G5R2	PA65G6R2 PA76G6R2
		PA14G2R3 PA16G2R3 PA21G2R3 PA22G2R3		PA43G4R3	PA53G5R2 PA54G5R2 PA57G5R2	PA69G6R2 PA72G6R2
	PA02G1R3 PA03G1R3 PA05G1R3	PA12G2R3 PA18G2R3	PA33G3R3	PA45G4R3	PA58G5R2	
				PA44G4R3	PA59G5R2	PA68G6R2 PA73G6R2
0					PA49G5R2	PA64G6R2 PA74G6R2
					PA52G5R2	
				PA36G4R3	PA47G5R2	PA67G6R2
					PA60G5R2	PA62G6R2
						PA75G6R2
					PA51G5R2	
						PA66G6R2
-1						
	Less demanding	Less demanding	Less demanding	Less demanding	Less demanding	Less demanding
Measr	+Judges G1 (AUDIO)	+Judges G2 (VIDEO)	+Judges G3 (VIDEO)	+Judges G4 (AUDIO)	+Judges G5 (F2F Subset)	+Judges G6(F2F full)

Round 3: Test Form B



Round 3 Pairwise interactions: (F2F Round 2)

Group	Group	Sig.
G1-A	G1-V	No
G1-A	G2-A	No
G1-A	G3-A	No
G1-A	G4-A	No
G1-A	G5-F2F	No
G1-A	G6-F2F	No
G1-V	G2-V	No
G1-V	G3-V	No
G1-V	G4-V	No
G1-V	G5-F2F	No
G1-V	G6-F2F	No

Group	Group	Sig.
G2-A	G2-V	No
G2-A	G3-A	No
G2-A	G4-A	No
G2-A	G5-F2F	No
G2-A	G6-F2F	No
G2-V	G3-V	No
G2-V	G4-V	No
G2-V	G5-F2F	No
G2-V	G6-F2F	No
G3-A	G3-V	No
G3-A	G4-A	No

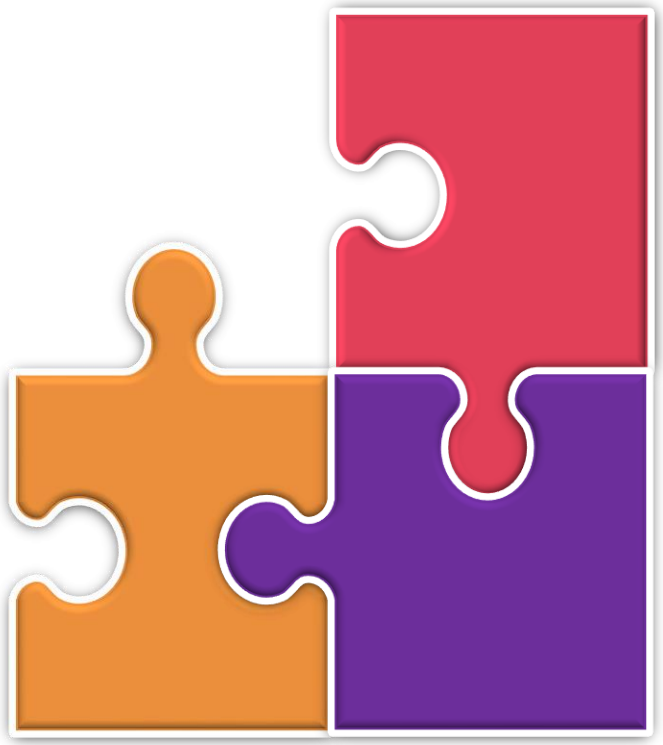
Group	Group	Sig.
G3-A	G5-F2F	No
G4-A	G6-F2F	No
G3-V	G4-V	No
G3-V	G5-F2F	Yes
G3-V	G6-F2F	Yes
G4-A	G4-V	No
G4-A	G5-F2F	No
G4-A	G6-F2F	No
G4-V	G5-F2F	No
G4-V	G6-F2F	No
G5-F2F	G6-F2F	No

Findings

Within rounds, differences in mean measures between …

not statistically significant

Group 3 idiosyncrasies more likely …



Round 2 F2F vs. Round 3 Virtual, differences in mean measures …

statistically significant Group 3-Video only

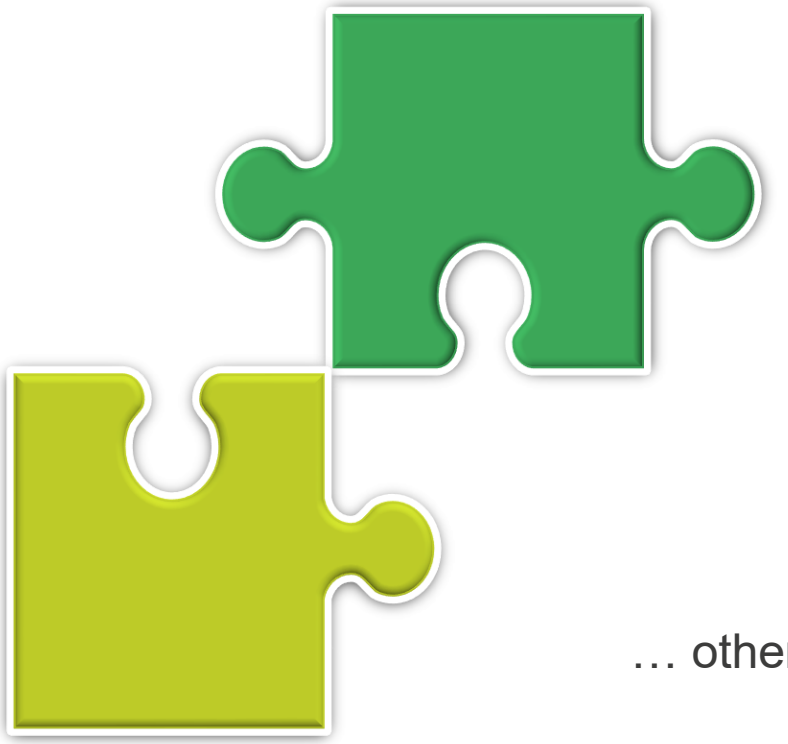
The unchartered VSS waters

Need to investigate …

… other skills:

- listening
- speaking
- writing

… discussion in virtual environment



… other CEFR levels

… other SS methods

References

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalised examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education,* 12(4), 343-366. doi: 10.1207/S15324818AME1204_2

Buckendahl, C. W., Ferdous, A. A., Gerrow, J. (2010). *Recommending cut scores with a subset of items: An empirical illustration Practical Assessment, Research & Evaluation*, 16(6). doi:10.7275/tv3s-cz6

Dunlea, J., & Figueras, N. (2012). Replicating results from a CEFR test comparison project across continents. In D. Tsagari, & C. Ildikó (Eds.), *Collaboration in language testing and assessment* (Vol. 26, pp. 31-45). Frankfurt: Peter Lang.

Ferdous, A. A, Plake, S. B. (2005). The use of subsets of test questions in an Angoff standard-setting method. *Educational and Psychological Measurement*, 65(2), 185-201. doi:10.1177%2F0013164404264852

Ferdous, A. & Smith, R. W. (2008, March). Considerations for using subsets of items for standard setting. Paper presented at the annual meeting of the National Council on Measurement in Education. New York, NY.

Harvey, A. L., & Way, W. D. (1999). *A comparison of web-based standard setting and monitored standard setting.* Montreal: Paper presented at the annual meeting of the National Council of Measurement in Education.

References cont.

Katz, I. R., & Tannenbaum, R. J. (2014). Comparison of web-based and face-to-face standard setting using the Angoff method. *Journal of Applied Testing Technology, 15*(1), 1-17. Retrieved from www.jattjournal.com/index.php/atp/article/view/52751

Katz, I. R., Tannenbaum, R. J., & Kannan, P. (2009). Virtual standard setting. *Clear Exam Review, 20*(2), 19-27.

Kollias, C. (2012). *(BCCE) examination: Setting a Common European Framework of Reference (CEFR) B1 cut score. Technical Report.* Hellenic American University.

Kollias, C. (2017). Virtual standard setting in language testing: Exploring the feasibility and appropriateness of setting cut scores in synchronous audio and audio-visual environments [Unpublished Ph.D. thesis*]. Lancaster University.

Kollias, C. (2018). Virtual standard setting in language testing: Exploring the use of synchronous audio and audio-visual environments. Bochum: Paper presented at the 15th EALTA Conference.

Tannenbaum, R. J. (2013). Setting standards on the TOEIC(R) listening and reading test and the TOEIC(R) speaking and writing tests: A recommended procedure. In *The research foundation for the TOEIC tests: A compendium of studies* (Vol. II, pp. 8. 1-8.12). Princeton: Educational Testing Service.

* Revised version with F2F to be published soon by Peter Lang

Thank You

The opinions expressed in this presentation are those of the presenter (Charalambos Kollias). They do not purport to reflect the *opinions* or *views* of the NFER or its members.

1946–2021 75 YEARS

NFER
National Foundation for Educational Research

Evidence for excellence in education

© **National Foundation for Educational Research 2021**

All rights reserved. No part of this document may be reproduced or transmitted in any form or by any means, electronic, mechanical, photocopying, or otherwise, without prior written permission of NFER.

The Mere, Upton Park, Slough, Berks SL1 2DQ
T: +44 (0)1753 574123 • **F:** +44 (0)1753 691632 • enquiries@nfer.ac.uk

www.nfer.ac.uk