Setting cut scores and evaluating standard setting judgments through the Many-Facet Rasch Measurement (MFRM) model

Charalambos (Harry) Kollias, Oxford University Press

Paraskevi (Voula) Kanistra, Trinity London College

13th Annual UK Rasch User Group Meeting, 21-03-19, Cambridge

"… the Rasch measurement approach basically construes raters or judges as individual experts, … It may thus be reasonable not to perform MFRM analyses in the later stages of standard setting where judges can be assumed to gravitate toward the group mean."

(Eckes, 2015 p.163)

questions

Q1: Do judges change their ratings across rounds? If yes, to what extent?

Q2: What do judges claim mainly influences their ratings?

Q3: Can we use MFRM to analyse Round 2 & Round 3 ratings?

Q4: Do judges remain independent experts across rounds?

Q5: What do we gain from MFRM analysis of standard setting data?

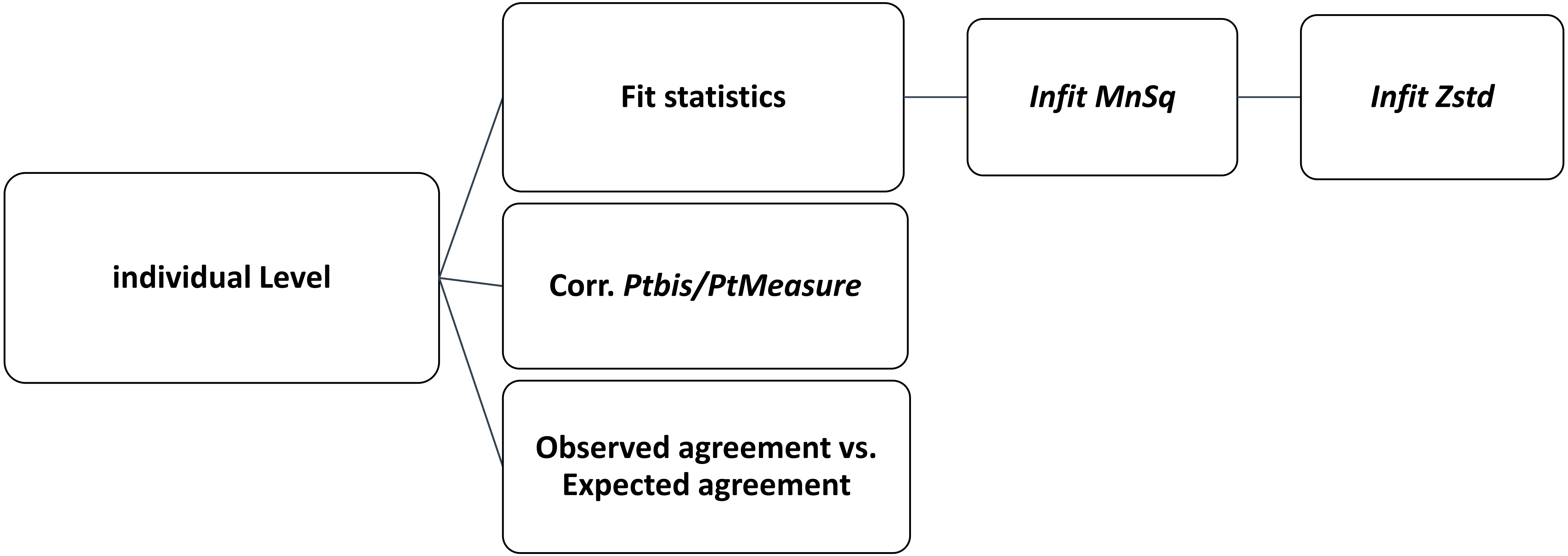
standard setting (SS) workshop stages

introduction stage	orientation stage	training in the method stage	judgment stage(s)
• welcoming & Introductions	• SS overview • familiarisation with CEFR • familiarisation with test instrument	• training & practice	• Round 1 judgments, feedback & discussion • Round 2**,** (empirical data), judgments, & feedback • Round 3 judgments, & feedback (when applicable)

TRINITY COLLEGE LONDON

OXFORD UNIVERSITY PRESS

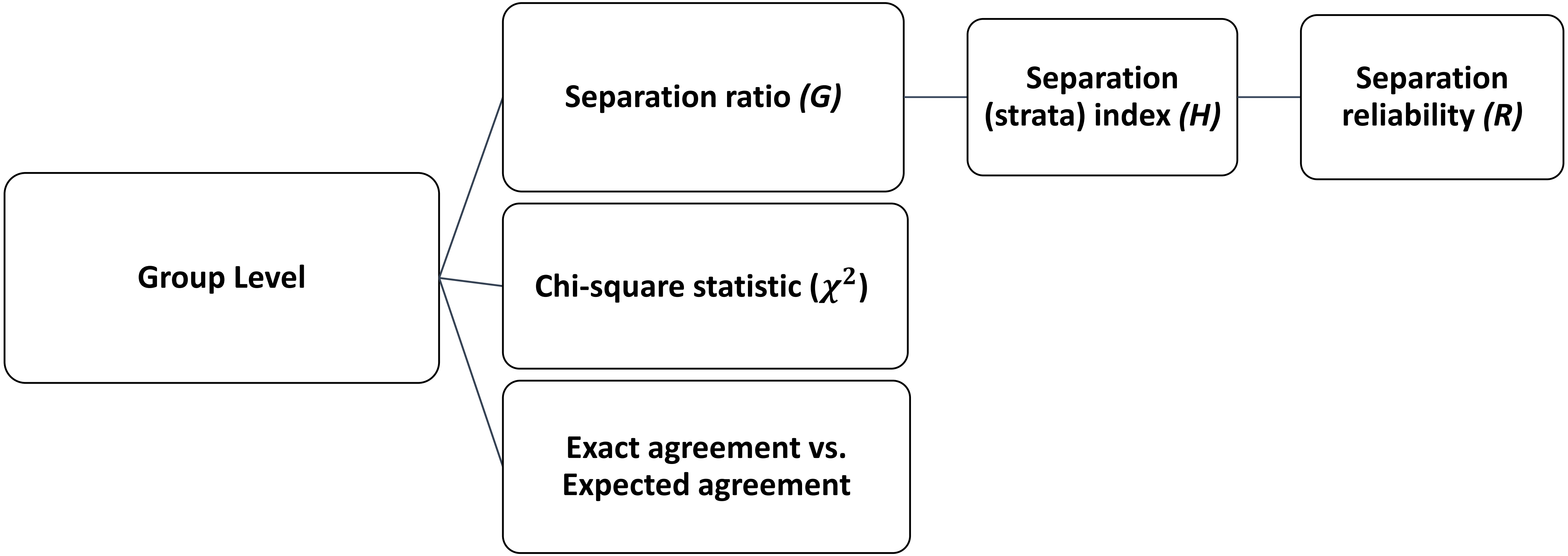
consistency evaluation framework (individual)



individual Level

- **Fit statistics**
  - ***Infit MnSq***
    - ***Infit Zstd***
- **Corr. *Ptbis/PtMeasure***
- **Observed agreement vs. Expected agreement**

consistency evaluation framework (group)



Group Level
Separation ratio *(G)*
Separation (strata) index *(H)*
Separation reliability *(R)*
Chi-square statistic (χ^2)
Exact agreement vs. Expected agreement

study 1

Participants

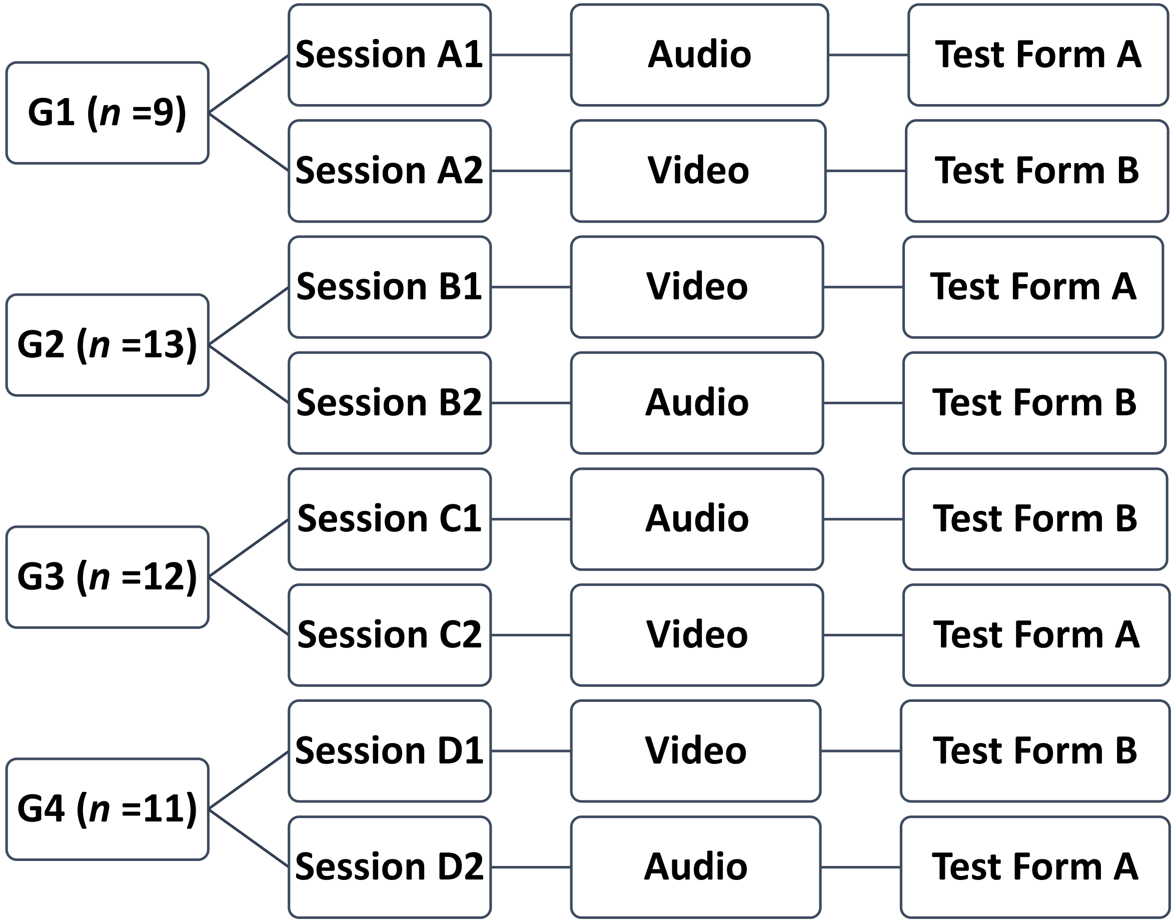
- 45 judges - 4 groups (G1 – G4)

SS Method

Yes/No Angoff (3 Rounds)

Instruments

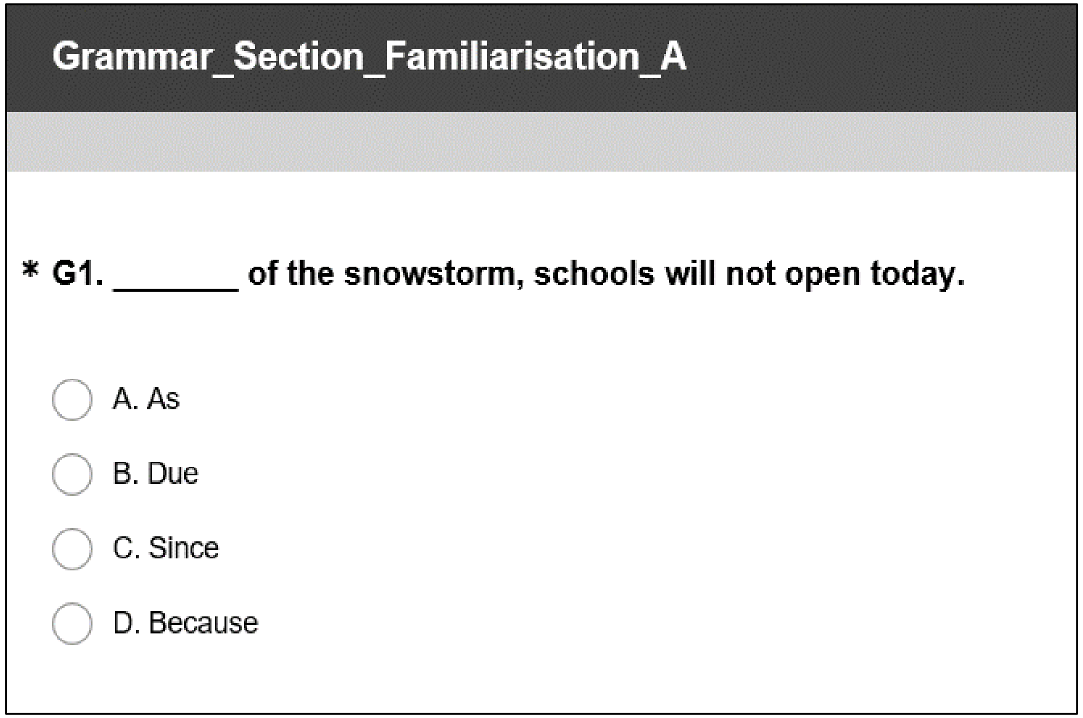
- 2 Rasch equated B1 (GVR) multiple choice test
  - Form A & Form B
- 45 items per instrument
  - 15 grammar (discrete)
  - 15 vocabulary (discrete)
  - 15 reading items 3 passages X 5 items



G1 (*n* =9)	**Session A1**	**Audio**	**Test Form A**
	Session A2	**Video**	**Test Form B**
G2 (*n* =13)	**Session B1**	**Video**	**Test Form A**
	Session B2	**Audio**	**Test Form B**
G3 (*n* =12)	**Session C1**	**Audio**	**Test Form B**
	Session C2	**Video**	**Test Form A**
G4 (*n* =11)	**Session D1**	**Video**	**Test Form B**
	Session D2	**Audio**	**Test Form A**

example of familiarisation & R1 rating form

Test familiarisation



Grammar_Section_Familiarisation_A

*** G1. _______ of the snowstorm, schools will not open today.**

○ A. As

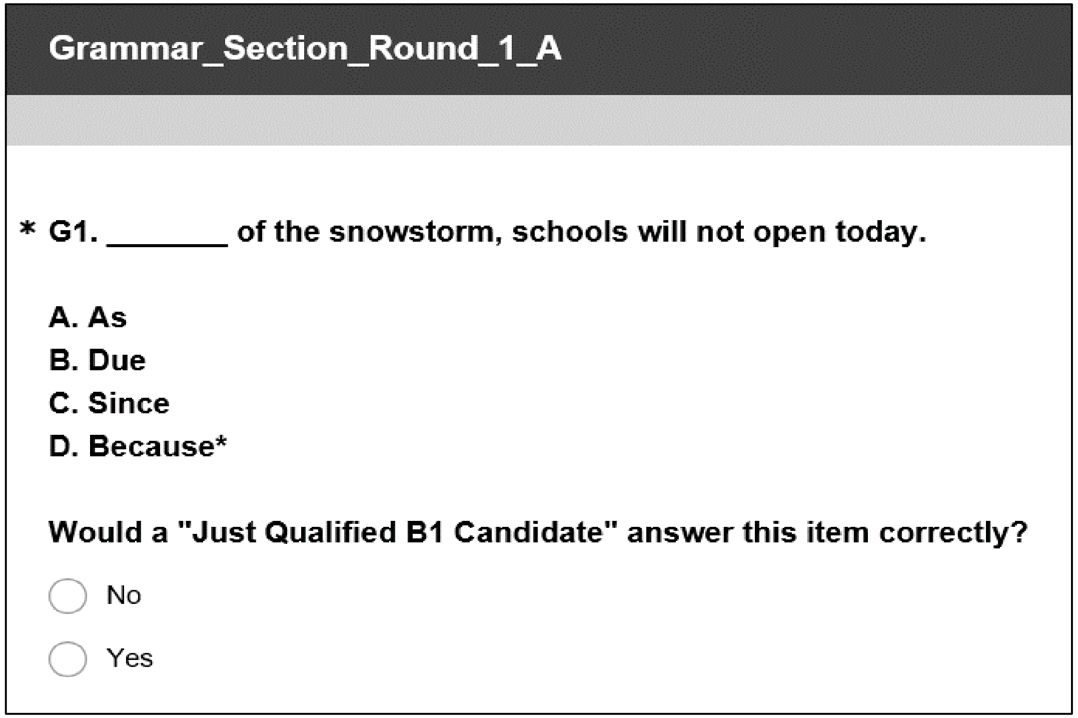
○ B. Due

○ C. Since

○ D. Because

Source: Hellenic American University (n.d.)

Round 1 rating form



Grammar_Section_Round_1_A

*** G1. _______ of the snowstorm, schools will not open today.**

A. As
B. Due
C. Since
D. Because*

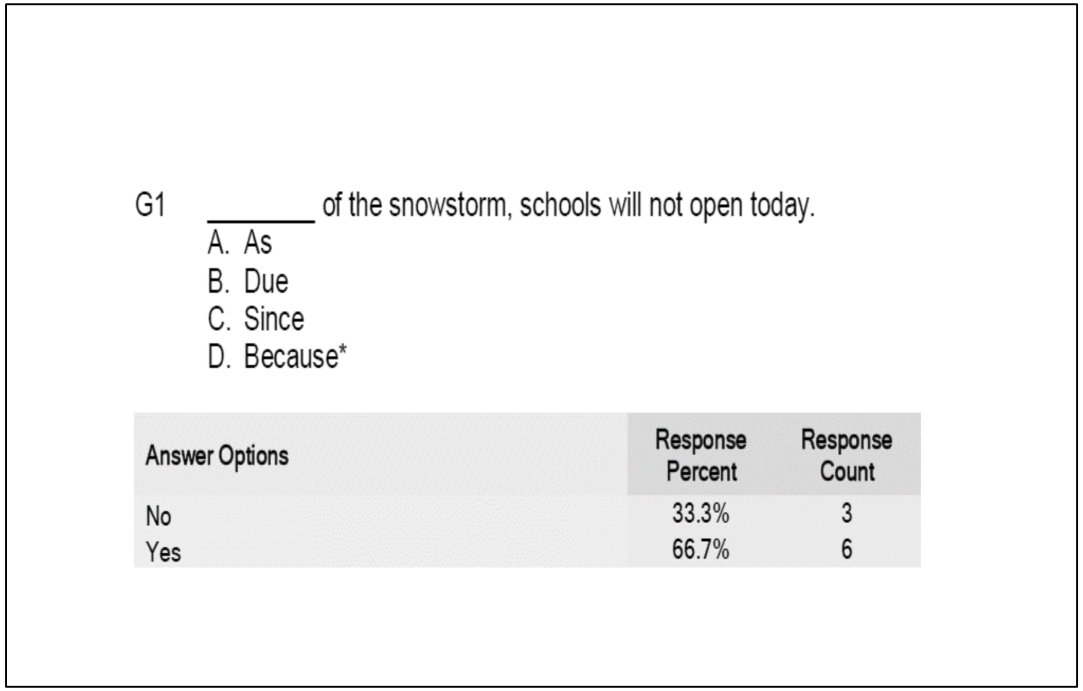
Would a "Just Qualified B1 Candidate" answer this item correctly?

○ No

○ Yes

example of R1 feedback & R2 rating form

Round 1 discussion feedback

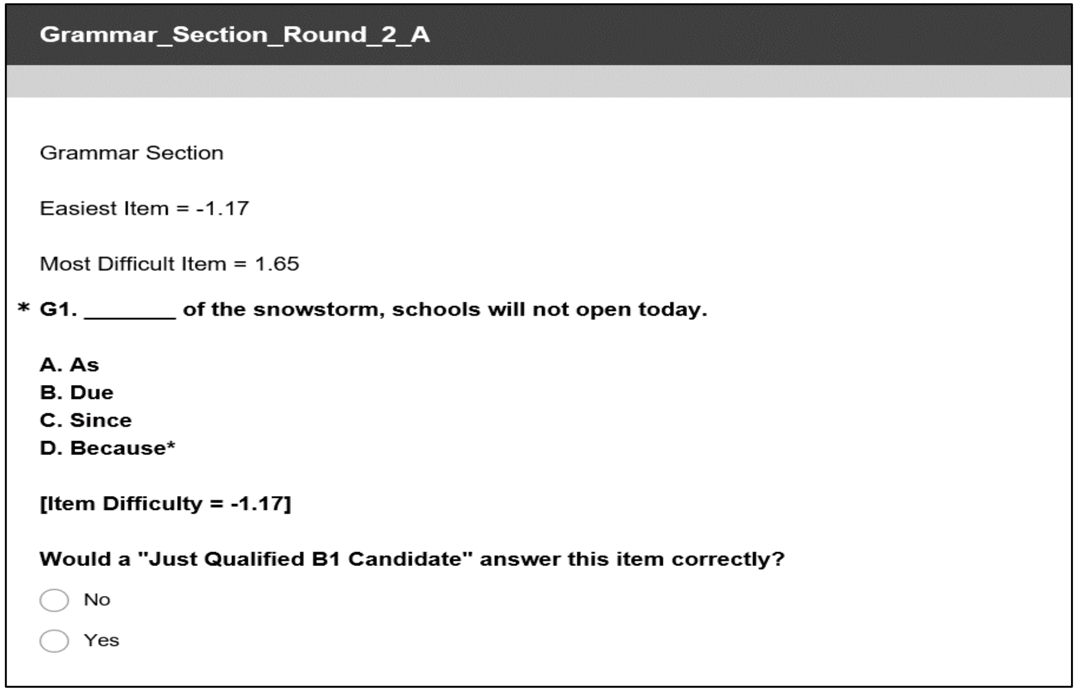


G1 ________ of the snowstorm, schools will not open today.

A. As
B. Due
C. Since
D. Because*

Answer Options	Response Percent	Response Count
No	33.3%	3
Yes	66.7%	6

Round 2 rating form



Grammar_Section_Round_2_A

Grammar Section

Easiest Item = -1.17

Most Difficult Item = 1.65

*** G1. ________ of the snowstorm, schools will not open today.**

A. As
B. Due
C. Since
D. Because*

[Item Difficulty = -1.17]

Would a "Just Qualified B1 Candidate" answer this item correctly?

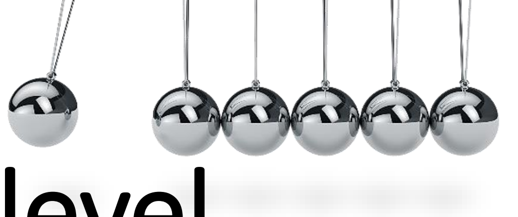
○ No

○ Yes

group/round pairwise interactions

	G1 (n = 9)		**G2 (n = 13)**		**G3 (n = 12)**		**G4 (n = 11)**	
Round 1	G1R1 mean: .27		G2R1 mean: .35		G3R1 mean: .43		G4R1 mean: .28	
	min. -.60	max. 1.33	min. -.40	max. 1.07	min. -.10	max. 1.63	min. -.40	max. 1.07
Round 2	G1R2 mean: .45		G2R2 mean: .50		G3R2 mean: .61		G4R2 mean: .56	
	min. .00	max. 1.80	min. -.10	max. 1.33	min. -.20	max. 1.20	min. -.20	max. 1.99
Welch *t* *(d.f)*	G1 Welch *t*: -1.15 (807)		G2 Welch *t*: .-1.27 (1167)		G3 Welch *t*: -1.47 (1077)		G4 Welch *t*: -1.88 (987)	
	min. -1.56 (87)	max. .94 (87)	min. -1.57 (87)	max. .67 (87)	min. -2.03 (87)	max. .89 (97)	min. -1.61 (83)	max. .89 (87)
prob.	G1 *prob.:* .25		G2 *prob.:* .20		G3 *prob.:* .14		G4 *prob.:* .06	
	min. .12	max. 1.00	min. .12	max. 1.00	min. .05	max. 1.00	min. .11	max. .81
change (n=45)	min. 5		min. 2		min. 0		min. 4	
	max. 16		max. 18		max. 23		max. 19	

OXFORD
UNIVERSITY PRESS



R2 consistency of judgments: individual level

R1: Infit range:
.50 – 1.50
(Linacre, 2018)

	G1 (*n* = 9)		G2 (n = 13)		G3 (n = 12)		G4 (n = 11)	
Infit (*Zstd*)	min. .79 (-2.0)	max. 1.45 (4.0)	min. .69 (-3.1)	max. 1.24 (2.1)	min. .76 (-2.5)	max. 1.15 (1.4)	min. .72 (-3.2)	max. 1.18 (1.7)
Outfit (*Zstd*)	min. .74 (-.6)	max. 1.50 (4.0)	min. .62 (-2.5)	max. 1.38 (2.6)	min. .73 (-1.5)	max. 1.17 (1.4)	min. .70 (-3.1)	max. 1.26 (2.1)
Corr. Ptbis	min. -.01	max. .68	min. .04	max. .85	min. .21	max. .69	min. -01	max. .79
Obs % - Exp%	min. -4.80	max. 13.20	min. -.3.50	max. 19.10	min. .80	max. 11.3	min. -4.60	max. 16.70
Rasch – Kappa	min. -.11	max. .29	min. -.08	max. .38	min. .02	max. .28	min. -.10	max. .40

OXFORD
UNIVERSITY PRESS

R2 consistency of judgments: group level

	G1 (*n* = 9)	**G2 (*n* = 13)**	**G3 (*n* = 12)**	**G4 (*n* = 11)**
Separation ratio *(G)*	1.19	.27	.47	1.40
Separation (strata) index *(H)*	1.92	.69	.96	2.20
Separation reliability (*R*)	.59	.07	.18	.66
χ^2 (d.f.)	15.5 (8)	12.8 (12)	14.8 (11)	26.0 (10)
χ^2 prob	.05	.39	.19	.00
Observed agreement (%)	63.3	67.7	63.4	67.0
Expected agreement (%)	56.2	57.2	58.0	56.9
Rasch – Kappa	.16	.25	.13	.23

OXFORD
UNIVERSITY PRESS

inter/ intra judge consistency:

	G1 (*n* = 9)	G2 (*n* = 13)	G3 (*n* = 12)	G4 (*n* = 11)
Internal consistency [SEc/RMSE ≤ .50]	.43	.24	.27	.44
Ratings correlated with empirical item difficulties	.58*	.77*	.73*	.72*

*all correlations significant at the .05 level (2-tailed)

judge feedback

Rank order, from least (1) to most (7), the following sources of information that advised your judgments. Select one (1) for the source of information you relied on the least to make your judgment and seven (7) for the source you relied on the most.

	G1 (*n* = 9)		**G2 (*n* = 13)**		**G3 (*n* = 12)**		**G4 (*n* = 11)**	
	T.Score	**Rank**	**T.Score**	**Rank**	**T.Score**	**Rank**	**T.Score**	**Rank**
My experience taking the test	40	2	48	4	41	6	54	1
My own experiences with real students	57	1	67	1	50	4	52	2
The Performance Level Descriptors (PLDs)	25	7	43	7	47	5	48	3
The item performance info. (e.g., p-values)	31	5	45	6	38	7	45	4
The panel discussions	38	3	50	3	51	2	37	5
The normative info. (e.g. judge ratings)	28	6	46	5	51	2	37	5
The consequences info. (i.e. impact data)	33	4	65	2	58	1	35	7

study 2

Participants

- 9 judges

SS Method

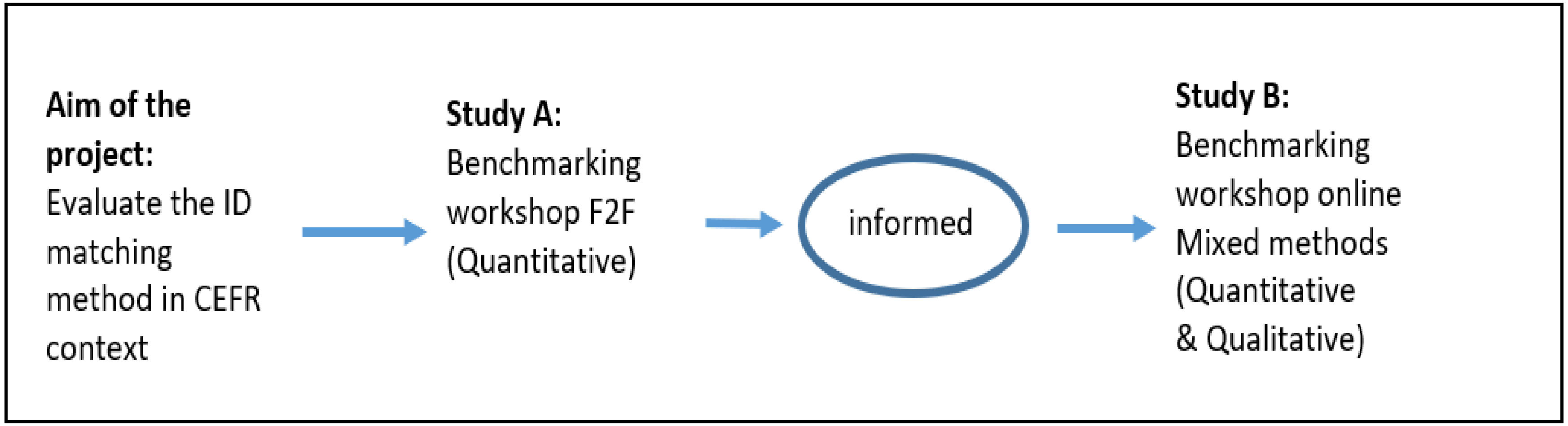
Item Descriptor (ID) matching method

Instruments

- Integrated Skills in English II (ISE, B2)
- Reading Section
  - (11 items)
  - Ordered Item Booklet (OIB)

Part of PhD project

The mixed methods multiphase evaluation design



Aim of the project: Evaluate the ID matching method in CEFR context

Study A: Benchmarking workshop F2F (Quantitative)

informed

Study B: Benchmarking workshop online Mixed methods (Quantitative & Qualitative)

(Creswell, 2014; Creswell & Clark, 2018; Plano Clark & Ivankova, 2016)

TRINITY COLLEGE LONDON

Universität Bremen*

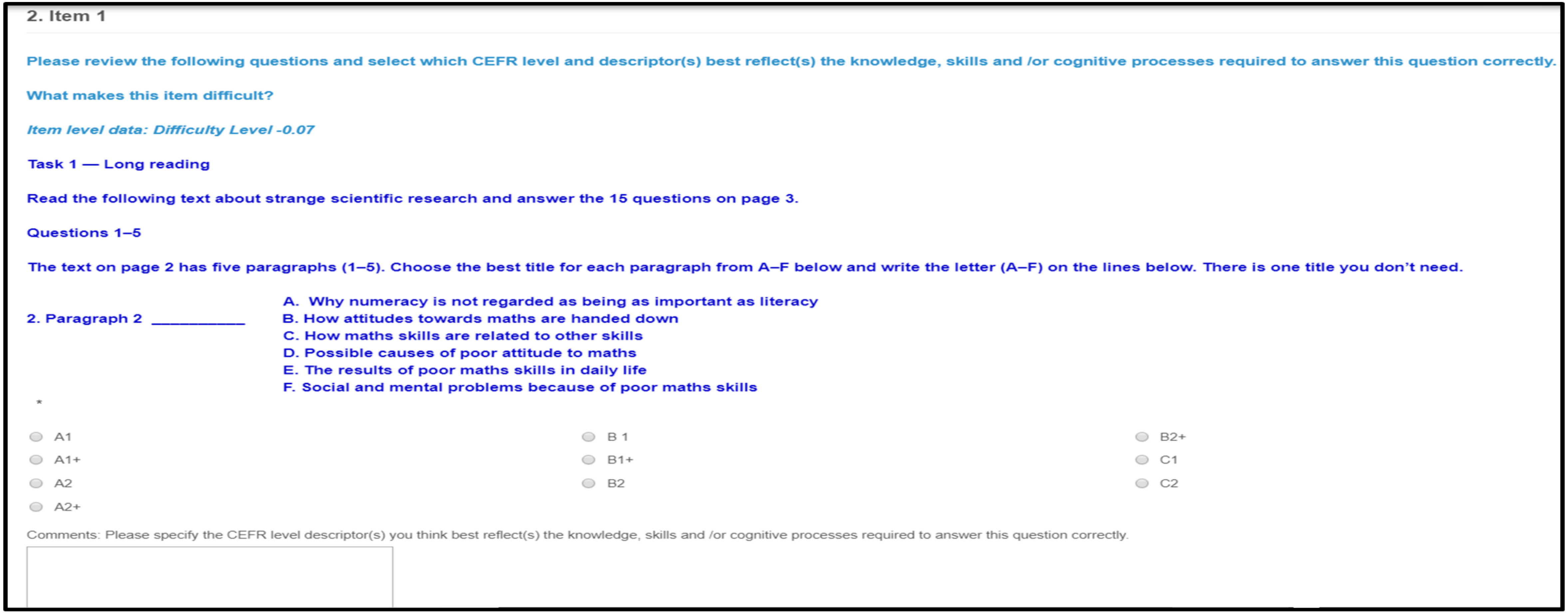
the ID matching method

Judge task:

i. Which performance level descriptor(s) most closely match(es) the knowledge and skills required to respond successfully to this item (or score level for constructed response items)?

ii. What makes this item more difficult than the ones that precede it?

example of OIB rating form



2. Item 1

Please review the following questions and select which CEFR level and descriptor(s) best reflect(s) the knowledge, skills and /or cognitive processes required to answer this question correctly.

What makes this item difficult?

Item level data: Difficulty Level -0.07

Task 1 — Long reading

Read the following text about strange scientific research and answer the 15 questions on page 3.

Questions 1–5

The text on page 2 has five paragraphs (1–5). Choose the best title for each paragraph from A–F below and write the letter (A–F) on the lines below. There is one title you don't need.

2. Paragraph 2 ___________

A. Why numeracy is not regarded as being as important as literacy
B. How attitudes towards maths are handed down
C. How maths skills are related to other skills
D. Possible causes of poor attitude to maths
E. The results of poor maths skills in daily life
F. Social and mental problems because of poor maths skills

*

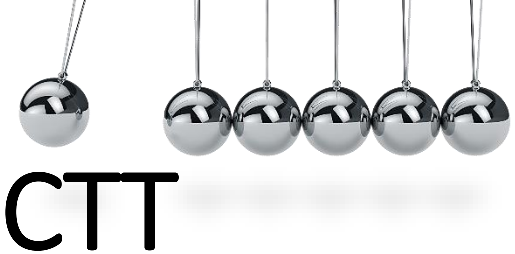
○ A1
○ A1+
○ A2
○ A2+

○ B 1
○ B1+
○ B2

○ B2+
○ C1
○ C2

Comments: Please specify the CEFR level descriptor(s) you think best reflect(s) the knowledge, skills and /or cognitive processes required to answer this question correctly.

TRINITY
COLLEGE LONDON



reliability & consistency of judgements: CTT

	Round 1	Round 2
Cronbach's Alpha	.90	.91
ICC (absolute agreement)	.83	.87

consistency of judgments: Rasch

R1: Infit range: Infit mean ± 2SD
(Pollitt & Hutchinson, 1987)
-.35 to 1.97

R2: Infit range: Infit mean ± 2SD
(Pollitt & Hutchinson, 1987)
-0.33 to 1.95

	Round 1		**Round 2**	
Infit **(*Zstd*)**	min. .27 (-1.6)	max. 1.93 (1.5)	min. .28 (-1.50)	max. 1.92 (1.50)
Outfit (*Zstd*)	min. .28 (-1.5)	max. 1.68 (1.0)	min. .28 (-1.5)	max. 1.70 (1.1)
Corr. PtMeasure	min. .00	max. .98	min. .55	max. .94
Obs % - Exp%	min. -9.5	max. 6	min. -.8.2	max. 8.5
Rasch –Kappa	min. -.15	max. .10	min. -.11	max. .14
Change (*n* = 11)			min. 0	max. 6

TRINITY
COLLEGE LONDON

R2 consistency of judgments: group level

	Round 1	**Round 2**
Separation ratio *(G)*	1.96	1.52
Separation (strata) index *(H)*	2.95	2.36
Separation reliability (*R*)	.79	.70
χ^2 (d.f.)	40.6 (8)	31.3 (8)
χ^2 prob	.01	.00
Observed agreement (%)	31.3	35.1
Expected agreement (%)	32.6	34.4
Rasch – Kappa	-.02	.01

TRINITY
COLLEGE LONDON

judge feedback

Please consider which of the source information listed below advised your judgement the most and rank order them from the most important (6) to the least important (1).

	Total score	**Overall Rank**
The samples of actual test takers’ responses (oral or written, item difficulties)	35	1
The CEFR level descriptors	24	2
The group discussions	23	3
Other participants’ ratings	22	4
My own experiences with real students	22	4
My experience taking the test	21	6

concluding

Q1: Do judges change their ratings across rounds? If yes, to what extent?

Q2: What do judges claim mainly influences their ratings?

Q3: Can we use MFRM to analyse Round 2 & Round 3 ratings?

Q4: Do judges remain independent experts across rounds?

Q5: What do we gain from MFRM analysis of standard setting data?

references

Creswell, J. W. (2014). *Research design: Qualitative, quantitative and mixed methods approaches.* California: Sage.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3 ed.). London: SAGE Publications Ltd.

Eckes, T. (2015). *Introduction to Many-Facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd revised and updated ed.). Frankfurt: Peterlang

Ferrara, S., Perie, M., & Johnson, E. (2008). Matching the judgemental task with standard setting panelist expertise: The Item-Descriptor (ID) matching method. Journal of Applied Testing Technology, 9(1), 1-20.

Hellenic American University. (n.d.). *Basic Communication Certificate in English (BCCE): Official past examination Form A test booklet.* Retrieved from: https://hauniv.edu/images/pdfs/bcce_past_paper_form_a_test_booklet2.pdf

Linacre, J. M. (2018). A user's guide to FACETS Rasch-model computer programs (Program manual 3.81.0). Retrieved from http://www.winsteps.com/ manuals. htm.

Plano Clark, V. L., & Ivankova, N. V. (2016). *Mixed methods research: A guide to the field.* California: Sage.

Pollitt, A., & Hutchinson, C. (1987). Calibrated graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing, 4*(1), 72-92. doi:10.1177/026553228700400107

Thank you!

Charalambos.Kollias@oup.com

voula.kanistra@trinitycollege.com

OXFORD
UNIVERSITY PRESS

TRINITY
COLLEGE LONDON