Assessing raters' ratings' quality in writing assessment

Under holistic and analytic scoring. An empirical example

Psychometric Society

IMPS 2023

College Park, Maryland, US

July 25

2023

Carrasco, Diego, PhD,
Centro de Medición MIDE UC
Pontificia Universidad Católica de Chile

Ávila, Natalia, PhD; Castillo, Carolina PhD(c); Escribano, Rosario, PhD,
Facultad de Educación
Pontificia Universidad Católica de Chile

Espinosa, María Jesús, PhD,
Facultad de Educación
Universidad Diego Portales

Figueroa, Javiera, PhD,
Facultad de Educación
Universidad Alberto Hurtado

Centro UC
Medición - MIDE

https://github.com/dacarras/imps_2023_wri

Raters under different rating schemes

Introduction

why

https://github.com/dacarras/imps_2023_wri

Why

- Holistic rubrics are favored in writing assessment when writing quality is to be judge (White, 1984; Elliot, 2005).
- Analytic scoring is often dismissed in the **holistic tradition**, due to its over-reliance on superficial characteristics of written samples, including indicators such as word complexity, or spelling (e.g., Hamp-Lyons, 2016), instead of writing communicative attributes.
- However, **"analytic rubrics" in its more descriptive meaning refers to scoring methods** using two or more indicators (Frey, 2018). These different views on rubrics, regarding what is holistic or analytic**, confound scoring methods with writing assessment approaches**.

Figure 1: how and attribute is conceived and its assumed method

Holisticism

Calling for an overall view of the phenomena of interest

↓

Holistic scoring

~~"Analyticism"~~

What summative aspects can we distinguish

↓

Analytic scoring

https://github.com/dacarras/imps_2023_wri

Present Study

- We assume we can **separate** the **conceptual model** of the attribute of interest (e.g., writing quality, communicative purpose fulfilment), from the characteristics of the different measurement tools out there: including the **measurement process**, and its **response model**.

- We want to know which scoring method is more convenient, within a common conceptual framework of the attribute of interest. Thus, ***using a "holistic" attribute concept of writing assessment, two rubrics were generated***: a single indicator, and a multiple indicator rubric.

- In essence, we are applying a **pragmatic framework** of measurement (Torres-Irribarra, 2021).

Figure 2: separating the concept of the attribute with the properties of the measurement model

Dimension 1

Conceptual model regarding the attribute of interest

Holisticism

Calling for an overall view of the phenomena of interest

~~"Analyticism"~~

What summative aspects can we distinguish

Dimension 2

Measurement tool used to represent the attribute

Single indicator

(Holistic Rubric)

Multiple Indicators

(Analytic Rubrics)

https://github.com/dacarras/imps_2023_wri

Present Study

Figure 3: separating the concept of the attribute with the properties of the measurement model

Dimension 1

Conceptual model regarding the attribute of interest

Dimension 2

Measurement tool used to represent the attribute

	Holisticism Calling for an overall view of the phenomena of interest	**~~"Analyticism"~~** What aspects are relevant
Single indicator (Holistic Rubric)	Option A	Option C
Multiple Indicators (Analytic Rubrics)	Option B	Option D

https://github.com/dacarras/imps_2023_wri

Present Study

Figure 3: separating the concept of the attribute with the properties of the measurement model

Dimension 1

Conceptual model regarding the attribute of interest

Dimension 2

Measurement tool used to represent the attribute

	Communicative Purpose Given a task, can writers fulfill their purpose	~~**Formalism of the written samples**~~ e.g., Orthography, Spelling, Vocabulary
Single indicator (Holistic Rubric)	Option A	Option C
Multiple Indicators (Analytic Rubrics)	Option B	Option D

https://github.com/dacarras/imps_2023_wri

Present Study

Figure 3: separating the concept of the attribute with the properties of the measurement model

Dimension 1

Conceptual model regarding the attribute of interest

Communicative Purpose

Given a task, can writers fulfill their purpose

~~**Formalism of the written samples**~~

e.g., Orthography, Spelling, Vocabulary

Dimension 2

Measurement tool used to represent the attribute

Single indicator

(Holistic Rubric)

Multiple Indicators

(Analytic Rubrics)

	Focus of interest	Outside the focus of interest
Single indicator	Option A	Option C
Multiple Indicators	Option B	Option D

Comparing Raters ratings quality under different schemes

Method

How

https://github.com/dacarras/imps_2023_wri

Method

Design

- **Fully crossed**: 90 raters (88 observed raters) x 2 rubrics x Order (H-A, A-H).
- 2 rubrics (Holistic 6 levels, Analytic: 6 indicators (2 and 3 ordinal categories), to rate 30 written samples from 5th graders
- Design helps us to teased out rater variance from written sample proficiency variance

Written samples

- From a previous study, we had more than 400 written samples from 4th grade students, scored with a holistic rubric of six performance levels
- We select 5 written samples of each level with no legibility problems, and no vocabulary marks (e.g., common uses considered too informal)
- The selected written samples are aimed to gives us **maximum coverage on writing proficiency** (as possible).

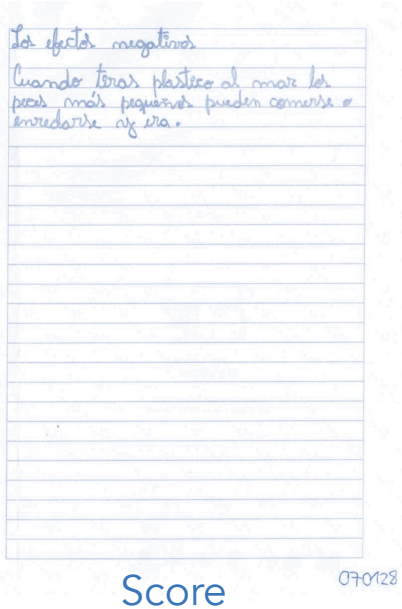
Figure 4: Research questions and their method traditions

Wind & Peterson (2018)

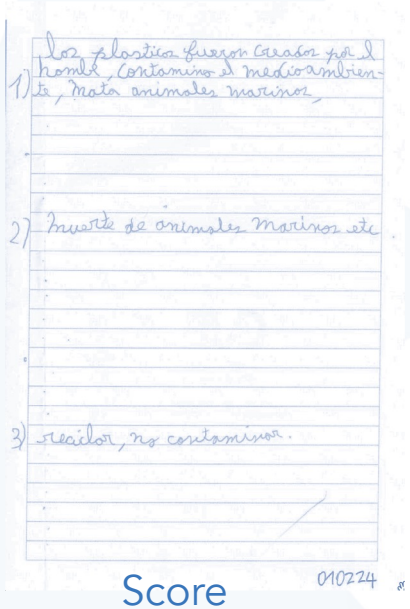
Research traditions

Observed Ratings Total score	**Scaled Ratings** Response Models
Variance Partitions Models	Accuracy Model (comparison to master raters)
How much uncertainty comes from the different rating process?	How accurate are the raters scores under different rating process?

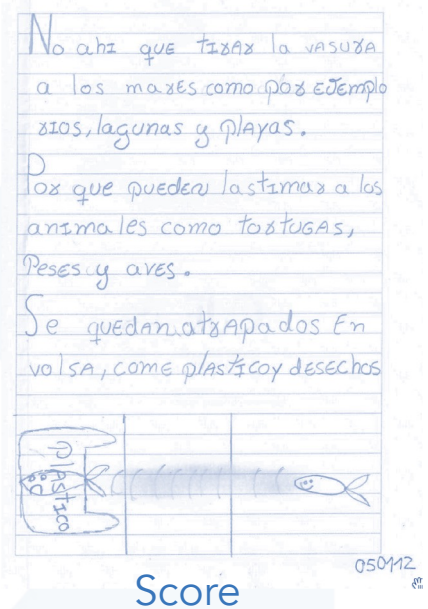
Communicative Purpose is unfulfilled → Communicative Purpose is fulfilled



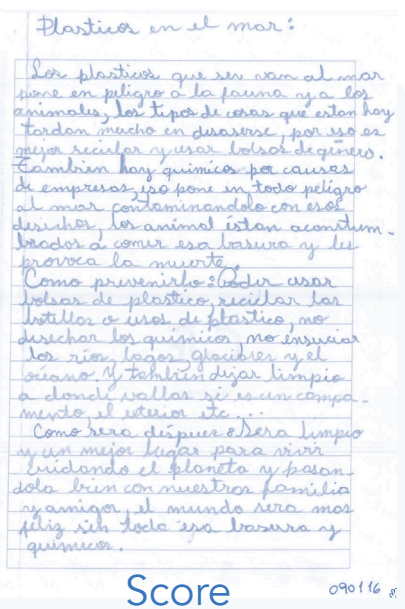
Score 1



Score 2



Score 4



Score 6

connection	
purpose	
transcription	
progression	
demarcation	

Task: write up a brochure describing the effects of sea plastic pollution, and how to prevent it.

Centro UC Medición - MIDE

Note: diagrams are referential.

Variance Partition Models

Total Score Uncertainty

How much uncertainty comes from the different rating process?

https://github.com/dacarras/imps_2023_wri

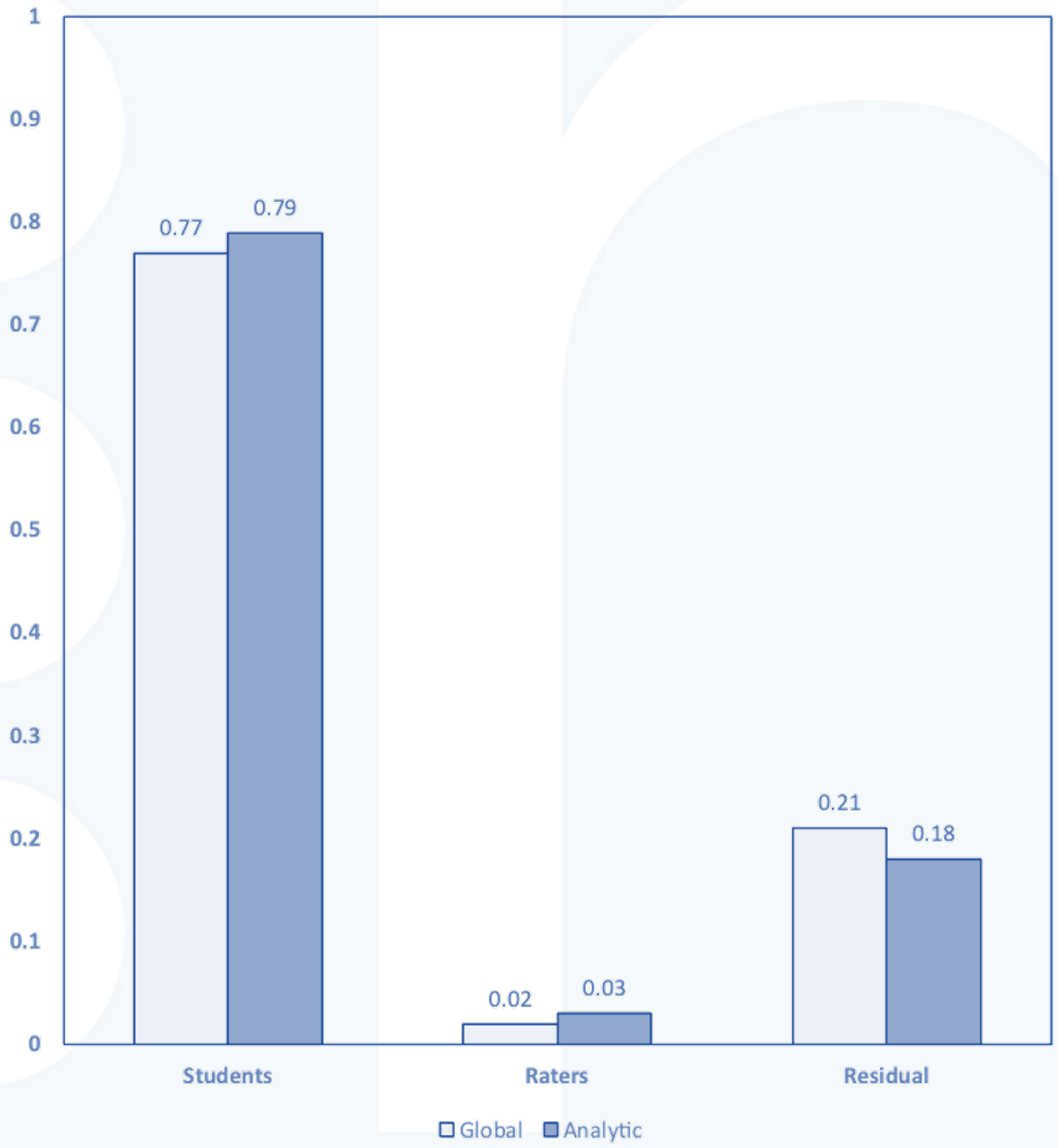
Study 1: raters' error

Approach

- Total scores for each rubric where computed, and z score
- The term of interest is the relative variance attribute to the raters under each rating process
- We use a cross classified model to get random terms for students and for raters, with Bayes estimators (and non informative priors).
- We found no substantial differences between the two different rating process

	Global			Analytic		
	E	LL	UL	E	LL	UL
Students	0.77	0.67	0.87	0.79	0.69	0.87
Raters	**0.02**	**0.01**	**0.03**	**0.03**	**0.02**	**0.05**
Residual	0.21	0.13	0.3	0.18	0.11	0.26

Figure 4: Variance Partition Model Point Estimates



https://github.com/dacarras/imps_2023_wri

Accuracy Models

Raters' Accuracry

How accurate are the raters ' scores under different rating process?

https://github.com/dacarras/imps_2023_wri

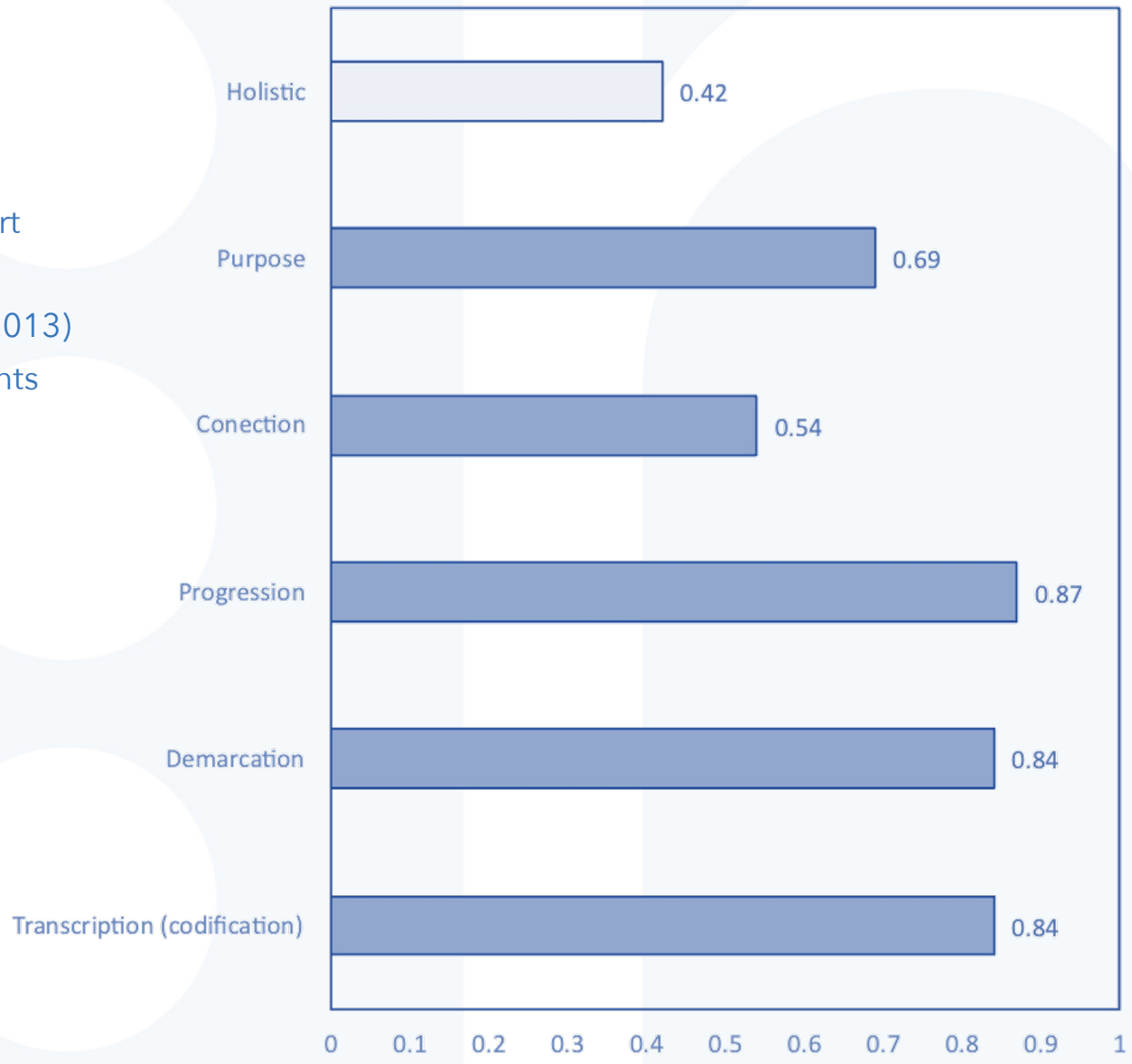
Study 2: raters' accuracy

Approach

- We created a long format response table. Each observed rater response, was coupled with an expert rater response (three expert raters, with total agreement). Each response was classified as "accurate" if matched the benchmark score (Wind & Engelhard, 2013)
- We fitted a cross classified mixed logistic model, including students and raters as random terms. Estimates are logits.

```
=============================================
                          Model 1
---------------------------------------------
(Intercept)                  1.21 (0.11) ***
rubric(holistic)            -1.56 (0.05) ***
---------------------------------------------
AIC                       17481.17
BIC                       17511.85
Log Likelihood            -8736.59
Num. obs.                 15840
Num. groups: rater           88
Num. groups: person          30
Var: rater (Intercept)        0.02
Var: person (Intercept)       0.35
=============================================
*** p < 0.001; ** p < 0.01; * p < 0.05
```

Figure 5: percentage of matched responses to master scores



https://github.com/dacarras/imps_2023_wri

Comparing rating processes

Summary

Final remarks

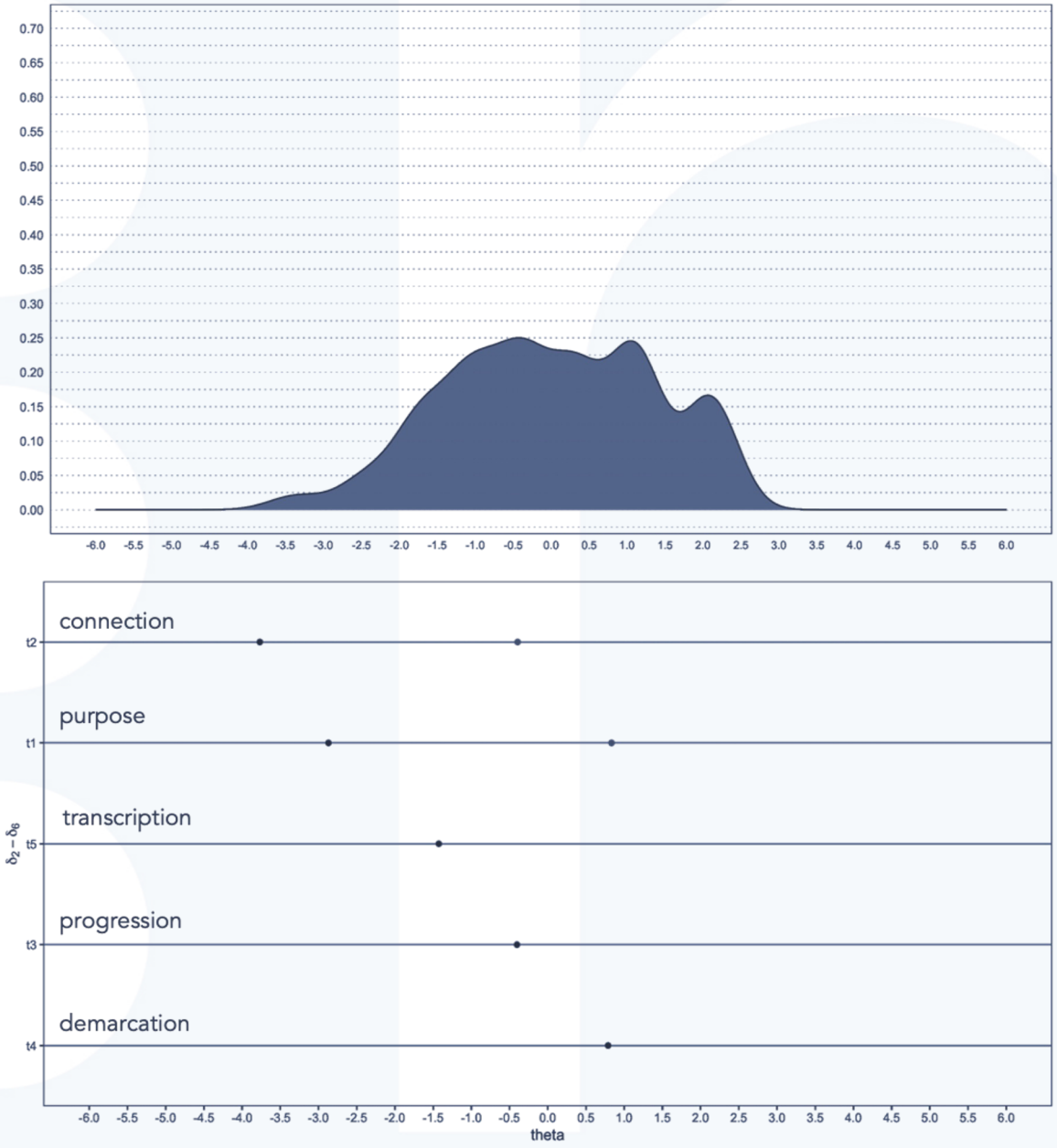
https://github.com/dacarras/imps_2023_wri

Limitations

- Study 1. A common limitation of observed score tradition approaches, is its difficulty to separate ability and rater variance. In the present study we isolate ability by design, using the same set of responses with uniform ability.

- Study 2. Although rating accuracy is higher to the multiple indicator rubric in general, the accuracy varies between indicators. Purpose and Connection are indicators with the less accurate results. Raters' training may need to focus more thoroughly in these two indicators.

Figure 6: expected item person map for the multiple indicator rubric



https://github.com/dacarras/imps_2023_wri

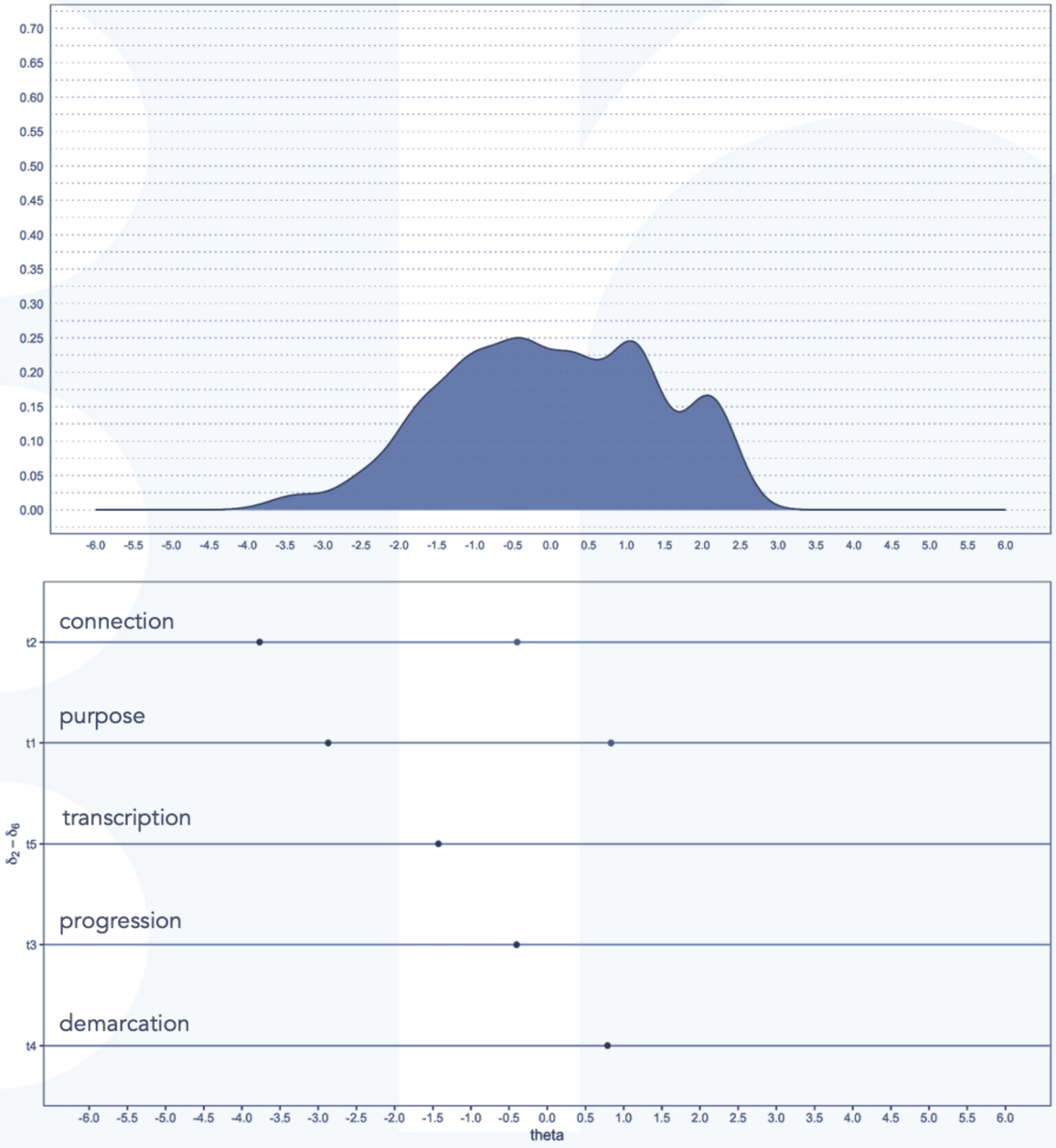
Closing remarks

- The present study is part of a larger endeavor (2-year grant funded project). Our aim is to reach an optimal design informed by different studies (i.e., evidence-based assessment design). **Any ideas are welcome!**

- We opted for the multiple indicator rubric.
  - Not due to the rater uncertainty (which is negligible)
  - Arguably more accurate rates
  - **Notably is more informative**: each indicator is informative to the communicative purpose, because each indicator was design with the theoretical interest in mind.

- In multidisciplinary work and assessment development, we believe is of upmost importance to **distinguish** the **concept** of the attribute, from the **means** of **research** (i.e., the scoring tool, and the response model).

Figure 6: expected item person map for the multiple indicator rubric



https://github.com/dacarras/imps_2023_wri

Muchas gracias!

Carrasco, D., PhD,

Centro de Medición MIDE UC

Pontificia Universidad Católica de Chile

https://dacarras.github.io/

https://github.com/dacarras/imps_2023_wri

Referencias

Elliot, N. (2005). On a Scale. A social History of Writing Assessment in America. Peter Lang Publishing Inc.

Frey, B. B. (2018). The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation (B. B. Frey (ed.)). SAGE Publications, Inc.

Hamp-Lyons, L. (2016). Farewell to Holistic Scoring? Assessing Writing, 27, A1–A2. https://doi.org/10.1016/j.asw.2015.12.002

Torres Irribarra, D. (2021). A Pragmatic Perspective of Measurement. Springer International Publishing. https://doi.org/10.1007/978-3-030-74025-2

White, E. M. (1984). Holisticism. College Composition and Communication, 35(4), 400. https://doi.org/10.2307/357792

Wind, S. A., & Engelhard, G. (2013). How invariant and accurate are domain ratings in writing assessment? Assessing Writing, 18(4), 278–299. https://doi.org/10.1016/j.asw.2013.09.002

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. Language Testing, 35(2), 161–192. https://doi.org/10.1177/0265532216686999