Addressing two methodological challenges
when students rate their classroom discussion.

Carrasco, D., Centro de Medición MIDE UC, Pontificia Universidad Católica de Chile
Treviño, E., Facultad de Educación, Pontificia Universidad Católica de Chile
López-Hornickel, N., Advanced Quantitative Methods, University of Bath
Castillo, C., Facultad de Educación, Pontificia Universidad Católica de Chile

https://github.com/dacarras/invalsi_2022_opd

Students rating their classroom
chapter

Students Ratings Their Open Classroom Discussion

Diego Carrasco, Ernesto Treviño, Natalia López Hornickel, and Carolina Castillo

1 Introduction

Past research in civic education has positioned open classroom discussion of political and social issues (OPD) as an essential factor for different citizenship outcomes, including civic knowledge (Isac et al., 2014), support of egalitarian values (Carrasco & Torres Irribarra, 2018), political efficacy (Knowles & McCafferty-Wright, 2015), among others. At the same time, it present hostile relations with youth alienation (Torney-Purta, 2009), authoritarianism endorsement (Hahn & Tocci, 1990), and tolerance of corruption (Carrasco et al., 2020).

OPD is a reflective measure of the learning environment and not a classical individual difference measure. It allows capturing students' experience as a collective (at the school level) through students' perceptions as individuals. Student responses are the source of information about their school practices, were students rate their learning environments (Carrasco & Torres Irribarra, 2018). OPD items are reference-shift items, and if their rating response nature is ignored, the compositional models can lead to the wrong conclusions. It is argued that compositional model specification produce an unnecessary correction of level 2 estimates for reference shift scale scores (Lüdtke et al., 2009). Additionally, OPD scores of schools are subject to students' inter-rater variability. As such, two different schools may receive the same OPD mean score, yet the students' OPD ratings can vary broadly (Schweig, 2016). How much students' rating variability is tolerable? Cut off scores may not be easily determined. Common advice in the organizational literature is to exclude

D. Carrasco (✉)
Centro de Medición MIDE UC, Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: dacarras@uc.cl

E. Treviño · N. López Hornickel · C. Castillo
Facultad de Educación, Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: ernesto.trevino@uc.cl; nvlopez@uc.cl; carolina.castillo@uc.cl

463

468   D. Carrasco et al.

Fig. 1 Graphical representation of $\gamma_c = \gamma_b - \gamma_w$

and γ_w. In Table 2 we include the estimates of γ_b, γ_w, and γ_c for all participating ICCS 2016 samples, using the disaggregated model.

In total 6 out of 24 countries, present a similar scenario of Italy. In Italy, Slovenia, Russia, Latvia, Norway, and Belgium, there is a risk of making the conclusion that OPD school levels are not relevant to explain civic knowledge between schools, if γ_c is wrongly interpreted as a learning environment effect.
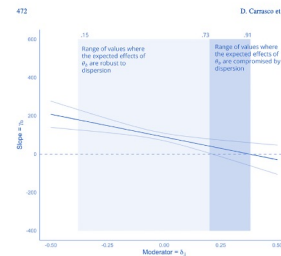
472   D. Carrasco et al.

Fig. 2 Johnson-Neyman plot for γ_b conditional to $\bar{x}_j - \bar{x}_{..}$ scores

the moderation effect of the dispersion model is small, estimates from Model 1 and Model 3 produces similar estimates of γ_b. As such, the dispersion effect model is advantageous for only some scenarios.

The dispersion effect model permits to estimate the critical point where a lack of consensus between students from the same school compromise the estimates of a referent-shift items scale scores. In Fig. 2, we plot the expected slope of OPD school levels, at the observed values of the dispersion score. The black line represent the point estimates of γ_b at different dispersion score values. The curve lines that accompany the points estimates are the 95% confidence intervals of γ_b. The model specification has the advantage that the OPD realizations and the dispersion score are in the same scale, thus its coefficients can be interpreted in a similar manner. Between 0.73 and 0.91 of dispersion, the estimates of γ_b are compromised as its lower confidence interval crosses zero.

Springer Proceedings in Mathematics & Statistics

Marie Wiberg · Dylan Molenaar
Jorge González · Ulf Böckenholt
Jee-Seon Kim *Editors*

Quantitative Psychology

The 85th Annual Meeting
of the Psychometric Society, Virtual

Springer

https://github.com/dacarras/invalsi_2022_opd

https://link.springer.com/chapter/10.1007/978-3-030-74772-5_41

Topic

Classroom Discussion of political and social issues

Classroom discussion
of political and social issues

- The **sophistication hypothesis** expects that more educated adults develop higher political sophistication (Lipset, 1959), because education enables persons to organized concepts, facts and political opinions (Lyons, 2017). Political sophistication is needed of active citizenry: to elect fit politicians, participate in jury duty, denounce corruption and defend civil rights.

- On the other hand, the **intergenerational transmission** hypothesis (Schlozman, Verba, & Brady, 2012) asserts that political inequities between adults are inherit from parents to sons. Thus, **unless there is an intermediate action**, political sophistication of students can be guessed solely based on the family socioeconomic background.

- As such **Classroom discussion of political issues at school, is expected to compensate for the lack political socialization at home**. Its an school measure that tells us how often students discuss with their peers and their teachers (Ehman, 1969) about political and social issues during regular lessons (Carrasco & Torres Irribarra, 2018).

Students rating their classrooms
open classroom discussion

dati invalsi
uno strumento
per la ricerca
e la didattica

Q17 **When discussing political or social issues during regular lessons, how often do the following things happen?**

(Please tick only one box in each row.)

		Never	Rarely	Sometimes	Often
IS3G17A	a) Teachers encourage students to make up their own minds.	☐₁	☐₂	☐₃	☐₄
IS3G17B	b) Teachers encourage students to express their opinions.	☐₁	☐₂	☐₃	☐₄
IS3G17C	c) Students bring up current political events for discussion in class.	☐₁	☐₂	☐₃	☐₄
IS3G17D	d) Students express opinions in class even when their opinions are different from most of the other students.	☐₁	☐₂	☐₃	☐₄
IS3G17E	e) Teachers encourage students to discuss the issues with people having different opinions.	☐₁	☐₂	☐₃	☐₄
IS3G17F	f) Teachers present several sides of the issues when explaining them in class.	☐₁	☐₂	☐₃	☐₄

What is the referent for each of these assertions?

Reflective Measure
open classroom discussion

A different type of items

- **Reference-shift items** (Bliese et al., 2018) are items refer to the group or to an attribute of the learning environment (e.g., "**in my classroom there is a lot of noise**").

- These kind of items are different from referring to the student (e.g., "**Are you a girl or a boy?**")

- These are often used to produce scores of school features, by aggregating students' responses at the cluster level. These are often called reflective measures (Stapleton, et al 2016).

Methodological challenges for reflective measures

- **Problem 1:** Ignoring items come from a reflective measure design, can lead to **wrong model specification**, and wrong conclusions regarding school practices (Lüdtke et al. 2009).

- **Problem 2**: while fixing problem 1, the aggregation of scores to the cluster of interest, is subject to cluster inter rater variance. **How much agreement is needed to trust these scores?** Single rule of thumbs may not apply (Lüdtke et al. 2009) [problem 2].

- In the following presentation we will address these two problems.

General problem

under estimation of school attributes estimates when students are informants

Let's see an illustration of problem 1

classroom discussion and civic knowledge

Italy, ICCS 2016

Illustration of problem 1
Italy (ICCS, 2016)

Table 1: unstandardized estimates on civic knowledge (Italia, ICCS 2016)

	Model 1	Model 2	Model 3
Intercepts	-.05 (.04)	-.05 (.04)	-.04 (.03)
Classroom discussion (School)	**.26 (.10) ****	**.02 (.09)**	**-.04 (.07)**
Classroom discussion (students)		.24 (.02) ***	.22 (.02) ***
SES (students)			.30 (.02) ***
Random part			
Students	.85 (.03) ***	.80 (.03) ***	.74 (.02) ***
Schools	.15 (.04) ***	.15 (.04) ***	.10 (.02) ***

*** p < 0.001, ** p < 0.01, * p < 0.05

Illustration of problem 1
Italy (ICCS, 2016)

Table 1: unstandardized estimates on civic knowledge (Italia, ICCS 2016)

	Model 1	Model 2	Model 3
Intercepts	-.05 (.04)	-.05 (.04)	-.04 (.03)
Classroom discussion (School)	.26 (.10) **	.02 (.09)	-.04 (.07)
Classroom discussion (students)		.24 (.02) ***	.22 (.02) ***
SES (students)			.10 (.02) ***
Random part			
Students	.85 (.03) ***	.80 (.03) ***	.74 (.02) ***
Schools	.15 (.04) ***	.15 (.04) ***	.10 (.02) ***

*** p < 0.001, ** p < 0.01, * p < 0.05

Problem 1

Wrong Model Specification

Estimands representation
what the estimands stand for

High level
school attribute

Average level
school attribute

γ_c

γ_b

γ_w

Low level
school attribute

y_{ij} = Learning outcome
(civic knowledge)

x_{ij} = Students responses regarding an attribute of the learning environment
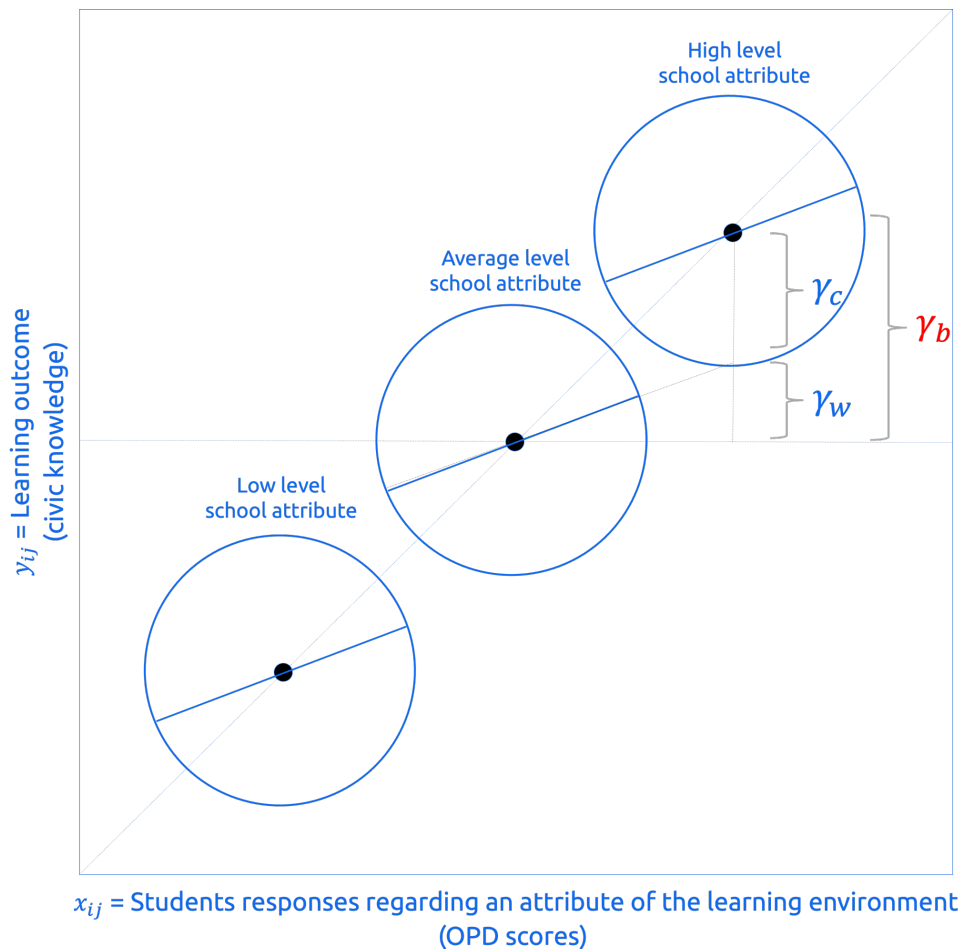(OPD scores)

Fig. 1 Graphical representation of $\gamma_c = \boldsymbol{\gamma_b} - \gamma_w$

Wrong estimand (compositional model)

• The compositional model at the between level, produces $\boldsymbol{\gamma}_c$. This **coefficient underestimate the effect of interest**, because its overcorrecting the effect by the student's perception scores across schools.

  • $y_{ij} = \alpha + \gamma_w\left(w_{ij} - \overline{w}_{..}\right) + \gamma_c\overline{w}_{.j} + u_j + \epsilon_{ij}$

Alternative Model (estimand of interest).

• When we compare learning environments, we are interested in the **difference between classroom environments as a whole**. That is, we are interested in the relative distance that its capture by $\boldsymbol{\gamma}_b$

  • $y_{ij} = \alpha + \gamma_w\left(w_{ij} - \overline{w}_{.j}\right) + \boldsymbol{\gamma_b}\overline{w}_{.j} + u_j + \epsilon_{ij}$

Notes

• We believe the compositional model is often used over reflective measures in large scale assessment, because the generated scores of **referent-shift item-scales** are usually generated by a **latent trait model**, in the same vein as other students' attributes.

• Alternatively, we could use a measurement model that includes **students and schools as relevant sources of variance**.

Alternative Response Model
retrieving open classroom discussion scores in a different way

Q17 **When discussing political or social issues during regular lessons, how often do the following things happen?**

(Please tick only one box in each row.)

		Never	Rarely	Sometimes	Often
OP2	a) Teachers encourage students to make up their own minds.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
OP3	b) Teachers encourage students to express their opinions.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
OP4	c) Students bring up current political events for discussion in class.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
OP5	d) Students express opinions in class even when their opinions are different from most of the other students.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
OP6	e) Teachers encourage students to discuss the issues with people having different opinions.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
OP7	f) Teachers present several sides of the issues when explaining them in class.	$\square_1$	$\square_2$	$\square_3$	$\square_4$

By using a multilevel item response theory model, we can retrieve orthogonal scores for open classroom discussion; and use these scores in the inferential model.

Torres-Irribarra, P. & Carrasco, D. (2019)
A multilevel partial credit model to retrieve classroom diferences, when students are informants of their classrooms.
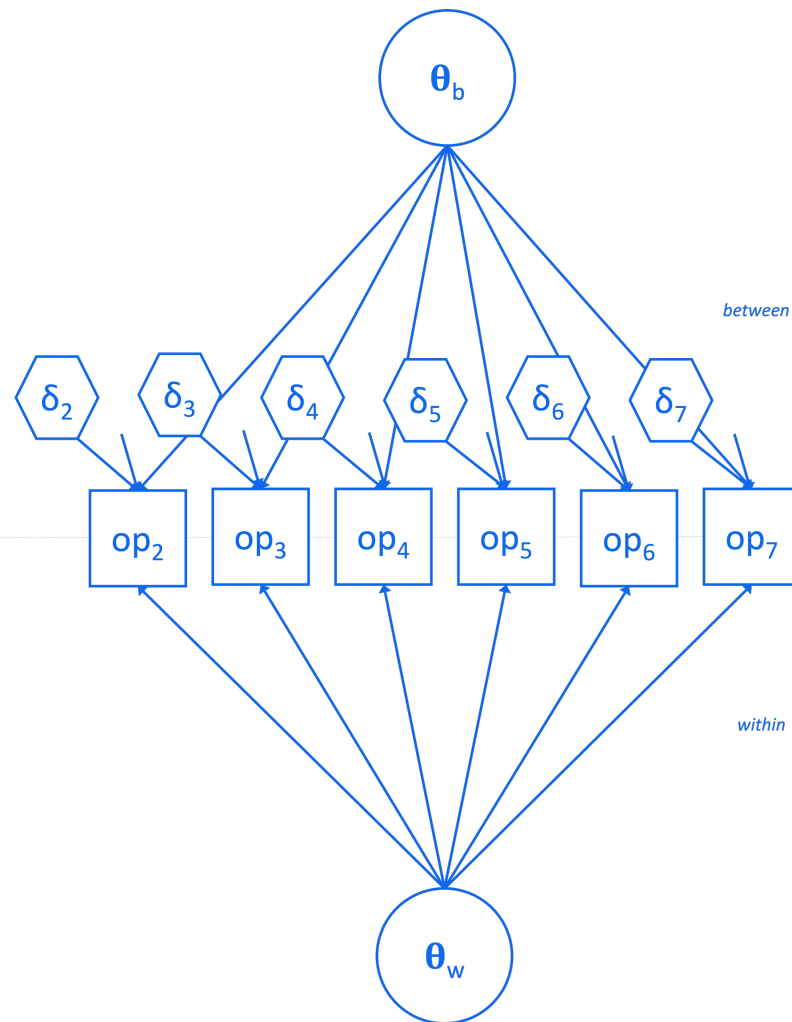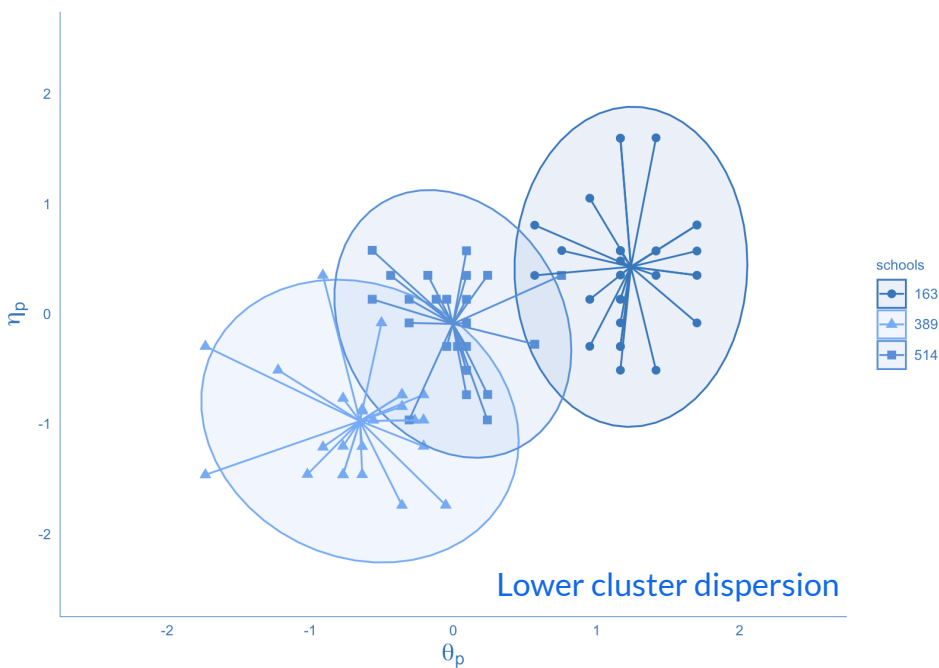

$$\theta_p = \theta_w + \theta_b$$

For Problem 1

Alternative Model

$$y_{ij} = \alpha + \gamma_w \boldsymbol{\theta}_w + \gamma_b \boldsymbol{\theta}_b + u_j + \epsilon_{ij}$$

Problem 2

What if informants vary in their ratings?

Informants' agreement as score dispersion

Lower cluster dispersion

If Schools show low dispersion, then there is not so much of a problem.

If schools do not overlap greatly, it seems easy to draw a main trend between schools to make inferences.

Higher cluster dispersion
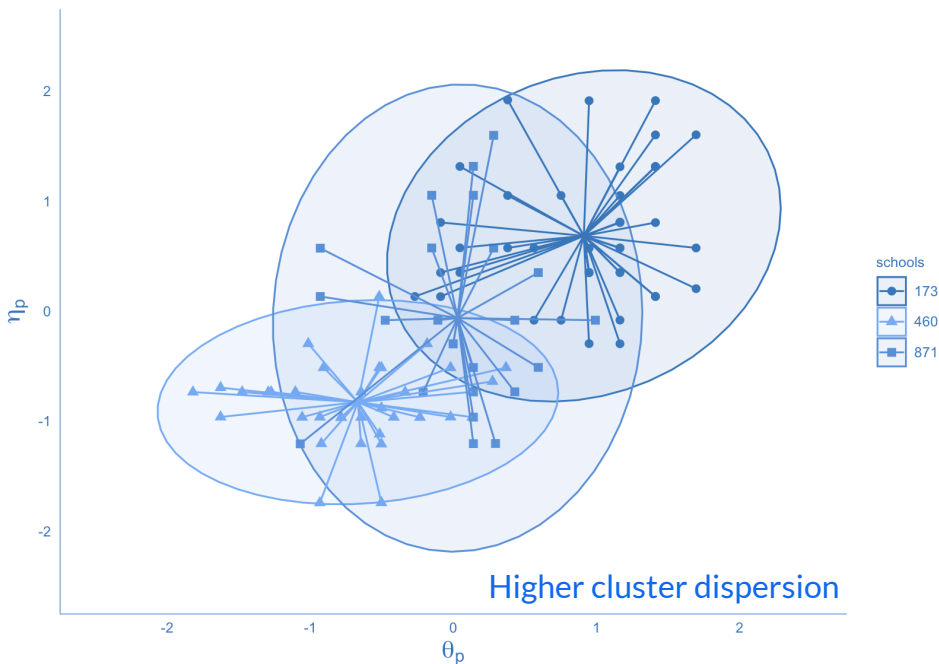
In contrast, if schools display higher dispersion, then we can find much more overlap between clusters.

With varying dispersion among clusters, is less clear what is the main trend between schools

Agreement as dispersion
clusters with varying agreement

How much agreement is needed between informants?

- A common advice in the literature of organizational behavior **is to assure certain level of cluster agreement** (rwg < .70). In practice, this leads to sample lost.

- Moreover, the uncertainty around agreement indexes depends on the intra class correlation and the group size of the clusters (Lüdtke & Robitzsch, 2009). **Thus, rules of thumb are not applicable.**

We can address this problem with a **moderation term**

- Using an index of disagreement between students from the same cluster, we can **penalize clusters with low agreement.** And using Johnson-Neyman plots we can identify how much disagreement is tolerable before it compromise the expected effects.

- index of disagreement between informants: standard deviation of $\boldsymbol{\theta}_p$ by cluster ($\delta_{.j}$).

- $y_{ij} = \alpha + \gamma_w \boldsymbol{\theta}_w + \gamma_b \boldsymbol{\theta}_b + \beta_b(\delta_{.j} - \bar{\delta}_{..}) + \lambda_b(\delta_{.j} - \bar{\delta}_{..})\boldsymbol{\theta}_b + u_j + \epsilon_{ij}$
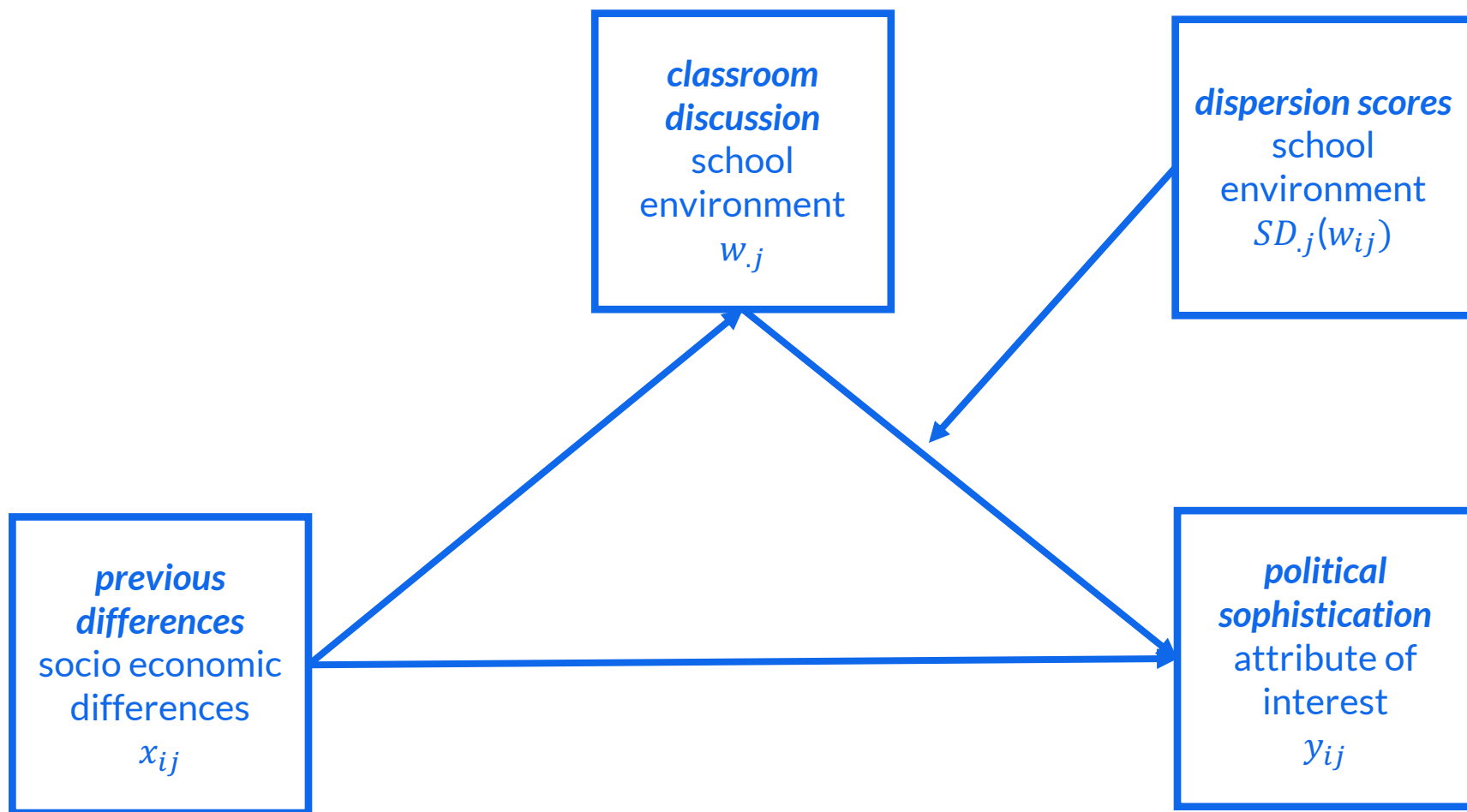
For Problem 2

Dispersion score model

$$y_{ij} = \alpha + \gamma_w \mathbf{\theta}_w + \gamma_b \mathbf{\theta}_b + \beta_b \left(\delta_{.j} - \bar{\delta}_{..} \right) + \lambda_b \left(\delta_{.j} - \bar{\delta}_{..} \right) \mathbf{\theta}_b + u_j + \epsilon_{ij}$$

Let's see an illustration of problem 2

classroom discussion and civic knowledge

Perú, ICCS 2016

Conceptual model
dispersion score model

classroom discussion school environment $w_{.j}$

dispersion scores school environment $SD_{.j}(w_{ij})$

previous differences socio economic differences x_{ij}

political sophistication attribute of interest y_{ij}

$$y_{ij} = \alpha + \underbrace{\pi_w ses_w + \pi_b ses_b}_{gap} + \underbrace{\gamma_w \boldsymbol{\theta}_w + \gamma_b \boldsymbol{\theta}_b}_{practice} + \underbrace{\beta_b(\delta_{.j} - \bar{\delta}_{..})}_{dispersion} + \underbrace{\lambda_b(\delta_{.j} - \bar{\delta}_{..})\boldsymbol{\theta}_b}_{moderation} + u_j + \epsilon_{ij}$$

Illustration of problem 2
Perú (ICCS, 2016)

dati invalsi
uno strumento
per la ricerca
e la didattica

Table 2: Unstandardized estimates on release civic knowledge scores (Perú, ICCS 2016)

Fixed effects	Gap	Practice	Dispersion	Moderation
intercept	432.21 ***	427.81 ***	427.83 ***	428.98 ***
socio-economic status (student deviations)	14.23 ***	13.81 ***	13.81 ***	13.80 ***
socio-economic status (school means)	64.84 ***	48.45 ***	48.31 ***	45.98 ***
θ_w classroom discussion (student deviations)		26.03 ***	26.03 ***	26.04 ***
θ_b classroom discussion (school means)		81.31 ***	81.51 ***	89.58 ***
dispersion score $(\delta_{\cdot j} - \bar{\delta}_{\cdot\cdot})$			-2.41	-24.26
Interaction $(\delta_{\cdot j} - \bar{\delta}_{\cdot\cdot})\theta_b$				-237.13 **
Random Part				
students	4602.20 ***	4401.68 ***	4399.69 ***	4397.39 ***
schools	1899.69 ***	1089.53 ***	1096.82 ***	1002.91 ***

Illustration of problem 2
Perú (ICCS, 2016)

Range of values where the estimand of γ_b is robust to dispersion (i.e., lack of agreement)

Range of values where the estimand of γ_b are compromised by dispersion (i.e., lack of agreement)

.15   .73   .91

Slope = γ_b

Moderator = $\delta_{.j}$
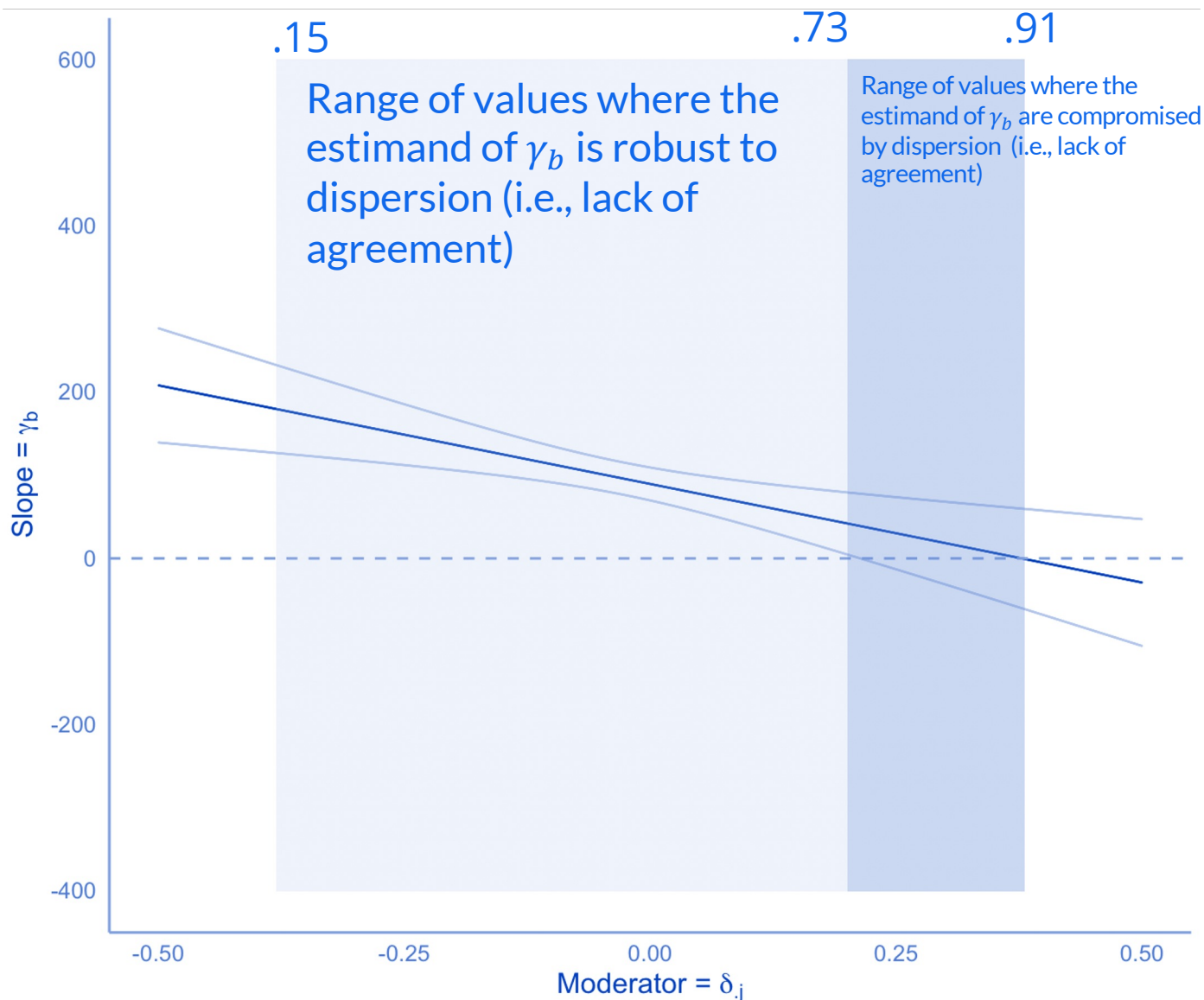
Conclusions

**Take-home message for
the two problems**

Summary
open classroom discussion

dati invalsi
uno strumento
per la ricerca
e la didattica

Conclusion

- To conclude, with the present results we can argue, that in spite of schools students composition (e.g. SES) and students varying levels of inter-rater agreement, **open classroom discussion has a relationship effect on the civic knowledge of the students.**

    - Yet in schools environments where the inter-rater agreement is too low, then the expected effect of classroom discussion is compromised.

- If the reflective nature of OPD is ignored and a compositional model is used, in 9 out of 24 countries **we are at risk of making wrong conclusions and wrong recommendations.**

- How much dispersion between informants can be tolerated? It can be estimated. So far, we've found similar results for dispersion effects in 6 countries out of 24 in ICCS 2016.

In summary

- **The under estimation of school factors can be avoided.** But we need to specify adequate inference models. Otherwise, using wrong models over reflective measures can lead to wrong conclusions, and **limit possible public policy recommendations.**

Grazie mille!

References

Bliese, P. D., Maltarich, M. A., Hendricks, J. L., Hofmann, D. A., & Adler, A. B. (2018). Improving the measurement of group-level constructs by optimizing between-group differentiation. Journal of Applied Psychology. https://doi.org/10.1037/apl0000349

Carrasco, D., & Torres Irribarra, D. (2018). The Role of Classroom Discussion. In A. Sandoval-Hernández, M. M. Isac, & D. Miranda (Eds.), Teaching Tolerance in a Globalized World (pp. 87–101). https://doi.org/10.1007/978-3-319-78692-6

Ehman, L. H. (1969). An Analysis of the Relationships of Selected Educational Variables with the Political Socialization of High School Students. American Educational Research Journal, 6(4), 559–580. https://doi.org/10.3102/00028312006004559

Lüdtke, O., & Robitzsch, A. (2009). Assessing Within-Group Agreement: A Critical Examination of a Random-Group Resampling Approach. Organizational Research Methods, 12(3), 461–487. https://doi.org/10.1177/1094428108317406

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. Contemporary Educational Psychology, 34(2), 120–131. https://doi.org/10.1016/j.cedpsych.2008.12.001

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. Learning Environments Research, 9(3), 215–230. https://doi.org/10.1007/s10984-006-9014-8

Schlozman, K. L., Verba, S., & Brady, H. E. (2012). The Unheavenly Chorus: Unequal Political Voice and the Broken Promise of American Democracy. Princeton University Press.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct Meaning in Multilevel Settings. Journal of Educational and Behavioral Statistics, 41(5), 481–520. https://doi.org/10.3102/1076998616646200

Torres-Irribarra, P., & Carrasco, D. (2019). Ajuste de modelos multinivel de crédito parcial, para la comparación en ambientes escolares: Una aplicación. Primer Encuentro Nacional de Métodos y Técnicas de Investigación Social. Rancagua, Chile.

Carrasco, D., PhD,

Centro de Medición MIDE UC
Pontificia Universidad Católica de Chile
https://dacarras.github.io/