To PV or not PV

How to get random slopes with ILSA studies

Carrasco, D., Centro de Medición MIDE UC, Pontificia Universidad Católica de Chile
Torres Irribarra, D., Escuela de Psicología, Pontificia Universidad Católica de Chile

VII Seminar:
"Promuovere l'utilizzo dei dati INVALSI nella ricerca scientifica e nella didattica"
ROME, October 17th- 19th, 2024

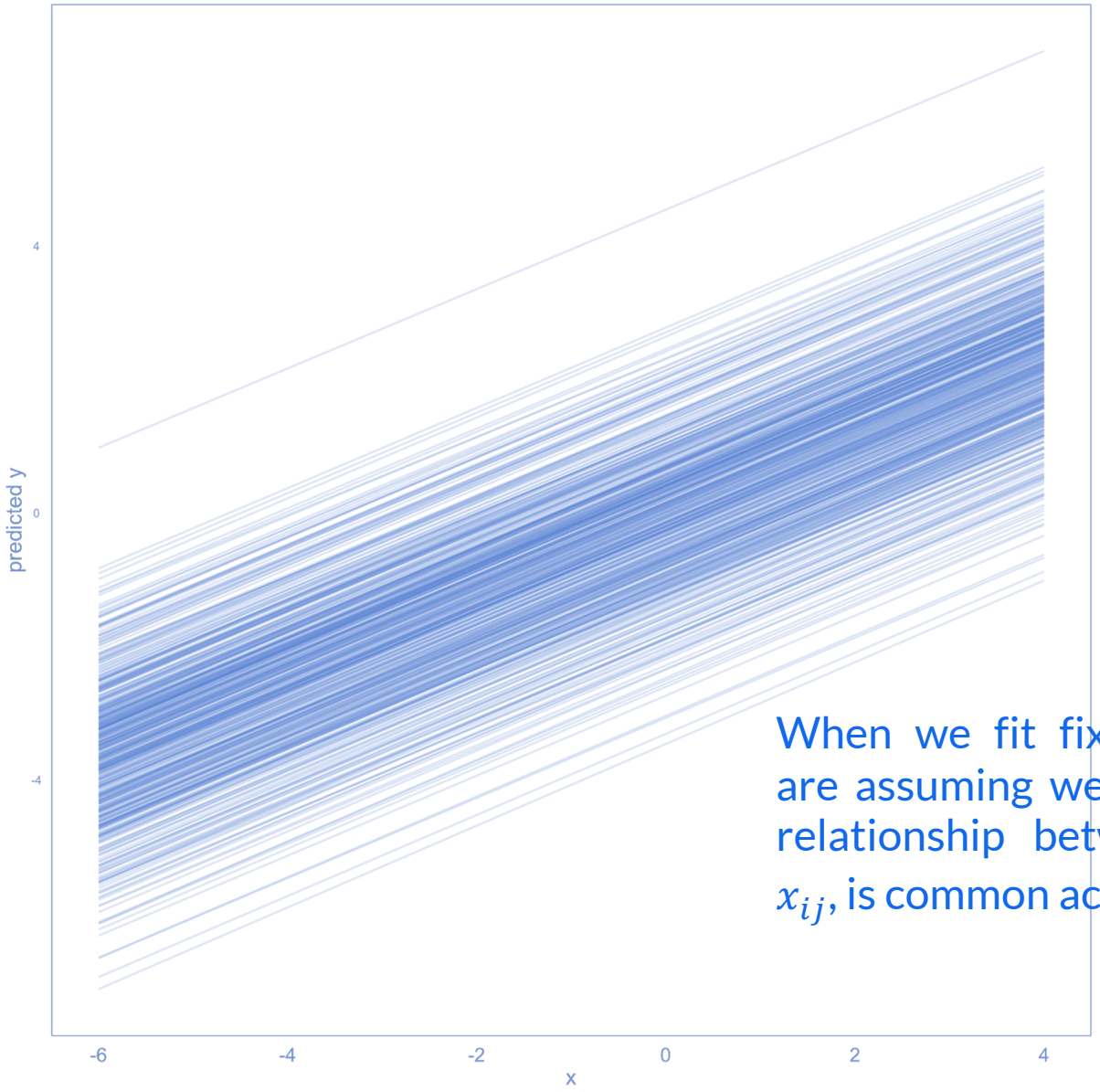
https://github.com/dacarras/invalsi_2024_pv

Introduction

School variability and random slopes

Introduction

How fixed slopes looks likes

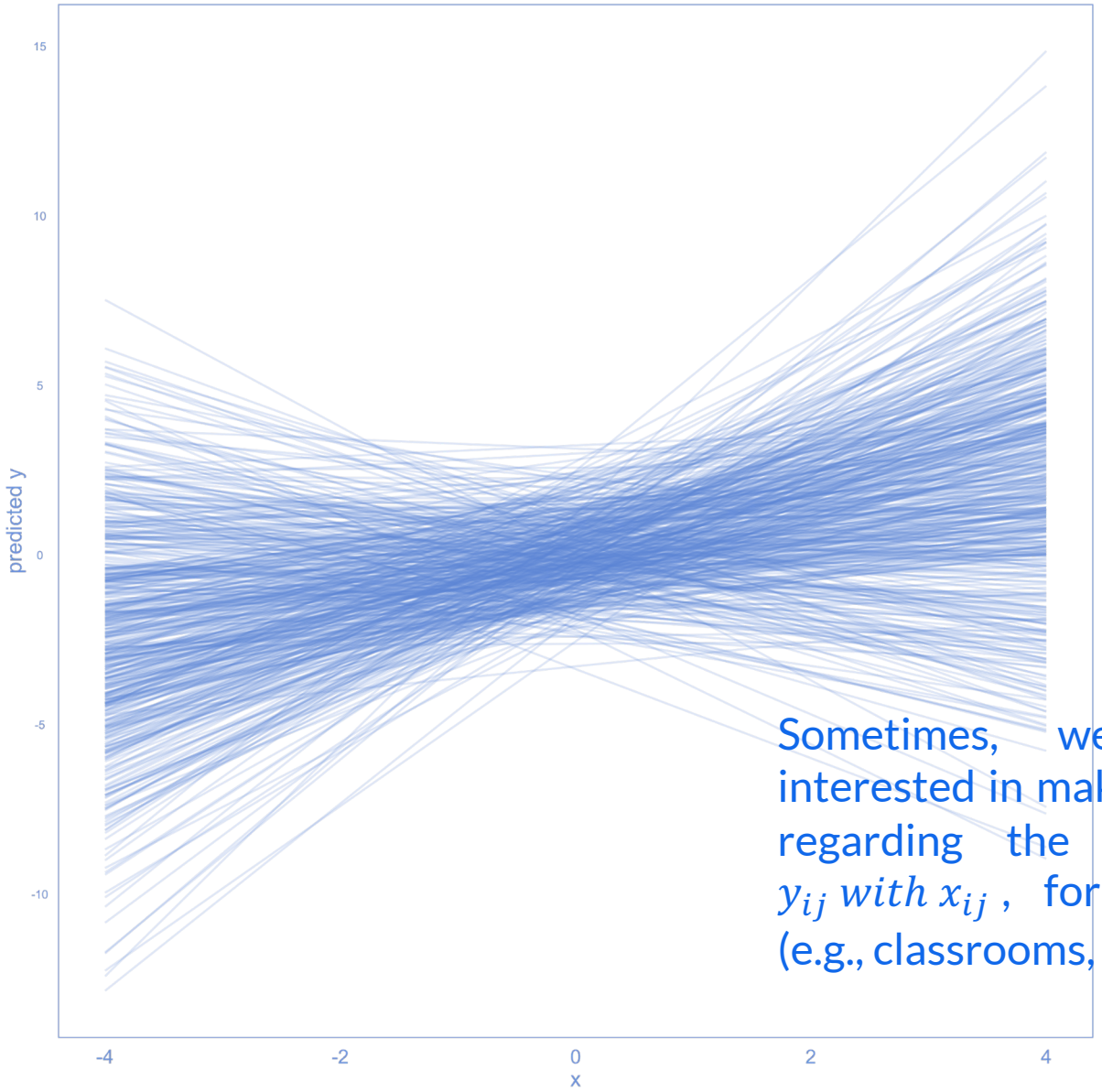


When we fit fixed slopes, we are assuming we can claim the relationship between y_{ij} and x_{ij}, is common across clusters.

https://github.com/dacarras/invalsi_2024_pv

Introduction

How random slope looks like



Sometimes, we might be interested in making inferences regarding the variability of $y_{ij}\ with\ x_{ij}$, for each cluster (e.g., classrooms, schools).

https://github.com/dacarras/invalsi_2024_pv

Introduction

Random Slopes in Large Scale assessment

Random slopes are of special interest in the study of school effectiveness research

- Specially for **compensatory hypothesis** (e.g., Campbell, 2008)
  - Are there school factors that can ameliorate the socioeconomic gap?
- And also, for **differential effectiveness** research questions (e.g., Caro et al., 2016)
  - Are there any teacher practices that benefit students differently?

Large scale assessment studies are **ideal scenario** for these research inquiries

- These includes a **large collection of schools/classrooms** (150 or more), with a **sampling design** that can maximize variance of interest
- Yet, there is a **caveat** regarding the scores of interest
  - Test scores are represented with **plausible values.**

Problem

Limitations of the Plausible values

Problem

Random slopes are not easily retrievable

Test outcomes scores in ILSA are very special

- Test outcomes are represented by **Plausible values**
- Plausible values are response **model realizations** (Wu, 2005)
- These can help us to retrieve expected population estimates of **congenial models** (i.e., nested models onto the generative model for the realizations) (Braun et al., 2017).
  - Yet, if the researcher model is not congenial, the obtained estimates may not recover the estimands he or she is looking for.
  - Specially, random slopes and interaction terms.
- Current conditional models used to generate plausible values **do not include terms to represent slope variabilities** (Zheng, 2024).
  - As such, model fitted onto the plausible values would not "recover" the expected random slopes estimates and variance.

This limitation compromise different research strands such as:

- **Compensatory hypothesis (e.g., Campbell, 2008), Differential Effectiveness hypothesis (e.g., Caro et al, 2016), … among other research inquiries that requires random slopes fitted onto plausible values.**

Method

How are we going to illustrate the problem?

Illustration

Plausible Values are not a good tool to retrieve random slopes

Methods

- To illustrate the problem we are pointing out, we will use data from International Civic and Citizenship Education Study 2009 (ICCS 2009).
  - We will fit a series of mixed models including civic knowledge as the response variable. We will use sex and socioeconomic status (ses) as covariates for different models, for each country. We will fit models of the following form:

$$y_{ij} = \alpha + \beta_{1j}(x_{ij} - \bar{x}_{.j}) + u_{0j} + u_{1j}(x_{ij} - \bar{x}_{.j}) + \epsilon_{ij}$$

- We will fit the model of interest **onto the five plausible values** as common guidelines would suggest (Von Davier et al., 2009; Rutkowski et al., 2010; Carsten et al., 2010).
  - We will use a **Likelihood Ratio test adapted for imputed data** (Grund et al., 2023) to make inferences regarding the random slope variance.
- We will proceed similarly, fitting the same model onto the IRT WLE scores also shared by the ICCS 2009 study (i.e., NWLCIV scores).
- We standardized estimates using the score respective standard deviation (PV = 100, WLE = 10).

Illustration

Plausible Values are not a good tool to retrieve random slopes

Rationale

- The IRT WLE scores are expected to not to be affected by the limitations of the conditioning model.
- IRT WLE scores do not carry the measurement model error or uncertainty, yet these may provide attenuated estimates (Bhaktha et al, 2021).
- However, if these scores are generated with (Diakow, 2013)
  - A large number of items (e.g. more than 25 items)
  - Present high reliability
  - And good spread of item locations (i.e., high test information coverage)
  - **These bias should be ameliorated.**

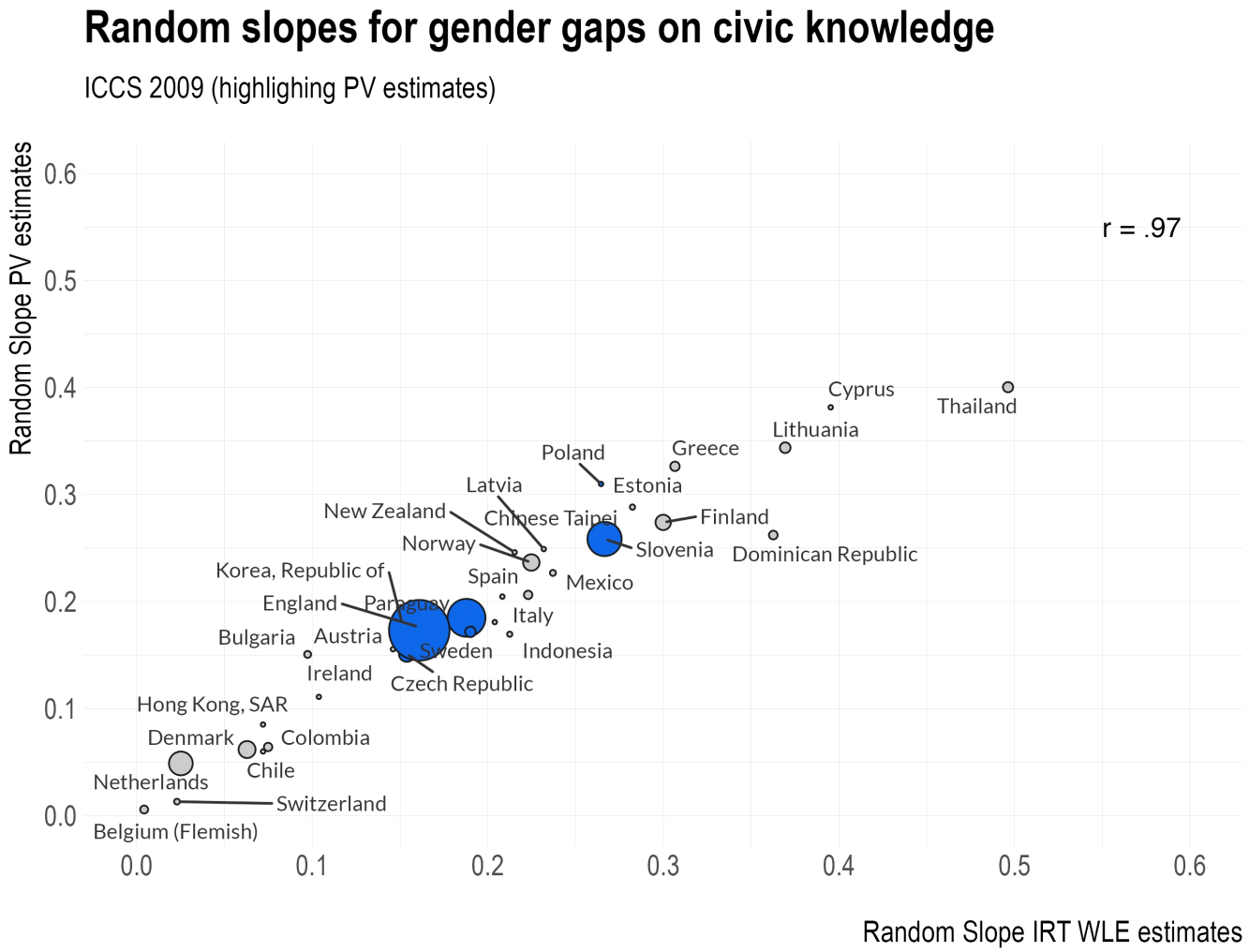
- IRT WLE from ICCS 2009 fulfill all these requirements (i.e., generated with 79 items, with a good spreading of difficulties, and high reliability (> .8) (See Schulz et al., 2011),
- Thus, comparing the results of models fitted onto PV scores, and onto IRT WLE scores can give use an approximation of how risky is to rely on PV scores alone when our research inquiry requires random slopes.

Results I

Random slopes for gender gaps

Random Slopes

Gender gaps on civic knowledge

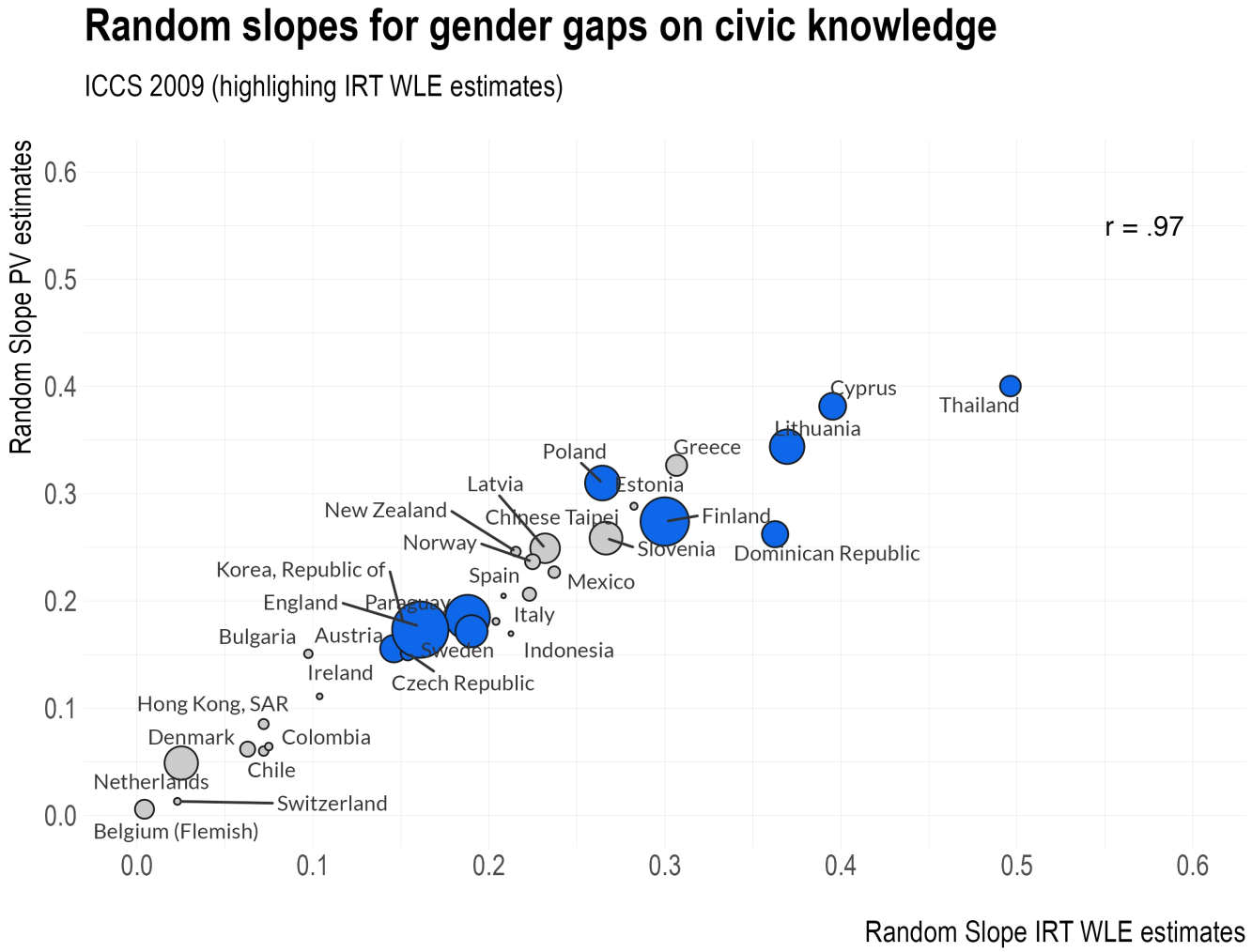


Note: blue dots highlights random slopes via D2 stat for PV (Grund et al., 2023); bubble size is relative variance of the random slope.

https://github.com/dacarras/invalsi_2024_pv

Random Slopes

Gender gaps on civic knowledge



Note: blue dots highlights random slopes via LRT mixture on IRT WLE scores; bubble size is relative variance of the random slope.

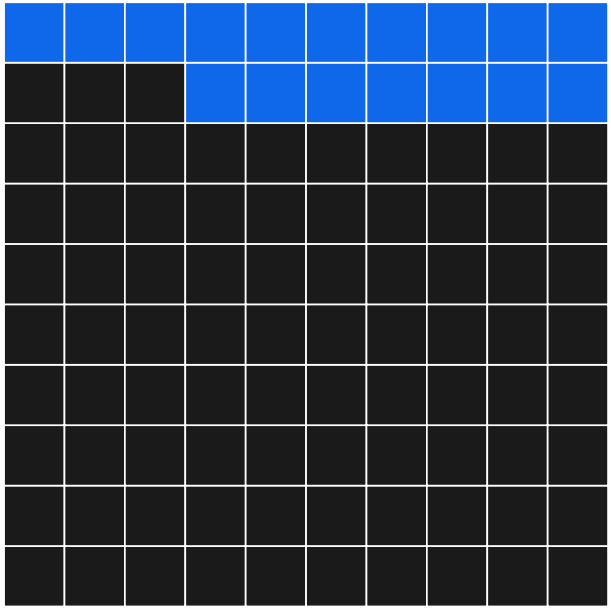
https://github.com/dacarras/invalsi_2024_pv

Random Slopes

Gender gaps on civic knowledge

Estimates with Plausible Values

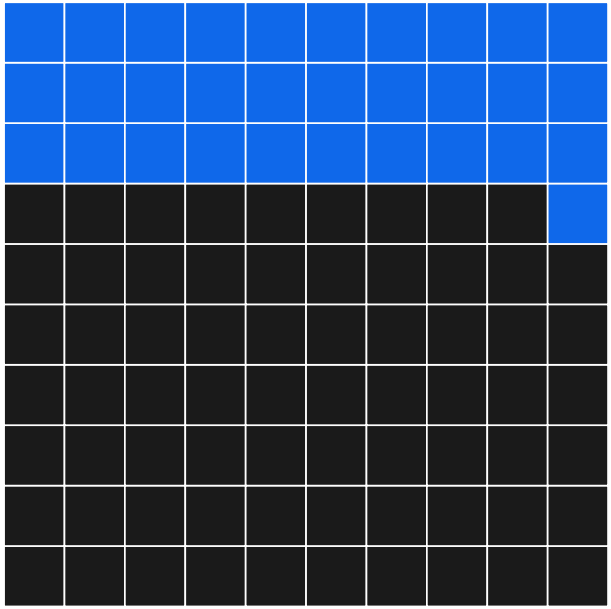
D2 stat (Grund, 2023), 6 of 35 are random



Fixed slope (83%) Random slope (17%)

Estimates with IRT WLE scores

LRT mixture, 11 of 35 are random



Fixed slope (69%) Random slope (31%)

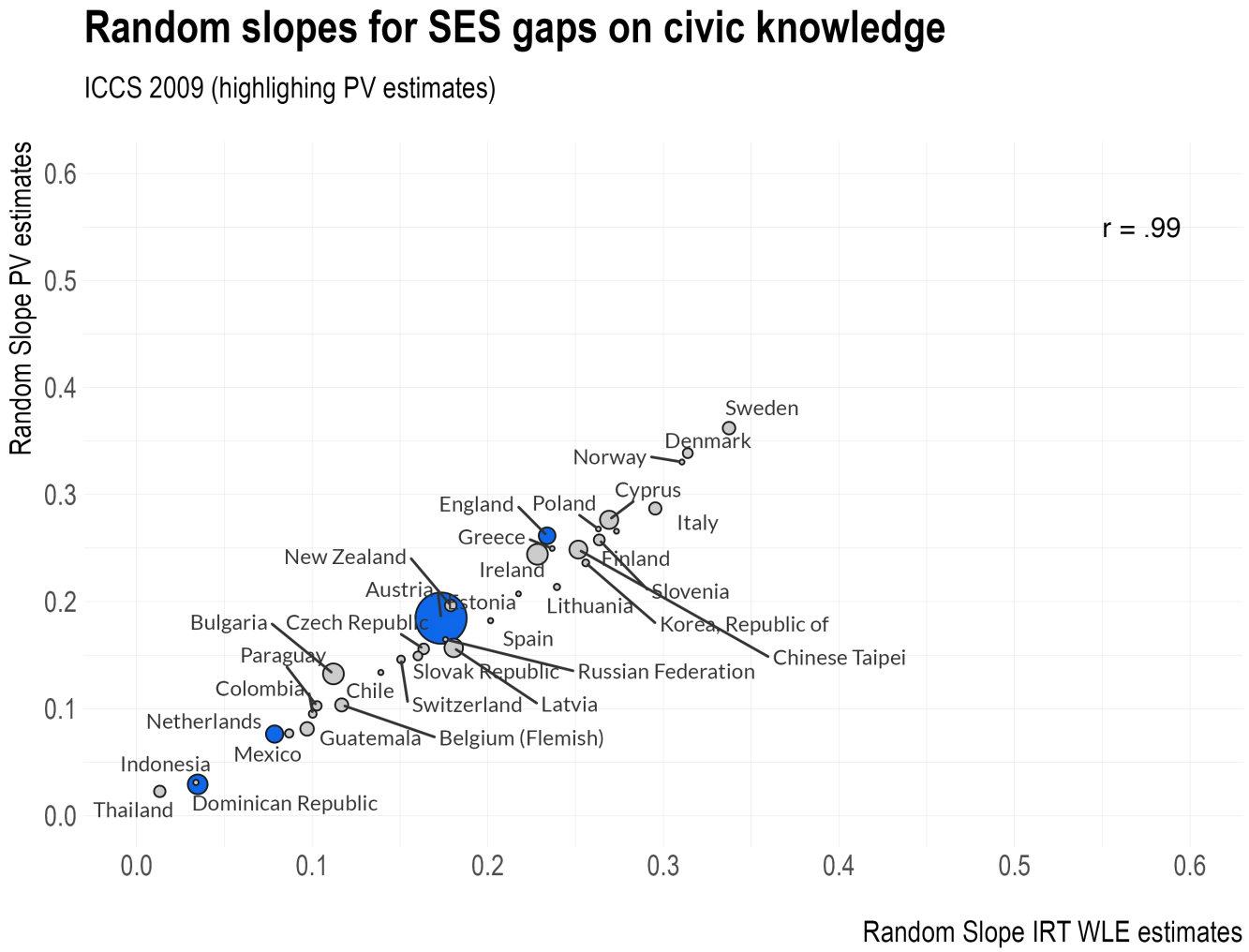
https://github.com/dacarras/invalsi_2024_pv

Results II

Random slopes for socioeconomic gaps

Random Slopes

SES gaps on civic knowledge

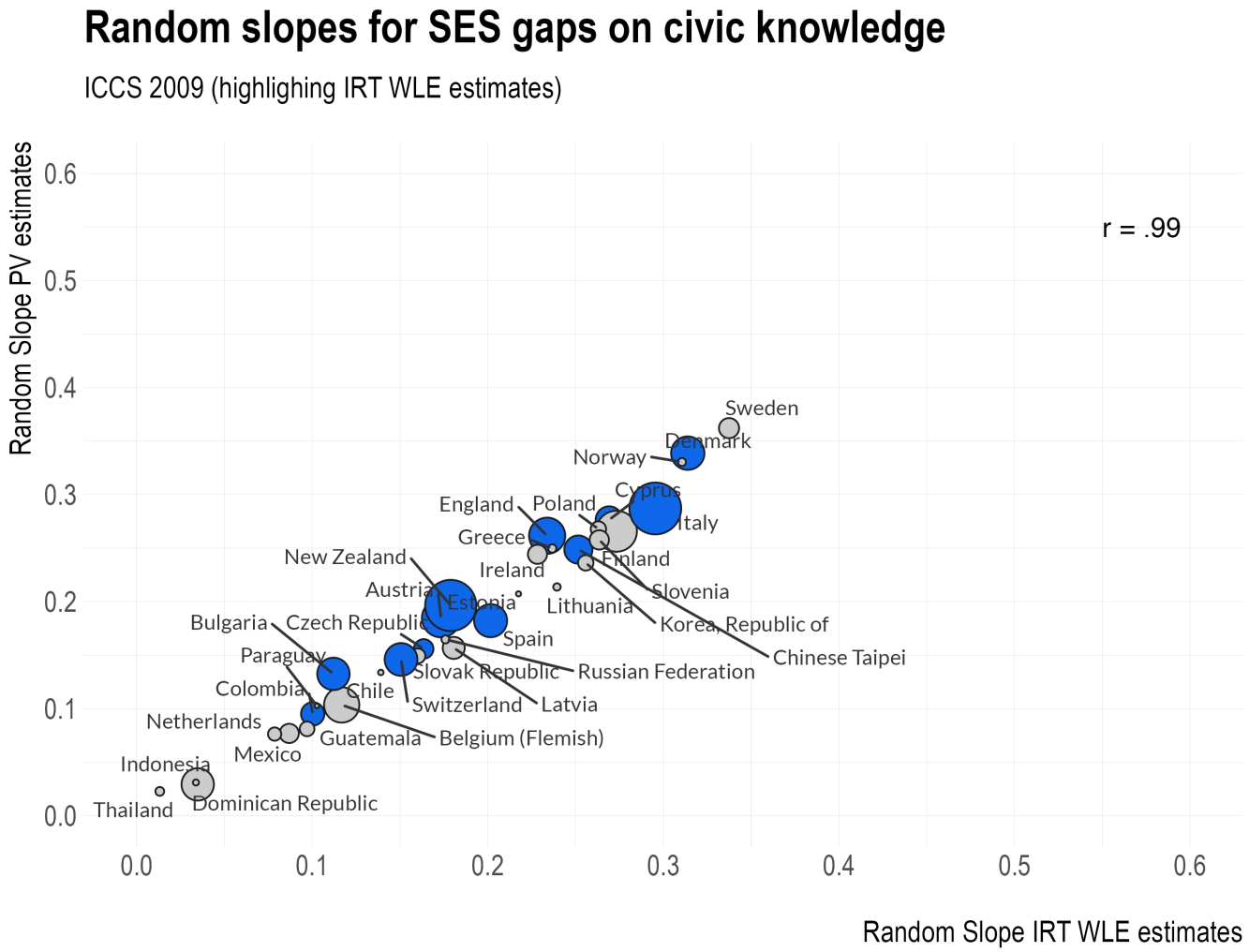


Note: blue dots highlights random slopes via D2 stat for PV (Grund et al., 2023); bubble size is relative variance of the random slope.

https://github.com/dacarras/invalsi_2024_pv

Random Slopes

SES gaps on civic knowledge



Note: blue dots highlights random slopes via LRT mixture on IRT WLE scores; bubble size is relative variance of the random slope.

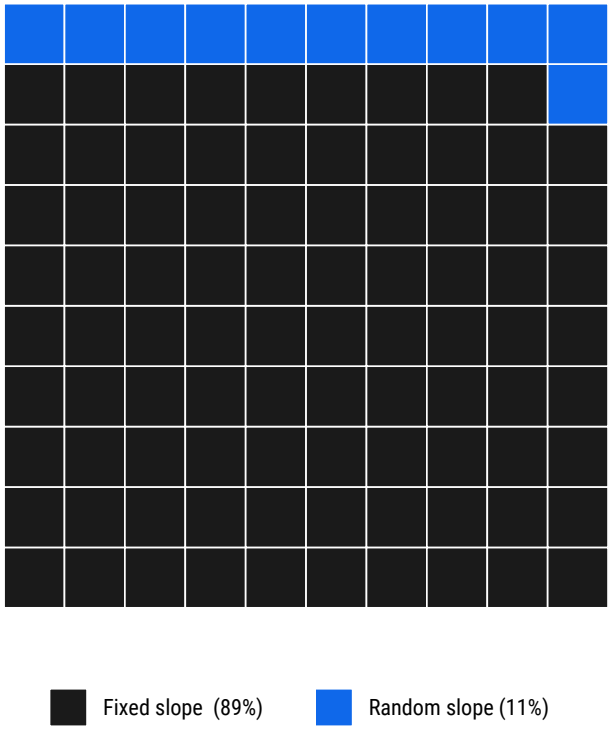
https://github.com/dacarras/invalsi_2024_pv

Random Slopes

SES gaps on civic knowledge

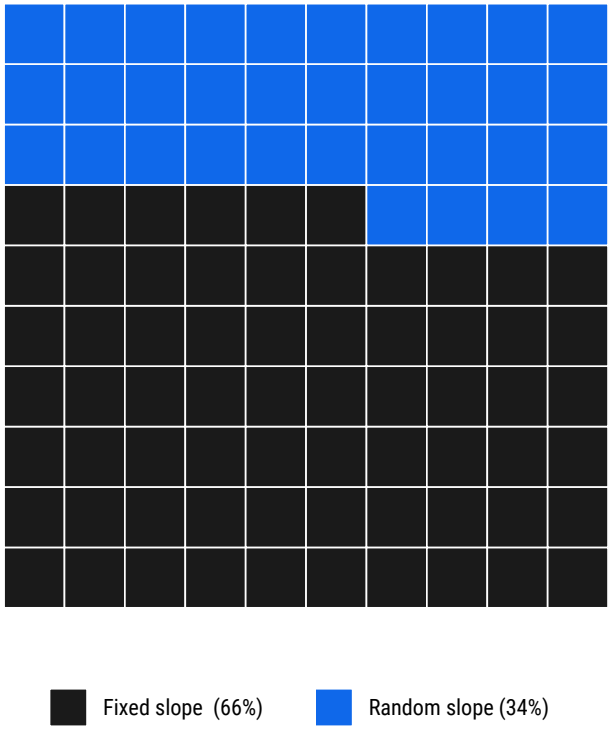
Estimates with Plausible Values

D2 stat (Grund, 2023), 4 of 35 are random



Estimates with IRT WLE scores

LRT mixture, 12 of 35 are random



https://github.com/dacarras/invalsi_2024_pv

Discussion

Summary and take-home message

Summary

Random slopes estimates with large scale assessment

Conclusion

- Inferences based on random slopes estimated with PV scores can be misleading.
- PV estimates do not correctly recover the variance of the random slope estimates(Zheng, 2024).
- For the covariates in this study, **it is 2 to 3 times more likely to recover statistically significant random slopes when using IRT WLE scores than when using PV scores.**
- If the test has a good number of items, well-distributed item locations, and high reliability, **an alternative approach is to generate random slope estimates using the IRT WLE scores of the target test.**

https://github.com/dacarras/invalsi_2024_pv

Summary

Random slopes estimates with large scale assessment

Take home message for secondary users:

- For research inquiries requiring random slopes, using PV scores may not be advisable. These scores seems to to underestimate the random slope variance.

- If available, a better choice is to use the IRT WLE scores for the same purpose
- If not available, is better to generate the IRT WLE scores first, and fit models with these new scores in a second step.
- Fit a full fledge explanatory response model, including the desire random slope
  - Bear in mind the time needed to fit these models may be greater than fitting random slope models onto observed scores.

Summary

Random slopes estimates with large scale assessment

Take home message for ILSA providers:

- PV scores enables the generation of results for many research inquiries of interest.
- However, these scores do not provide adequate estimates outside the space of the congenial models.
- A simple aid for secondary users is to provide the IRT WLE scores of the target test, and its standard errors.
  - These two elements may help researchers to retrieve estimates of interest outside the congenial space of the conditional models of the generated plausible values (Diakow, 2013).

Further research

- Complement the present results with simulations studies to identify magnitude of bias, and other boundary conditions (e.g., estimate size, variance size, floor and ceiling effects of test item locations, survey weights scaling).

https://github.com/dacarras/invalsi_2024_pv

Grazie mille!

Carrasco, D., PhD,

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

References

Bhaktha, N., & Lechner, C. M. (2021). To Score or Not to Score? A Simulation Study on the Performance of Test Scores, Plausible Values, and SEM, in Regression With Socio-Emotional Skill or Personality Scales as Predictors. Frontiers in Psychology, 12(October). https://doi.org/10.3389/fpsyg.2021.679481

Braun, H., & von Davier, M. (2017). The use of test scores from large-scale assessment surveys: psychometric and statistical considerations. Large-Scale Assessments in Education, 5(1), 17. https://doi.org/10.1186/s40536-017-0050-x

Campbell, D. E. (2008). Voice in the Classroom: How an Open Classroom Climate Fosters Political Engagement Among Adolescents. Political Behavior, 30(4), 437–454. https://doi.org/10.1007/s11109-008-9063-z

Caro, D., Lenkeit, J., & Kyriakides, L. (2016). Teaching strategies and differential effectiveness across learning contexts: Evidence from PISA 2012. Studies in Educational Evaluation, 49, 30–41. https://doi.org/10.1016/j.stueduc.2016.03.005

Carstens, R., & Hastedt, D. (2010). The effect of not using plausible values when they should be : An illustration using TIMSS 2007 grade 8 mathematics data. 4th IEA International Research Conference IRC 2010, 12.

Grund, S., Lüdtke, O., & Robitzsch, A. (2023). Pooling Methods for Likelihood Ratio Tests in Multiply Imputed Data Sets. Psychological Methods.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International Large-Scale Assessment Data: Issues in Secondary Analysis and Reporting. Educational Researcher, 39(2), 142–151. https://doi.org/10.3102/0013189X10363170

Stapleton, L. M. (2013). Incorporating Sampling Weights into Single- and Multilevel Analyses. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), Handbook of International Large scale Assessment: background, technical issues, and methods of data analysis (pp. 363–388). Chapman and Hall/CRC.

Schulz, W., Ainley, J., & Fraillon, J. (2011). ICCS 2009 Technical Report (W. Schulz, J. Ainley, & J. Fraillon (eds.)). International Association for the Evaluation of Educational Achievement (IEA). http://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/ICCS_2009_Technical_Report.pdf

Von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 2, 9–36. http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf

Wu, M. (2005). The role of plausible values in large-scale surveys. Studies in Educational Evaluation, 31(2–3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Annexes

Method details

Methods

Sample selection and estimate

Method details

- We exclude Malta, Lichtenstein and Luxembourg country samples, due to low count of clusters (n < 50).
  - We assume clusters are primary sample using within strata (i.e., jackknife zones).
  - Malta, Lichtenstein and Luxembourg present IDSCHOOLS crossed with jackknife zones, which may entail the generation of “pseudo clusters”. These are generated primary sample units (PSU) generated for error correction, which may not represent observed schools or intact classrooms.
  - As such, we exclude these students' sample for the current illustration, so we can assure to the best of our knowledge that the estimated random slopes refer to school/classrooms grouping students' students throughout a school year.
- Estimates are pseudo maximum likelihood estimates
  - We fit mixed models, while including disaggregated survey sample weights
  - Cluster survey weights were scaled to the effective sample size (Method B, Stapleton, 2013).
- Software tools
  - We use Mplus 8.11 software to fit the specified models.
  - We use MplusAutomation to fit models in batches and retrieve estimates to generate plots