USAID
FROM THE AMERICAN PEOPLE

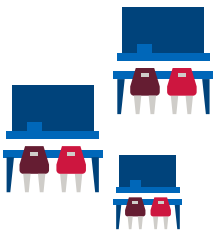
Intracluster Correlations for Early Reading Evaluations

By Alicia Menendez and Alejandro Ome | NORC at the University of Chicago

Power analysis is a critical design component of any impact evaluation. A sample that is too small will fail to estimate a meaningful impact with acceptable precision, while samples that are too big are costly in terms of financial resources and respondent fatigue. Selecting an appropriate sample size requires researchers to assume certain parameters, like the expected program impact. Furthermore, in evaluations where treatment is assigned to units larger than individual participants, like schools, clustering also needs to be taken into account in power calculations. In these cases, researchers must make assumptions about the intracluster correlation (ICC).

In general, researchers base their assumptions on previous experience or expert knowledge. A popular rule of thumb is to assume that the ICC is 0.10, but previous work by Kelcey, Shen and Spybrook (2016) finds that this might be too low for education samples in low and middle income countries. These authors document ICC for reading assessments of 6th-grade students in several African countries and find that ICC are between 0.08 and 0.60. To provide estimates for ICC for early grade reading, we use data from numerous evaluations that collected early grade reading assessments (EGRA). We also produced estimates for how much of the variation in the outcomes of interest is explained by observable characteristics (R-squared). This parameter, although less critical for power analysis, also needs to be considered in power calculations. We focus on oral reading fluency (ORF), as this is the outcome most often referenced in early grade reading assessments.

Intracluster correlation



ICC refers to the proportion of the variance of the outcome of interest that is explained by the variance between groups. If the overall variance is fully explained by variance between groups (i.e., ICC is 1), then individuals within each group are identical, and in terms of power analysis the sample size is not the number of individuals but the number of groups. When very little of the variance is explained by variation between groups and most is explained by variation within groups (i.e., ICC is close to 0), the clustered structure has a small effect on power, because it is closer to sampling students randomly and not following a clustered design.

The table below shows the parameters for each evaluation. All these evaluations had two waves of data collection, some follow the participants longitudinally, and others are repeated cross sections. Most projects were fielded in multiple geographical and/or language regions within the same country. The table also shows the parameters by grade as appropriate.

To calculate the ICC we used the Stata command loneway. The results show how different the ICC can be in different contexts.

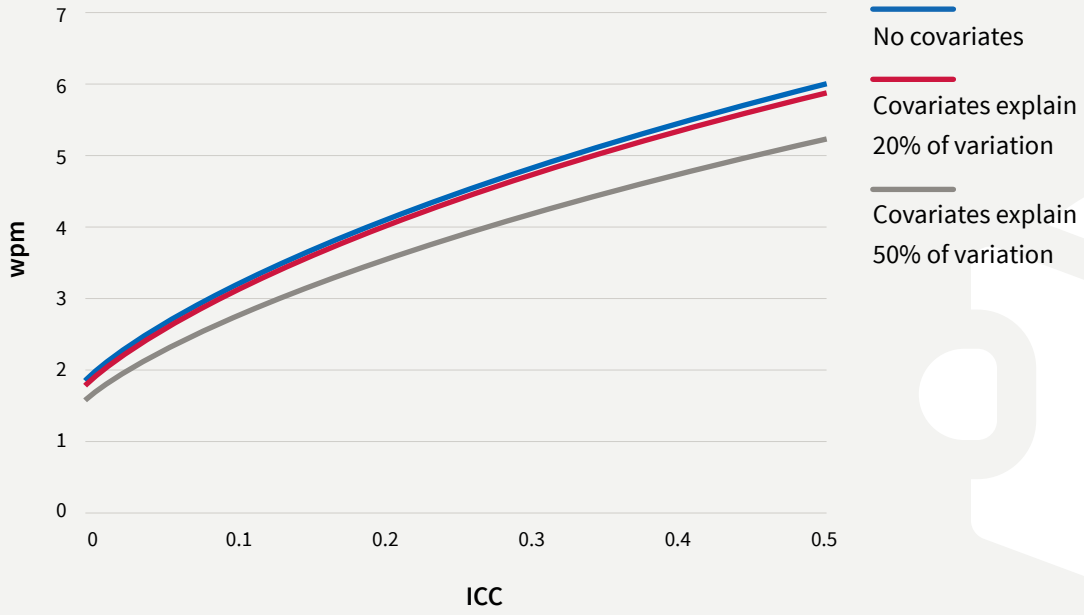
Table 1: ICC and R-squared for Reading and Access Impact Evaluations

Project (Region/Language)		Grade at Baseline	Grade at Endline	ICC at Baseline	R-squared	Longitudinal/ Cross sectional
Zambia Makhalidwe Athu		2-3	3-4	0.15	0.44	Longitudinal
Reading for Ethiopia's Achievement Developed Community Outreach program	Amhara	2	4	0.14	0.57	Longitudinal
	Oromia	2	4	0.18	0.51	
South Africa Story Powered School Program	Eastern Cape	2	3	0.16	0.41	Longitudinal
		3	4	0.08	0.64	
		4	5	0.07	0.71	
	KwaZulu-Natal	2	3	0.11	0.49	
		3	4	0.11	0.65	
		4	5	0.05	0.65	
National Early Grade Reading Program in Nepal		1	1	0.35	0.23	Cross sectional (School panel)
		2	2	0.43	0.28	
		3	3	0.40	0.36	
Read Liberia		2	3	0.33	0.21	Cross sectional (School panel)
Pakistan Reading Project	Khyber Pakhtunkhwa	3	3	0.27	0.33	Cross sectional (School panel)
	Balochistan	3	3	0.34	0.38	
Uganda Literacy Achievement and Retention Activity	Luganda	1	3	0.06	0.29	Cross sectional (School panel)
	Runyankore/Rukiga	1	3	0.18	0.28	
	English	1	3	0.16	0.42	

While the ICC is relatively low in the Ugandan and South African studies, in the Nepal study it was much higher. To calculate the R-squared we follow different approaches depending on whether we have longitudinal or cross-sectional data. For longitudinal data we extract the adjusted R-squared from a regression where the dependent variable is ORF at endline, and the covariates are ORF at baseline and socioeconomic indicators, namely student's sex, age, and a household asset index. In the case of cross-sectional data we extracted the R-squared from school fixed effects regressions that included the same sociodemographic variables that were included in the longitudinal analyses. No treatment variables were included in any of the models to calculate the R-squared so we capture only the variation explained by covariates. The results for the R-squared show that longitudinal data provide more predictive power than repeated cross sections. In effect most R-squared for cross sectional data are below 40 percent, while for longitudinal data most are above 50 percent. While higher R-squared is one of the advantages of collecting longitudinal data, a major risk of using longitudinal data is survey attrition, which also needs to be considered in power analysis. For the three projects for which we have longitudinal data, attrition rates vary between 10 and 50 percent. Ome, Ardington and Menéndez (2021) use data from these three projects to discuss methods to correct for attrition in program evaluations.

To show how sensitive sample size calculations are to different levels of ICC, we simulate scenarios for different values of ICC to estimate what would be the minimum detectable effect size (MDES) in terms of words per minute. The figure shows simulations for an experiment where there are 40 schools in the treatment group, 40 schools in the control group, and in each school 20 children are surveyed. To focus on changes of the ICC we keep sample size, number of clusters, alpha (5 percent) and power constant (80 percent). The blue line shows results where no covariates are included in the impact evaluation model. The results indicate that if the ICC is low, say 0.1, then the program would have to have an impact of at least 3 wpm for this sample to estimate it with acceptable precision, while if the ICC is 0.5 the MDES would be 6 wpm. If covariates are included in the analysis, and they explain 20 percent of the variation in ORF, the associated MDES are about the same as if no covariates are included at each level of ICC; but if covariates explain 50 percent of variation the MDES is a bit lower, between 0.2 and 1 wpm depending on the ICC, than when no covariates are included.

Figure 1: Minimum Detectable Effect Size for Different Levels of ICC and R-squared



These results show that ICC can vary greatly between studies and that in many cases the 0.10 rule of thumb may be too low, resulting in underpowered samples. Researchers planning evaluations should consider the results provided in this brief and, when available, other sources, to inform their decision about a specific ICC, so that samples are large enough to detect the expected program impact.

References

Ardington, Cally, Ursula Hoadley, and Alicia Menendez (2019) "Impact Evaluation of USAID/South Africa Story Powered School Program Endline Report" NORC at the University of Chicago – USAID https://pdf.usaid.gov/pdf_docs/PA00Z3KH.pdf

Keaveney, Erika, Carlos Fierros, Alexander Rigaux and Alicia Menendez (2021). "USAID Reading and Access: Pakistan Reading Project (PRP) Endline Report - Khyber Pakhtunkhwa." https://pdf.usaid.gov/pdf_docs/PA00Z7MW.pdf

Keaveney, Erika, Carlos Fierros, Alexander Rigaux and Alicia Menendez (2021). "USAID Reading and Access: Pakistan Reading Project (PRP) Endline Report - Balochistan." https://pdf.usaid.gov/pdf_docs/PA00Z7MQ.pdf

Kelcey, Ben, Zuchao Shen, and Jessaca Spybrook (2016). "Intraclass Correlation Coefficients for Designing Cluster- Randomized Trials in Sub-Saharan Africa Education". Evaluation Review Vol. 40(6) 500-525.

Menendez, A., R. Nayyar-Stone, I. Rojas, C. Fierros, L. Onyango, and S. Downey (2020) "Uganda Performance and Impact Evaluation for Literacy Achievement and Retention Activity (LARA) Midterm impact and final performance evaluation report" NORC at The University of Chicago. USAID. https://pdf.usaid.gov/pdf_docs/PA00XF5X.pdf

Menendez, Alicia and Gregory Haugan (2020). "USAID Reading And Access Endline Evaluation Report Impact Evaluation Of The National Early Grade Reading Program (NEGRP) In Nepal". https://pdf.usaid.gov/pdf_docs/PA00Z3FH.pdf

Menendez, Alicia, Ursula Hoadley and Anna Solovyeva (2021). "READ Liberia Impact Evaluation Endline Report". https://pdf.usaid.gov/pdf_docs/PA00Z92Q.pdf

Ome, Alejandro and Alicia Menendez (2018). USAID/Ethiopia Impact Evaluation of Reading for Ethiopia's Achievement Developed Community Outreach (READ CO) Program. Endline Evaluation Report. https://pdf.usaid.gov/pdf_docs/PA00Z83M.pdf

Ome, Alejandro and Alicia Menendez (2018). "Impact Evaluation of the USAID/Makhalidwe Athu Project (ZAMBIA). https://pdf.usaid.gov/pdf_docs/PA00SZJS.pdf

Ome, Alejandro, Alicia Menendez and Cally Ardington (2021). "Attrition in Randomized Control Trials: Evidence from Early Education Interventions in Sub-Saharan Africa. NORC Working Paper Series WP-2021.02. https://www.norc.org/PDFs/Working%20Paper%20Series/WPS_OAM_2021.02.pdf

For more information about the impact report, contact:
Alicia Menendez | menendez@uchicago.org
Alejandro Ome | ome-alejandro@norc.org