

October 2021

Using Generalizability and Rasch Measurement Theory to Ensure Rigorous Measurement in an International Development Education Evaluation

Louise Bahry
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_2



Part of the [Applied Statistics Commons](#), [Educational Assessment, Evaluation, and Research Commons](#), [Educational Methods Commons](#), [International and Comparative Education Commons](#), [Social Statistics Commons](#), and the [Statistical Methodology Commons](#)

Recommended Citation

Bahry, Louise, "Using Generalizability and Rasch Measurement Theory to Ensure Rigorous Measurement in an International Development Education Evaluation" (2021). *Doctoral Dissertations*. 2269.
<https://doi.org/10.7275/24148853> https://scholarworks.umass.edu/dissertations_2/2269

This Open Access Dissertation is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

USING GENERALIZABILITY AND RASCH MEASUREMENT THEORY TO ENSURE
RIGOROUS MEASUREMENT IN AN INTERNATIONAL DEVELOPMENT EDUCATION
EVALUATION

A Dissertation Presented

by

Louise Marie Bahry

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2021

College of Education

Research, Educational Measurement,
and Psychometrics

Using Generalizability and Rasch Measurement Theory to Ensure Rigorous Measurement in an
International Development Education Evaluation

A Dissertation Presented

By

LOUISE MARIE BAHRY

Approved as to style and content by:

Jennifer Randall, Chair

Lisa Keller, Member

Ian Barron, Member

Ezekiel Kimball
Associate Dean of Academic Affairs
College of Education

DEDICATION

This one's for me.

ACKNOWLEDGEMENTS

My friends from high school

Married their high school boyfriends

Moved into houses

In the same ZIP codes where their parents live

But I

I could never follow

No I

I could never follow

I hit the highway

In a pink RV with stars on the ceiling

Lived like a gypsy

Six strong hands on the steering wheel

I've been a long time gone now

Maybe someday, someday I'm gonna settle down

But I've always found my way somehow

By takin' the long way

Takin' the long way around

I met the queen of whatever

Drank with the Irish and smoked with the hippies

Moved with the shakers

Wouldn't kiss all the asses that they told me to

No I

I could never follow

It's been two long years now

Since the top of the world came crashing down

And I'm gettin' it back on the road now

But I'm takin' the long way

Takin' the long way around

I'll just take my time, I won't lay down

And take long way around

Well, I fought with a stranger and I met myself

I opened my mouth and I heard myself

It can get pretty lonely when you show yourself

Guess, I could've made it easier on myself

But I

I could never follow

No I

I could never follow

Well, I never seem to do it like anybody else

Maybe someday, someday I'm gonna settle down

If you ever want to find me I can still be found

Takin' the long way

Takin' the long way around

- The Chicks © 2006

I want to thank everyone in my life who helped me to get here, my way. You know who you are,
and I am eternally grateful.

ABSTRACT

USING GENERALIZABILITY AND RASCH MEASUREMENT THEORY TO ENSURE RIGOROUS MEASUREMENT IN AN INTERNATIONAL DEVELOPMENT EDUCATION EVALUATION

SEPTEMBER 2021

LOUISE M. BAHRY,

B.A. (HONOURS), SIMON FRASER UNIVERSITY

M.Ed., UNIVERSITY OF ALBERTA

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Jennifer Randall

Between the United States and Great Britain, over 30 billion USD was spent in 2018 on international aid, over a billion of which is dedicated to education programs alone. Recently, there has been increased attention on the rigorous evaluation of aid-funded programs, moving beyond counting outputs to the measurement of educational impact. The current study uses two methodological approaches (Generalizability (Brennan, 1992, 2001) and Rasch Measurement Theory (Andrich, 1978; Rasch, 1980; Wright & Masters, 1982) to analyze data from math and literacy assessments, and self-report surveys used in an international evaluation of an educational initiative in the Democratic Republic of the Congo. These approaches allow the researcher to identify and select pertinent facets and look at them in relation to one another, allowing us to attribute smaller or larger sources of variability to a particular facet, and using both provides additional insight to instrument development and validation efforts. A thorough analysis of five Early Grades Reading Assessment subtasks, five Early Grades Mathematics Assessment subtasks, and three sets of items from a survey administered to the girls in the study was completed. Results suggest that two factors were consistently flagged as contributing to error in the outcome measures: enumerators and language of administration/girl's home language. The results of this

study provide implications for several phases of evaluations of educational initiatives in developing countries: evaluation design development; the importance of a pilot in assisting in refining the design and sampling plan; and the importance of selecting the appropriate outcome measure, particularly in projects utilizing payment for success models. The results also indicate the utility and complementary nature of using Generalizability and Rasch Measurement Theory analytic procedures in assessing the quality of complex evaluation data. Evaluations such as the one used in this study are highly complex in nature, with more possible sources of error than those included in the current study. What these results indicate is that though there is a wish to standardize and assess in difficult settings, the fact that context affects not only the results of assessments like the EGMA and EGRA, but their utility, cannot be ignored.

TABLE OF CONTENTS

DEDICATION	iv
ACKNOWLEDGEMENTS	v
ABSTRACT.....	viii
LIST OF TABLES	xiv
LIST OF FIGURES	xvii
CHAPTER I. INTRODUCTION	1
Education in the Developing World	1
Poverty 1	
Challenging Geographies.....	2
Quality Education	3
Resources.....	7
Infrastructure.....	8
Conflict, Insecurity, and Instability	9
Gender 11	
Education in the Democratic Republic of the Congo	13
Programs to Support Girls Education in Developing Countries	16
Girls Education Challenge	17
VAS Y Fille! Program	41
VAS-Y Fille! Program Evaluation	49
Issues in International Development Evaluation	56
Summary	58
CHAPTER II. LITERATURE REVIEW.....	61
Educational Interventions in Developing Nations	61
New Schools and Infrastructure.....	63
Materials	64
Teachers.....	66
Management	67
Supply-Side Summary	68
Reduced Educational Costs	68

Increasing Preparedness.....	76
Providing Information.....	81
Summary	82
Challenges Surrounding Translation and Adaptation of Measures.....	83
Validity and Reliability.....	85
Validity Evidence	85
Reliability	86
Using Generalizability Theory to Establish Scale Reliability	89
Using the Rasch Measurement Theory to Establish Scale Reliability	93
Studies Using Both Rasch Measurement Theory and Generalizability Theory to Establish Scale Reliability	96
Summary	97
CHAPTER III. METHODS	99
VAS-Y Fille! Program Evaluation Design	100
School Sampling	102
Student Sampling.....	103
In-School Girls Sampling Protocol: Baseline (2013)	104
In-School Girls Replacement Sampling Protocol.....	105
Instruments.....	105
Early Grade Mathematics Assessment (EGMA)	106
Early Grade Reading Assessment (EGRA)	107
Girl-Friendliness Survey	108
Data Collection	109
Proposed Analysis.....	110
Generalizability Theory	110
Many-Facet Model.....	116
Summary	121
Subjective Measurement.....	121

Girl-Friendliness Survey.....	121
Objective Measurements.....	122
Early Grades Mathematics Assessment	122
Early Grades Reading Assessment	122
CHAPTER IV. RESULTS.....	123
Introduction.....	123
Descriptive Statistics.....	124
Objective Measures.....	124
Subjective Measures	130
Generalizability Theory	138
Determining Datasets for Analysis	138
Analytic Procedure and Interpretation	139
Generalizability Analysis Results	141
Baseline - Objective Measures	141
Early Grades Reading Assessment (EGRA).....	141
Early Grades Mathematics Assessment	153
Baseline – Subjective Measures.....	163
Annual – Subjective Measures	171
Many-Facet Model.....	175
Determining Datasets for Analysis	175
Analytic Procedure and Interpretation	175
Many-Facet Model Results	178
Baseline – Objective Measures	178
Early Grades Reading Assessment (EGRA).....	178
Early Grades Mathematics Assessment (EGMA).....	194
Baseline – Subjective Measures.....	211
Annual – Subjective Measures	221

Longitudinal – Subjective Measures.....	233
CHAPTER V. DISCUSSION, LIMITATIONS AND IMPLICATIONS FOR RESEARCH, AND IMPLICATIONS FOR PRACTICE AND POLICY	244
Summary of the Study	244
Discussion.....	245
Research Question 1: What are the largest sources measurement error in the current evaluation design, and how do they differ for subjective vs. objective measures?	249
Research Question 2: What is the effect the of non-standard translation and adaptation procedures used on the assessments throughout the VAS-Y Fille! evaluation?	251
Research Question 3: What facets are modifiable in a program such as VAS-Y Fille! that would allow for a decrease in the measurement error of the outcome measures?	252
Limitations	252
Implications	255
Implications for Evaluation in International Development.....	255
Implications for Measurement	260
APPENDICES	262
Appendix A. Early Grades Reading Assessment (EGRA)	263
Appendix B. Early Grades Mathematics Assessment (EGMA)	270
Appendix C. In School Girls’ Survey	274
REFERENCES	278

LIST OF TABLES

Table 1. Girls Education Challenge - Phase 1 Projects (UK Aid, 2015).....	19
Table 2. VAS-Y Fille! populations and intended program component exposure (UK Aid, 2013)	44
Table 3. Baseline study quantitative data collection plan (UK Aid, 2013)	48
Table 4. VAS-Y Fille! program evaluation questions (UK Aid, 2013).....	49
Table 5. VAS-Y Fille! project outputs with targets by year (UK Aid, 2013)	51
Table 6. VAS-Y Fille! cohort design (UK Aid, 2013)	54
Table 7. VAS-Y Fille! indicator matrix (UK Aid, 2013)	55
Table 8. VAS Y Fille! sample sizes for in-school girls per group and data collection instance.	104
Table 9. VAS-Y Fille sample composition at Midline and Endline by grade.....	104
Table 10. Reliability for EGRA and EGMA at Baseline	108
Table 11. Sources of variability in a two-facet fully-crossed design	112
Table 12. Sources of variability in a two-facet partially-nested design	113
Table 13. Descriptive Statistics for Baseline EGRA and EGMA Subtasks	124
Table 14. Descriptive Statistics for Baseline and Annual Girls' Survey Results.....	131
Table 15. Sample Size for Baseline EGRA - Letter Name Items.....	142
Table 16. GTheory Results for Baseline EGRA – Letter Name Items.....	142
Table 17. GTheory Results for Baseline EGRA – Non-Word Reading Items	144
Table 18. GTheory Results for Baseline EGRA – Oral Reading Fluency Items	146
Table 19. GTheory Results for Baseline EGRA –Reading Comprehension Items	148
Table 20. GTheory Results for Baseline EGRA –Listening Comprehension Items	150
Table 21. GTheory Results for Baseline EGMA –Number Identification Items	153

Table 22. GTheory Results for Baseline EGMA –Number Discrimination Items.....	155
Table 23. GTheory Results for Baseline EGMA –Missing Number Items.....	157
Table 24. GTheory Results for Baseline EGMA –Addition Items.....	159
Table 25. GTheory Results for Baseline EGMA –Subtraction Items	161
Table 26. Sample Size for Baseline General School Perception Survey Items.....	164
Table 27. GTheory Results for Baseline General School Perception Survey Items	164
Table 28. Sample Size for Baseline Teacher Perception Survey Items.....	166
Table 29. GTheory Results for Baseline Teacher Perception Survey Items	167
Table 30. Sample Size for Baseline School Violence Perception Survey Items	169
Table 31. GTheory Results for Baseline School Violence Perception Survey Items.....	170
Table 32. Sample Size for Annual General School Perception Survey Items.....	171
Table 33. GTheory Results for Annual General School Perception Survey Items	172
Table 34. Sample Size for Annual Teacher Perception Survey Items.....	173
Table 35. GTheory Results for Annual Teacher Perception Survey Items	173
Table 36. Sample Size for Annual School Violence Perception Survey Items	174
Table 37. GTheory Results for Annual School Violence Perception Survey Items.....	174
Table 38. Facets Results for Baseline EGRA – Letter Name Items.....	181
Table 39. Facets Results for Baseline EGRA – Non-Word Reading Items	184
Table 40. Facets Results for Baseline EGRA – Oral Reading Fluency Items.....	187
Table 41. Facets Results for Baseline EGRA – Reading Comprehension Items	190
Table 42. Facets Results for Baseline EGRA – Listening Comprehension Items.....	193
Table 43. Facets Results for Baseline EGMA – Number Identification Items	197
Table 44. Facets Results for Baseline EGMA – Number Discrimination Items	201

Table 45. Facets Results for Baseline EGMA – Missing Number Items	204
Table 46. Facets Results for Baseline EGMA – Addition Items.....	207
Table 47. Facets Results for Baseline EGMA – Subtraction Items.....	210
Table 48. Facets Results for Baseline General School Perception Survey Items.....	213
Table 49. Facets Results for Baseline Teacher Perception Survey Items.....	217
Table 50. Facets Results for Baseline School Violence Perception Survey Items.....	220
Table 51. Facets Results for Annual General School Perception Survey Items.....	224
Table 52. Facets Results for Annual Teacher Perception Survey Items.....	228
Table 53. Facets Results for Annual School Violence Perception Survey Items.....	232
Table 54. Facets Results for Longitudinal General School Perception Survey Items.....	235
Table 55. Facets Results for Longitudinal Teacher Perception Survey Items.....	239
Table 56. Facets Results for Longitudinal School Violence Perception Survey Items	243
Table 57. GTheory Results by Subtask with the Number of Analyses Out of the Total Analyses Completed Including the Facet Accounting for Over 5% of the Total Variance.....	246
Table 58. Facets Results for Single Time Points with the Number of Flags Indicating Variability of the Elements in the Facet on the Logit Scale	248
Table 59. Summary Facets Results for Longitudinal Analyses.....	249
Table 60. General School Perception Items from Girls' Survey.....	274
Table 61. Teacher Perception Items from Girls' Survey	275
Table 62. Perception of School Violence from Girls' Survey	277

LIST OF FIGURES

Figure 1. Framework for Quality Education (Pigozzi, 2006).....	4
Figure 2. School infrastructure comparison between the world and Sub-Saharan Africa (U.N., 2019b).....	9
Figure 3. VAS-Y Fille! Program theory of change (UK Aid, 2013).....	43
Figure 4. Approaches to Educational Interventions in Developing Nations (Krishnaratne et al., 2013)	62
Figure 5. Reliability standards from Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014).....	89
Figure 6. Sources of variability for a two-facet fully crossed design.....	112
Figure 7. Sources of variability for a two-facet partially-nested design.....	114
Figure 8. Simple one-facet Wright map.	118
Figure 9. Wright map for Many-Facet Model.	120
Figure 10. EGMA - Number Identification Subtask Number Correct Distribution	125
Figure 11. EGMA - Number Discrimination Subtask Number Correct Distribution.....	126
Figure 12. EGMA - Missing Number Subtask Number Correct Distribution.....	126
Figure 13. EGMA - Addition Subtask Number Correct Distribution	127
Figure 14. EGMA - Subtraction Subtask Number Correct Distribution	127
Figure 15. EGRA - Letter Name Subtask Number Correct Distribution	128
Figure 16. EGRA - Nonword Reading Subtask Number Correct Distribution	128
Figure 17. EGRA - Oral Reading Fluency Subtask Number Correct Distribution	129
Figure 18. EGRA - Reading Comprehension Subtask Number Correct Distribution.....	129
Figure 19. EGRA - Listening Comprehension Subtask Number Correct Distribution	130
Figure 20. Girls' Survey General School Perception Baseline Total Score Distribution	132

Figure 21. Girls' Survey General School Perception Annual Total Score Distribution	132
Figure 22. Girls' Survey General School Perception Baseline Merged File Total Score Distribution	133
Figure 23. Girls' Survey General School Perception Annual Merged File Total Score Distribution	133
Figure 24. Girls' Survey Teacher Perception Baseline Total Score Distribution	134
Figure 25. Girls' Survey Teacher Perception Annual Total Score Distribution	134
Figure 26. Girls' Survey Teacher Perception Baseline Merged File Total Score Distribution	135
Figure 27. Girls' Survey Teacher Perception Annual Merged File Total Score Distribution	135
Figure 28. Girls' Survey Perception of School Violence Baseline Total Score Distribution	136
Figure 29. Girls' Survey Perception of School Violence Annual Total Score Distribution	136
Figure 30. Girls' Survey Perception of School Violence Baseline Merged File Total Score Distribution	137
Figure 31. Girls' Survey Perception of School Violence Annual Merged File Total Score Distribution	137
Figure 32. Wright map for Many-Facet Model	176
Figure 33. Variable Map for Baseline EGRA - Letter Name Items	180
Figure 34. Variable Map for Baseline EGRA - Non-Word Reading Items.....	183
Figure 35. Variable Map for Baseline EGRA - Oral Reading Fluency Items	186
Figure 36. Variable Map for Baseline EGRA - Reading Comprehension Items.....	189
Figure 37. Variable Map for Baseline EGRA - Listening Comprehension Items.....	192
Figure 38. Variable Map for Baseline EGMA - Number Identification Items.....	196
Figure 39. Variable Map for Baseline EGMA - Number Discrimination	200

Figure 40. Variable Map for Baseline EGMA - Missing Number Items	203
Figure 41. Variable Map for Baseline EGMA - Addition Items	206
Figure 42. Variable Map for Baseline EGRA - Subtraction Items.....	209
Figure 43. Variable Map for Baseline Survey - General School Perception Items.....	212
Figure 44. Variable Map for Baseline Survey - Teacher Perception Items.....	216
Figure 45. Variable Map for Baseline Survey – School Violence Perception Items	219
Figure 46. Variable Map for Annual Survey - General School Perception Items.....	223
Figure 47. Variable Map for Annual Survey - Teacher Perception Items.....	227
Figure 48. Variable Map for Annual Survey - School Violence Perception Items	231
Figure 49. Variable Map for Longitudinal Survey Items - General School Perception Items	234
Figure 50. Variable Map for Longitudinal Survey Items - Teacher Perception Items	238
Figure 51. Variable Map for Longitudinal Survey Items - School Violence Perception Items	242

CHAPTER I. INTRODUCTION

Education in the Developing World

While progress in achieving universal primary education has been made, there continue to be significant and consistent gaps across populations (U.N., 2018). Nine percent of school-aged children remain out of school, with little progress made to decrease this rate since 2009 (U.N., 2017). Despite this enrolment growth, children are not learning, with proficiency rates remaining dismal, particularly in Sub-Saharan Africa where 88% of children in primary and lower secondary school were not proficient in reading, and 84% were not proficient in math as of 2015 (U.N., 2019a). According to Educate a Child, a program of the Education Above All Foundation, there remain eight significant barriers to a child's education in a developing country: poverty; economic migration; challenging geographies; quality education; resources; infrastructure; refugees; gender; and conflict, insecurity, and instability (<https://educateachild.org/>). While the barriers are outlined separately below, it is imperative to note that they interact with one another significantly with individuals in many developing countries being affected by a number of circumstances.

Poverty

Poverty is one of the most reliable predictors of both enrollment and educational success with children from the poorest households four times as likely to be out of school as those in the wealthiest households (U.N., 2015a). There are three primary ways in which poverty can be a barrier to education: direct and indirect costs of education, child labor, and economic migration (<https://educateachild.org/>). According to a Millennium Development Goals report (U.N., 2015a), household wealth is a significant predictor of student attendance and enrollment, with poorer households more likely to have school-aged girls out of school than boys. In recent years, some countries have implemented free primary education but still must pay for uniforms or supplies,

lodging (if the school is too far away), travel, food, etc., in addition to losing out on possible income from the household chores, childminding, farm/business work they would complete.

Child labor may include work for their own family described above, particularly for those families living in poverty. However, children partaking in child labor may be deprived of an education altogether or forced to leave school early. According to the Sustainable Development Goals Report (U.N., 2017), for children not enrolled in school, one consistent reason remains the incidence of child labor. Overall, approximately 10% of children engage in child labor practices, and this incidence is doubled in sub-Saharan Africa with over 20% of children engaged in child labor. In all cases, over half of the children engaged in child labor are engaged in dangerous or hazardous work. As of 2016, 152 million children in the world were engaging in child labor, with nearly half doing hazardous work (I.L.O., 2017). Of the children engaging in child labor across the world, almost half are between 5 and 11 years old (primary school aged), 58% are male, and 70.9% participate in agricultural work.

Of the 258 million international migrants, over half (150.3 million) are classified as migrant workers (I.O.M., 2018). Economic migration is defined as migration which has not been compelled through force or displacement, such as in the case of refugees (Goldin, Pitt, Nabarro, & Boyle, 2018). Though the migration here is a deliberate act, there remains a lack of formal government protection of children who are a part of the populations migrating for economic reasons, resulting in a lack of access to health and education resources, or even child labor (van de Glind & Kou, 2013).

Challenging Geographies

Natural disasters kill 130 for every million people in developing countries compared to 18 for every million people in developed countries, and economic losses as a result of disasters are much higher in developing countries (U.N., 2019a). Challenging geographies may be

physical, demographic, or cultural. Physical challenges include mountains, rivers, volcanic or tectonic zones, deserts, or areas otherwise susceptible to extreme weather or geographic events. Communities built in challenging physical geographies may also be nomadic by necessity with children living in nomadic or semi-permanent locations throughout the year, leading to difficulties in attending a school consistently. Demographic challenges include school overcrowding due to increases in birth rates or migration into the area resulting in strained educational resources. And finally, cultural challenges include cultural, language of instruction, religious, or political differences resulting in difficulties in enrolling and attending schools.

Quality Education

One of the major barriers to education in the developing world rests in the widespread unavailability of quality education (<https://educateachild.org/explore/barriers-to-education/quality>; <https://www.globalcitizen.org/en/content/10-barriers-to-education-around-the-world-2/>; (Force, 2013). A 2014 meta-analysis of educational interventions in sub-Saharan Africa showed the highest effect sizes in those studies with pedagogical interventions, emphasizing the importance of quality teaching in educational success (Conn, 2017). UNESCO released a conceptual model for quality education looking toward the future of education globally (Pigozzi, 2006). The model, pictured in Figure 1 shows the necessary components of a quality education system required at the levels of the learner, and at the levels of the system, in order for learning to occur. A recent article noted the 10 greatest barriers to education (<https://www.globalcitizen.org/en/content/10-barriers-to-education-around-the-world-2/>), all of which align to aspects of the quality education Pigozzi outlines: lack of funding; no, or untrained teachers; no classrooms; lack of learning materials; exclusion of students with disabilities; being the ‘wrong’ gender; living in a country in or at risk of conflict; distance to school from home; hunger and poor nutrition; and, the expense of education for the individual.

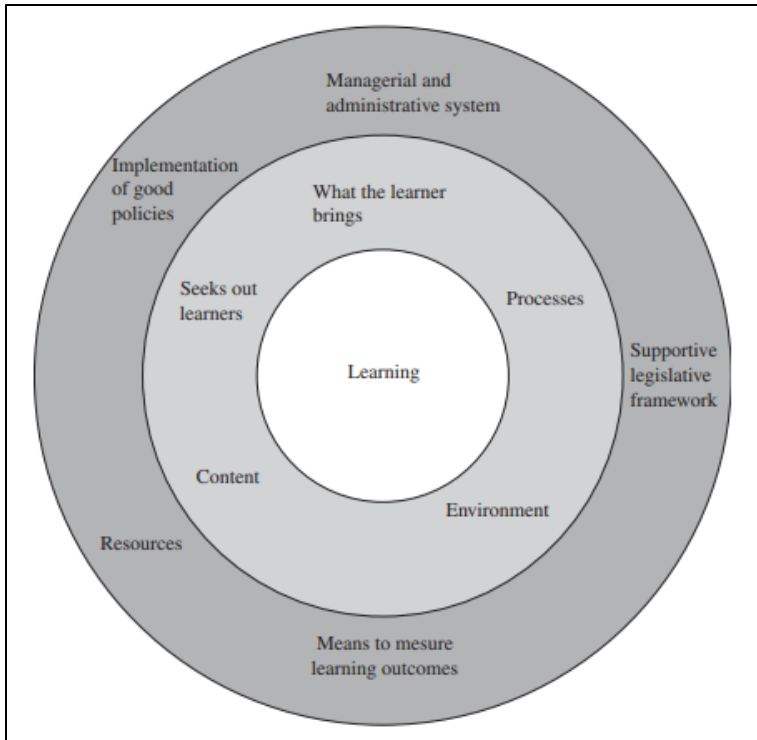


Figure 1. Framework for Quality Education (Pigozzi, 2006).

As shown in the model, at the level of the learner (the inner circle), there are 5 components:

- 1) Seeking out learners – underscores the need for education to be available without discrimination, and the general right to education;
- 2) What the learner brings – brings focus to the context from which each learner comes, be it traumatic experiences due to conflict or strong early childhood education, and requires an educational system to recognize this diversity;
- 3) Content – recognizes the need for evolving educational needs and content above and beyond the traditional reading, writing, and mathematics focus that has been pervasive. A quality education has content that is inclusive and materials that are accessible;
- 4) Processes – educational processes include the ways in which learning is facilitated by the teacher to ensure a learner-centered approach that minimizes issues of inequality.

The use of strong educational processes in the classroom requires well-trained teachers who are able to adapt their approach dependent on their learner context;

- 5) Learning environment – evidence has been gathered to suggest that a more holistic approach to a learning environment is needed. This approach includes physical classroom space, access to hygiene facilities and health and nutrition services, and general safety and security of students and teachers.

At the level of the system or organization (the outer circle), the 5 indicated components are as follows:

1. Managerial and administrative structure and processes – the structure of an education system provides an accountability framework requiring all key stakeholders to play an active role in enabling the system to function. This includes communities, parents, students, education departments or ministries, and teacher training institutions;
2. Implementation of good policies – as education exists within and is dependent on and for other systems and structures within a society, the policies and procedures should reflect these inter-dependencies, be consistent with other governmental policies, and be reflective of current state to remain relevant and understood;
3. Supportive legislative framework – a robust legislative framework will address access and quality of education, resource allocation, and overall expectations and accountability of the system in place;
4. Resources – high quality education requires a range of human and material resources as noted above and in the next section, which must be viewed as a long-term investment in the future state;
5. Means to measure learning outcomes – as the general view of what constitutes a quality education system has expanded greatly, the need for support in appropriate assessment of these more complex learning outcomes (i.e., values) increases.

Building upon the work done by Pigozzi, the Brookings Institute and UNESCO commissioned a task force to ensure that quality learning across the globe remained a priority post-2015 and the close of the Millennium Development Goals. The group recognized that while progress had been made in terms of universal access, as evidenced by increased enrollments, results in learning outcomes remained inconsistent and incomparable internationally. Therefore, a report was released (Force, 2013) with seven recommendations to move the world toward quality universal learning:

- 1) a global paradigm shift in focus from universal access to access and learning;
- 2) the development of learning competencies across seven domains of learning: physical well-being, social and emotional learning, culture and the arts, literacy and communication, learning approaches and cognition, numeracy and mathematics, and science and technology;
- 3) the development of learning indicators for global tracking including seven areas of measurement: learning for all, age and education matter for learning, reading, numeracy, ready to learn, citizen of the world, and breadth of learning opportunities;
- 4) supporting countries in strengthening their assessment systems by improving the technical, institutional, and political capacities;
- 5) measurement of learning with an explicit focus on identifying and addressing inequity;
- 6) championing assessment as a public good by making tools, documentations, and data publicly available; and,
- 7) encouraging all stakeholders to take actions to ensure the right to learn.

Resources

There are three types of resources that are supportive of providing quality education: human, material, and financial resources (<https://educateachild.org/explore/barriers-to-education/resources>). Human resources include a wide breadth of individuals from educational developers, administrators, teachers, mentors, and support staff. As noted in the previous section, research has shown that interventions including pedagogical changes show the greatest improvements in learning (Conn, 2014). Educate a Child notes three domains of support required for quality teaching: emotional support including positive connections between teachers and students and teacher sensitivity to student needs, organizational support including classroom and behavior management, and instructional support including appropriate learner strategies and quality feedback. Poor quality teaching leads to poor learning outcomes for students, and in developing countries where family financial resources for education are limited, students who do not show positive learning outcomes are more likely to be removed from school in order to save money.

As with human resources, quality education materials are not widely available in developing countries. There are many reasons these materials may not be developed or distributed throughout an education system. If a ministry of education does not have available funds or expertise to develop or re-develop education materials for their system, teachers and students will be required to make do with low-quality or out of date learning resources, or to rely on outside agencies to provide said materials. For those education systems where quality materials may be developed, distribution may also be a problem due to inadequate infrastructure for delivery of said resources.

Finally, financial resources for education may also be a barrier in some developing countries, whether at the family, community, or country level. As has been previously noted,

poverty is a major barrier to education, with the direct and indirect costs of education, prevalence and necessity of income through child labor, and families migrating for economic reasons and keeping children from attending school regularly. In addition to individual family financial resources, there may be systemic funding issues at the government level, leaving education funding for the nation lacking. A 2019 report by the United Nations noted that the poorest countries in the world would need to at least triple their education funding in order to meet funding requirements for universal primary education (U.N., 2019a), leaving a massive funding gap to be filled by external financial support.

Infrastructure

Directly related to the resource barrier, there are particular infrastructure needs for a sustainable and quality education system. Some of the more prevalent inadequacies seen in developing countries around infrastructure include: insufficient space per child and the adherence to reasonable teacher/student ratios; inadequate sanitary facilities for students and staff, including separate facilities for boys and girls; safe methods for students and staff to travel to and from the school; and, safe school sites. In addition, as education methods change and the overall connectedness of the world continues to grow, secure and stable internet connectivity is fast becoming a requirement for quality education delivery (<https://educateachild.org/explore/barriers-to-education/infrastructure>). Figure 2 shows a comparison between basic infrastructure availability in schools such as drinking water and handwashing facilities across the world vs. in Sub-Saharan Africa, the region most lagging in school infrastructure (U.N., 2019b). As has been noted, this region is also lagging most in school enrollment and achievement.

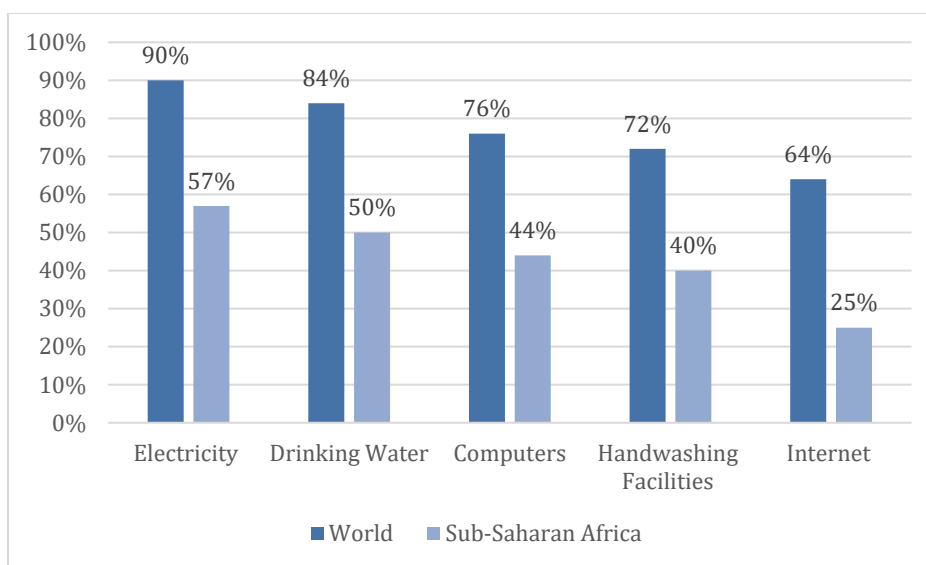


Figure 2. School infrastructure comparison between the world and Sub-Saharan Africa (U.N., 2019b).

Conflict, Insecurity, and Instability

Conflict-affected countries are home to more than a third of out of school children (UNESCO, 2014). Children in these countries are 30 percent less likely to complete primary school, and 50 percent less likely to complete lower secondary school, and they tend to have higher dropout rates, gender disparities, and out of school rates and lower completion rates and literacy levels (UNESCO, 2011). In some cases, conflict can even erase past educational gains. For example, the Syrian Arab Republic had achieved universal primary enrollment in 2000, but by 2013, 1.8 million children were out of school due to conflict (UNESCO, 2015).

[Educate a Child](#) indicates three main barriers to education related to conflict and instability: internally displaced persons, active armed conflict, and the use of child soldiers. In recent years, the challenge of educating children in conflict, post-conflict, and refugee contexts has risen exponentially. In fact, education is often directly a target of conflict and instability. The Global Coalition to Protect Education from Attack (GCPEA) reported in 2018 that between 2013 and 2017 there were attacks on education in 74 different countries. Eight countries were

identified as very heavily affected by attacks on education and military use of schools and universities between 2013 and 2017 (GCPEA, 2018). These are countries where reports documented 1,000 or more incidents of attacks on education or military use of educational facilities or 1,000 or more students and education personnel harmed by attacks on education, including the DRC, South Sudan, and Syria.

In 2018, the UNHCR reported that the number of people displaced by war, persecution, and conflict had exceeded 70 million, the highest level seen in almost 70 years (U.N., 2019b). This number includes internally displaced persons (IDPs), refugees, and asylum-seekers. Columbia had the largest displacement situation with over 9 million persons displaced, and Syria remained at the top of the list as the second largest displacement situation with 6.6 million Syrians displaced in some way due to the ongoing conflict. The DRC was the third-largest displacement situation with over 5 million Congolese displaced, comprising 4.4 million IDPs, 620,800 refugees and 136,400 asylum-seekers (UNHCR, 2019).

The movement of families and children from their home communities due to conflict and crisis places the burden of education onto other communities or countries in turn. Depending on the capacity and supports in the receiving community, the influx of children into a new education system stresses school infrastructure, resources, and even the most basic physical or emotional safety of the students (Moul, 2017). Key challenges of education in crisis and conflict contexts have been identified, and the United Nations continues to work to implement policy to ensure that children's education needs are being met.

Active armed conflict is most common in some of the world's poorest countries, and this has disastrous effects on education, both by way of infrastructure and resources. Schools can often be targeted during active conflict, causing damage or complete destruction of schools and infrastructure supporting schools such as roads and electricity (GCPEA, 2018). Active conflict

often results in school closures and absenteeism by children and teachers alike, and even when schools do remain open, the conflict can threaten their safety on the way to and from school. In a post-conflict area, there is often a resulting lack of qualified teachers, and inadequate policies and infrastructure in place for recruitment, training, and funding for teachers (including a lack of available funding for wages, or consistent and on-time delivery of wages; Bretherton, Weston, & Zbar, 2003).

Despite the 2008 ratification of the Child Soldiers Prevention Act (CSPA), 11 countries remain publicly listed as continuing to recruit, abduct, and use child soldiers in conflict. These countries include Afghanistan, DRC, Iran, Iraq, Syria, and Yemen. Child Soldiers are forced to be combatants, porters, servants, messengers, and spies for government armed forces, paramilitary organizations, and rebel groups. Girls may be forced to marry men in these organizations, may be subject to rape, sexual abuse, or other exploitation (U.S. Department of State, 2019).

Gender

While all of the barriers discussed above affect millions of boys and girls across the world, there is one barrier that disproportionately affects girls, their gender. Sixty-two million girls worldwide continue to be denied the right to attend primary and lower secondary school (UNESCO, 2015), with 118 girls out of school worldwide compared to 100 boys (U.N., 2019b). More specifically, for every 100 boys of primary school age out of school in 2017, 127 girls were denied the right to education in Central Asia, 121 in sub-Saharan Africa, and 112 in Northern Africa and Western Asia. The majority of girls in the world not enrolled in school live in Africa and Asia. In Sub-Saharan Africa, 16.6 million school-age girls are not enrolled in primary school, and 11.3 million are not enrolled in secondary school. In Asia, 8.1 million girls are not enrolled in primary school, and 16.3 million are not enrolled in secondary school (Clinton Foundation, Bill & Melinda Gates Foundation, & WORLD Policy Analysis Center, 2015). And, in Arab states, 2.6

million school-age girls are not enrolled in primary school and, 1.7 million are not enrolled in secondary school (UNESCO, 2015).

Girls are also more likely than boys never to enroll in school (48% vs. 37%), but once enrolled, they are more likely to reach the upper grades (UNESCO, 2015). For those girls that do enroll in school, making progress and completing even primary school remains difficult. For example, three-quarters of girls are enrolled in school in Sub-Saharan Africa, but only 8 percent finish (Winthrop, Anderson, & Cruzalegui, 2015). Girls who do attempt to attend school may face violence, and even death, in countries ranging from Nigeria to Pakistan, and many face sexual abuse on their way to or from school, or even when they are at school (GCPEA, 2014). Poor, rural girls are facing a triple disadvantage with women in rural areas in both low- and lower-middle-income countries spending less than three years in school (Clinton Foundation et al., 2015).

Child marriage and adolescent childbirth are major contributors to girls leaving school. In 2015, in sub-Saharan Africa, 36.6% of girls under the age of 18 were married, with 11.3% married under the age of 15, and in Central and Southern Asia, 43.2% were married under the age of 18 with 15.7% married under the age of 15 (U.N., 2017). According to data collected in 2017, 21% of women worldwide reported being married prior to their 18th birthday (U.N., 2018). However, data has also shown that for every additional year of secondary schooling, a girl is 7% less likely to enter into a child marriage (Wodon, Male, Onagoruwa, & Yedan, 2017). In addition to child marriage, girls are doing a disproportionate amount of unpaid labor, care, and domestic work, at times to the extent of 3 times as much as their male siblings (U.N., 2018).

In particular, developing sub-Saharan African countries continue to fall below the rest of the world when it comes to enrollment in early childhood education and primary education, as well as in secondary school graduation rates. And, while girls tend to outperform boys on

measures of reading proficiency, out-of-school rates for girls are still higher than for boys (U.N., 2017). As recently as 2012, in North Africa, sub-Saharan Africa, and Western Asia only 64% of developing countries had reached gender parity in primary education, with over half of those countries located in sub-Saharan Africa (U.N., 2015a). Despite these challenges, girls who do attend school are more likely than boys to learn how to read. Globally, for every 100 boys who achieved minimum proficiency in reading in 2015, 105 girls of primary school age and 109 adolescent women of lower secondary school age met at least the minimum standard (U.N., 2019b).

Education in the Democratic Republic of the Congo

The Democratic Republic of the Congo (DRC), is the second largest country in Africa, and ranks 175 of 188 on the Human Development Index (<http://hdr.undp.org/en/content/human-development-index-hdi>), making it one of the poorest countries in the world. While considered a post-conflict country, the DRC continues to experience unrest, continues to suffer from weak infrastructure and governance practices, and responded to its 11th outbreak of the Ebola virus since 1976 in the summer and fall of 2020 (World Health Organization, 2019). As a result, it is no surprise that there continue to be barriers to children achieving quality primary and secondary education. The main barriers to education in the DRC are: poverty, conflict and displacement, low levels of maternal education and perpetuation of gender inequality, early marriage pregnancy, and sexual violence, and the access to quality education (D. K. Evans & Popova, 2016; Groleau, 2017; UNESCO, 2014; UNICEF, 2016; USAID & ECCN, 2016). It should be noted, that each of these barriers tends to disproportionately disadvantage girls than boys. For example, when finances are tight, families keep their daughters home, as they can be helpful in the home taking care of younger siblings, older or ill family members, and aiding with cooking and cleaning.

The DRC is one of the poorest nations in the world, with a World Bank estimate of 63.9% of the population living below poverty in 2012 (<https://data.worldbank.org/country/congo-dem-rep>), more recent estimates place this at 71% of the population (Berthet, 2013). Limited financial resources leave families unable to afford school fees, supplies, and proper uniforms, and even the Ministry of Education operates with a budget of only 11% of the total government budget while serving 13.1 million primary age students (De Herdt & Titeca, 2016). Though primary education (grades 1 through 6) was made free and compulsory in 2010, large numbers of students remain out of school, with higher proportions of girls and rural children out of school. While free, there are costs that remain on the shoulders of families such as supplies and uniforms, as well as the possible lost income from the child that would be helping to support their family were they not in school, and the International Rescue Committee (IRC) estimates that 48% of girls aged 5 to 14 are engaged in some form of child labor.

In January of 2018, the United Nations appealed for 1.68 billion USD for urgent assistance of displaced people in the DRC. With a total of 4.3 million internally displaced people, and an increase of 1.7 million in 2017, the DRC has the largest number of displaced persons in Africa (U.N., 2018). Reasons for displacement are highly variable, but in any case, the poor infrastructure resulting from severe poverty and underfunding of education does not allow for sudden changes in student populations due to displacement.

The particular barrier of gender has been discussed previously, and this is no more prevalent than in the DRC where approximately 40% of girls who enroll in primary school do not complete it, as compared to only 20% of boys who fail to complete. Data shows gender discrimination is particularly strong in families where the mother has no education, they reside in a rural setting, and household income is extremely low. In these households, when faced with a decision to send male or female children to school, families will choose to enroll their sons first (Pereznieta, Magee, & Fyles, 2017).

Child marriage rates in the DRC are also high, possibly contributing to low secondary school completion rates. Nearly half of the women in the DRC marry before the age of 18, and 28% of married adolescent girls are either pregnant or have given birth already (Groleau, 2017). Girls who enter into marriage as children are likely to be victims of partner violence at a rate of 1 in 5 girls (U.N., 2018). Incidence of sexual violence is high in the DRC and has been studied in depth for years. However, the main focus of much of the research has been in relation to single-incidence sexual assault during the course of conflict. However, recent attention has been paid to the large incidence of school-based sexual violence perpetrated upon girls both within the schools by teachers and fellow students, and on the way to school or home from school.

As noted, the Ministry of Education in the DRC invests only 11% of the national budget into education. In 2013, this equated to only 2.5 percent of their GDP, which is less than half of the Sub-Saharan Africa average (Wodon et al., 2017). This low level of investment does have a negative effect on the country's ability to recruit, train, and retain qualified teachers. As a result, teachers are often un- or under-trained both in content and pedagogy, and with average wages for teachers in country being so low (between \$100 and \$150 USD per month), teachers have more incentives to leave the profession than stay (Groleau, 2017). Student achievement also lags in the DRC with two-thirds of students in 3rd and 4th grade being unable to read a single word in a sentence. Access to education is also problematic with 36% of girls aged 7 to 16 in poor, rural areas having no access to school (Groleau, 2017).

In order to address these barriers in a systematic way, in 2016, the Ministry of Education released the education sector strategy (2016-2025), which presents a framework for education reform in the country structured around three main strategic outcomes: 1) develop access and ensure equity, particularly around the expansion of the free and compulsory primary education policy implements in 2010; 2) improve the quality of learning by implementing quality assurance

and monitoring practices; and 3) improve governance and oversight of the education system (Groleau, 2017).

Programs to Support Girls Education in Developing Countries

Despite the gender gap ever present in education in developing countries, there are many documented benefits of educating adolescent girls in the developing world including, but not limited to, later marriage; lower fertility; healthier, more educated children; and even more rapid economic growth (Chaaban & Cunningham, 2011; Herz & Sperling, 2004; Rihani, Kays, & Psaki, 2006; Summers, 1992). For example, in the report *Girls Count*, from the Center for Global Development (Levine, Lloyd, Greene, & Grown, 2008), the authors outline a case for investing in girls and outline actions for governments of developing countries, the donor community, private employers, and civil society to follow. Their broad agenda includes three actions for these groups to consider moving forward: count girls to make them more visible to policymakers; invest in girls in a strategic and significant way; and, give girls a fair share across employment, education, and social programs.

In the sequel to *Girls Count*, *New Lessons: The Power of Educating Young Girls* (Lloyd & Young, 2009), the authors further explore the unique challenges and opportunities of educating adolescent girls in developing countries. The report outlines three distinct developmental phases of adolescent girls, and provides learning goals and pathways for each, as well as expanding upon the broad agenda proposed in *Girls Count* (Levine et al., 2008) to propose ten actions from which to move forward: collect and compile data on non-formal education, build and maintain a global database for education programs for adolescent girls, expand opportunities for girls to attend secondary school, support the non-formal education system, develop after-school tutoring and mentoring programs in both primary and secondary schools, produce curricula relevant to adolescent girls, offer post-secondary vocational programs, provide training and ongoing

incentives for women to enter and remain teaching, promote easy transition between non-formal and formal schools, and encourage and evaluate innovation. New Lessons puts forth a transformational view of educating adolescent girls, identifies actions to take, and encourages innovation in the education content and process, as well as evaluation of programs new and old.

More recently, the Brookings Institute compiled evidence on what works in educating girls (Sperling, Winthrop, & Kwauk, 2016). The book reviews the progress made by way of the MDGs, and outlines a path forward in focusing efforts and addressing five challenges in girls' education remaining as we move toward meeting the SDGs:

- to achieve actual learning and a high-quality education;
- to enable girls to complete secondary education and to address the learning needs of out-of-school adolescent girls;
- to help girls overcome violence and conflict;
- to assist girls in making school-to-work transitions; and
- to empower girls and women through education.

Girls Education Challenge

In response to added challenges of educating girls in developing countries, the Girl's Education Challenge (GEC) was implemented by the United Kingdom Department for International Development (DFID) in 2012. The GEC was implemented with a budget of nearly 350 million GBP and the intention of helping up to a million of the world's most marginalized girls, via 37 projects in 18 countries, improve their lives through education. The GEC focuses on supporting projects that plan to use innovative and effective ways of achieving this goal (<https://www.gov.uk/guidance/girls-education-challenge>) through three different funding mechanisms:

1. Step Change: scaling up successful interventions that are already having a positive impact;
2. Innovation: applying new interventions such as technological innovations, developing new partnerships, adapting proven solutions for new geographies, communities or age groups; and,
3. Strategic Partnerships: creating new partnerships with the private sector including, Discovery Communications, The Coca Cola Company, Avanti Communications and Ericsson.

Table 1 shows all 37 projects funded under this phase of the GEC spanning from projects focused on enrollment and attendance increases, teacher training, and curriculum focused on girl friendliness.

Table 1. Girls Education Challenge - Phase 1 Projects (UK Aid, 2015)

Country	Project	Education Focus	Project Description
Afghanistan	Establishing Community Based Girls Schools	Primary and lower secondary	<p>Establishing Community Based Girls Schools (CBGSs) across 10 provinces, enrolling out-of-school girls in each target community.</p> <p>Training government teachers from selected government schools in effective teaching methods and in the subjects they will teach.</p> <p>Training mentors from selected government schools to provide weekly support to their peers</p> <p>Mobilizing school management committees and communities to select girls from target schools to receive stipends.</p> <p>Increasing the capacity of communities, parents, local partners and local education departments to support girls' education in each target community across 10 provinces.</p>
	Steps Towards Afghan Girls' Education Stages	Pre-primary, primary, secondary and teacher training	<p>Establishing and supporting positive/conducive quality learning environments.</p> <p>Increasing demand for and engagement in quality education within communities, particularly for girls.</p> <p>Increasing literacy and engagement with learning among adults and communities.</p> <p>Increasing the capacity of teachers to apply effective, gender fair and relevant teaching methodologies.</p> <p>Strengthening relationships and capacity among national, provincial and district level education actors to sustain girls' education.</p>

Country	Project	Education Focus	Project Description
Burma	Empowering Marginalised Girls in Afghanistan	Primary	<p>Increasing access to primary school education for 2,400 marginalised girls in target districts</p> <p>Improving literacy for 12,240 marginalised girls in target districts through the provision of nine-month basic literary courses.</p> <p>Increasing income generation capacity for 720 marginalised girls in Faryab through the provision of six-month vocational training courses.</p>
	Equal Access to Education for Nomadic Populations in Northern Afghanistan	Lower Primary	<p>Offering summer and winter tuition to 1,200 girls and 800 boys regardless of their current enrolment status in winter government schools, whilst prioritising children who never enrolled or dropped out.</p> <p>Recruiting and training interested and literate women and men from migrating households to teach students in Grades 1 to 3 and Adult Learning Programmes. These teachers are supported by the NGO and Ministry of Education staff.</p> <p>Working with parents, in particular Education Shura members, to provide training and support and mobilise resources for their children's education. A mobile learning (M-learning) literacy program using Ustad mobiles will be used to increase parental involvement in children's education in school and home environments.</p> <p>Recruiting mentors from migrating households to work with parents and children offering additional and after-school learning opportunities for children.</p>
	Mobile Broadband and Education	Secondary	<p>Providing mobile broadband connectivity and ICT equipment (laptops and tablets) to up to 50 secondary schools in Burma (also known as Myanmar).</p> <p>Providing comprehensive teacher training, focusing on ICT skills for improved student learning (including the development of computer skills, pedagogical skills on student-centred and activity-based teaching, and support for teachers to</p>

Country	Project	Education Focus	Project Description
Democratic Republic of the Congo			develop their own teaching and learning materials adapted from existing online resources).
			Delivering an English language programme and a life skills programme to girls in secondary school, using mobile technology, to prepare girls with practical skills for the workplace and to build self-confidence.
			Providing up to 600 secondary school scholarships for marginalised girls.
	Valorisation de la Scholarisation de la Fille	Primary and lower secondary	<p>Increasing parental financial capacity to support girls in primary education.</p> <p>Improving girls' reading and maths skills through teacher training, tutoring and community reading programmes.</p> <p>Increasing community involvement, ensuring girls' access to quality education in a safe environment.</p> <p>Increasing civil society engagement in providing alternative learning opportunities that will allow out-of-school girls to complete primary education.</p>
	Securing Access and Retention into Good Quality Transformative Education	Primary and lower secondary	<p>Increasing the value attached to education by families of targeted marginalised girls (fathers and brothers especially) and their ability to develop more secure livelihoods to protect and support their daughters' education.</p> <p>Removing economic and psychosocial barriers that prevent girls being marginalised by early marriage, domestic labour, risky child migration and/or street-living from entering and remaining in primary school</p> <p>Supporting enrolled, marginalised girls to learn useful knowledge and skills</p> <p>Supporting the creation of a stimulating, safe, inclusive and child-friendly school learning environments for all girls and boys (including those with disability) in the 30 prioritised Kebeles.</p>

Country	Project	Education Focus	Project Description
			Supporting teachers, school administrators, parent groups, community leaders, community-based organisations and child protection structures to develop the skills and mechanisms that will assure and sustain access to good quality education for the targeted marginalised girls.
	Pastoralist Afar Girls' Education Support Projects	Primary	<p>Strengthening the provision of quality and gender-responsive alternative basic education and formal primary education services for girls in pastoralist communities.</p> <p>Improving physical infrastructure of schools including classrooms and access to water</p> <p>Improving life skills, literacy and confidence levels of marginalised girls and creating supportive community environments.</p> <p>Improving basic service delivery, coordination and livelihood opportunities, in order to minimise demand-side barriers to quality education for girls</p> <p>Strengthening government capacity to sustain and scale up project outcomes through strategic partnerships.</p>
	Life Skills and Literacy for Improved Girls Learning in Rural Wolaita Zone	Lower and upper primary	<p>Supporting the development and implementation of Gender Action Plans at woreda, cluster and school level, based on annual Gender Audits (aligned to the Ministry of Education National Girls Education Strategy)</p> <p>Developing local language audio-visual resources and supplementary readers for use in schools. Supporting extra-curricular clubs (Female Learners Forums, Girls Clubs and Reading Clubs); providing sanitary pads and upgrading sanitation facilities; support HIV/AIDS Circles; providing tutorial classes for girls 'at risk' of failure or dropping out</p> <p>Training teachers, school directors, PTA members, Girls Education Advisory Committee (GEAC) members and School Improvement Committee members in</p>

Country	Project	Education Focus	Project Description
Ghana			gender mainstreaming. Training teachers in gender-friendly methodology. LCDE also provides teachers with specific training in basic reading and numeracy. School development plans are expected to include activities around strengthening basic reading skills.
			Developing a school management simulation game to explore the challenges and benefits of girls' education. Lessons from the project will be shared through study tours, girls' education newsletters, zone girls' education conferences and regional and federal dissemination events.
			Building parents support for girls education by helping GEAC to carry out community awareness campaigns including the use of audio-visual resources.
	Making Ghana Girls Great!	Primary school distance learning	Introducing an interactive, distance-learning project to schools across the two districts. This is designed to respond to the scarcity of teachers in these districts.
			Equipping two classrooms in each school with a computer, projector, satellite modem and solar panels to provide reliable power (six hours a day, five days a week). A studio in Accra will be used as an interactive, distance-learning platform, to deliver both formal in school teaching and informal after school training.
			Addressing demand-side barriers to girls' education through an additional set of activities undertaken for two hours per day after school. The activities will follow a programme of lectures, readings, group activities and discussions covering girls' rights, sexual harassment, menstruation, malaria prevention, health, family planning and careers.
			Providing residential training for government teachers, recognising the important role played by facilitators in the classrooms.

Country	Project	Education Focus	Project Description
Ghana, Kenya, and Nigeria	Discovery Project	Primary and Junior Secondary	<p>Improving the quality of education through media in the classroom and teacher professional development (reaching over 11,900 teachers and 528,000 students in 1000 schools across Kenya, Ghana and Nigeria)</p> <p>Training and supporting communities in developing and implementing action plans to address barriers to education and gender marginalisation, including formation of clubs and other activities to connect out-of-school girls with educational opportunities and support girls to succeed in school</p> <p>Producing national television programmes that aim to change knowledge, attitudes and practices around education, especially for girls and women (reaching over 10 million people through locally-produced national broadcasts in Ghana, Kenya and Nigeria)</p>
	Kenya Equity in Education Project	Primary and lower secondary	<p>Working with refugee communities to improve girl-friendly school environments by guaranteeing there are separate latrines for girls to ensure privacy and safety.</p> <p>Providing girls with items they are lacking that will enable them to stay in school and improve learning, such as uniforms, stationary, solar lamps, and sanitary wear.</p> <p>Targeting support for female learners by providing remedial academic training and secondary school scholarships.</p> <p>Building parent and community support for girls' education by adopting multi-media (SMS, films, radio) strategies to share information and generate discussion on girls' education</p>
	Kenya Wasichana Wote Wasome	Primary	<p>Sustaining the capacity of communities to support the education of marginalised girls</p> <p>Sustaining the capacity of households to support their daughters' education.</p>

Country	Project	Education Focus	Project Description
			<p>Developing schools' capacity to provide a safe and supportive environment for girls' learning.</p> <p>Improving girls' health, self-confidence and aspirations to learn.</p> <p>Increasing the ability of the Ministry of Education to support education for marginalised girls.</p>
	Empowering Pioneering Inclusive Education Strategies for Disabled Girls in Kenya	Lower and Upper primary	<p>Addressing discrimination and stigma of disabled girls' education.</p> <p>Building the awareness and capacity of service providers on the rights and potential of disabled children by training education officials, politicians, media representatives and representatives of local civil society and faith based organisations; establishing parent groups linked to each school to engage parents/carers in quarterly meetings and training on practical care; and establishing child-to-child clubs to encourage children with and without disabilities to mix.</p> <p>Improving disabled girls' access to formal education by making schools physically accessible and training teachers in inclusive strategies and Kenyan sign language.</p> <p>Partnering with the LCD research centre at University College London to yield lessons on the barriers for disabled girls in transition from primary to secondary education.</p>
	Improved School Attendance and Learning for Vulnerable Kenyan Girls through an	Upper & lower primary and secondary	<p>Improving the capacity of school management committees to raise funds and form corporate partnerships.</p> <p>Strengthening the role of families and communities to encourage girls to pursue an education in 60 communities.</p>

Country	Project	Education Focus	Project Description
	Integrated Intervention		<p>Strengthening schools to improve the quality of education through providing training to teachers in curriculum delivery and gender; training management committees in gender policies and training teachers in data analysis.</p> <p>Increasing resources to improve the physical infrastructure of schools to ensure girls attend, stay in school and learn.</p> <p>Implementing Ministry of Education pro-gender policies to improve the quality of education. These include: School Management Committees, the Back-to-School Policy (for young mothers) and the Sanitary Towel Provision policy.</p> <p>Improving the positive portrayal of women to ensure girls stay in school and learn by training secondary school students as life skills peer educators and establishing mentoring clubs.</p> <p>Hosting large motivational mentoring events for girls and their mothers.</p> <p>Tracking student and teacher attendance, student performance and other metrics through a biometric system which will be rolled out to all 60 intervention schools.</p>
	The iMlango Project	Primary	<p>The iMlango project (derived from the Swahili word, ‘mlango’ which means doorway or portal) aims to deliver improved educational outcomes in maths, literacy skills and life skills for marginalised girls.</p> <p>The project combines: high speed internet connectivity to schools; provision of tailored online educational content; electronic attendance monitoring with a conditional payment to families to improve non-attendance and drop-out rates at school; in-field capacity in technology and support resources; and real-time project monitoring/measurement.</p> <p>The high-impact education programme aims to improve learning outcomes 25,675 marginalised girls across 195 Kenyan primary schools.</p>

Country	Project	Education Focus	Project Description
Malawi	Empowering Young Female Teachers to Create Inclusive Learning Environments for Marginalised Girls	Upper Primary	The end-to-end solution is made possible by a unique combination of satellite broadband and e-commerce technology, supported by interactive educational and IT resources.
			At the heart of the iMlango projectsits a dynamic internet learning platform, accessed through satellite connectivity, where partners provide students with interactive educational content.
			Identifying 350 outstanding young female teachers and training them as Agents of Change, capable of identifying and supporting girls who are at risk of dropping out, or who have left school, to improve their sexual and reproductive health, self-confidence and literacy and numeracy.
Mozambique	Promoting Advancement of Girls' Education in Mozambique	Primary and lower secondary	Posting 315 Agents of Change teachers to 225 rural and peri-urban primary schools. With the support of the Head Teachers and School Management Committees these Agents of Change will lead a range of extra-curricular activities including Girls' Clubs and Radio Listening Clubs. These activities are specifically designed to support the learning and sexual and reproductive health knowledge, attitudes and skills among girls aged 11 to 15. In addition the Agent of Change teachers will work in their local communities, reaching girls who have dropped out of school, and empowering parents to be more deeply involved in their children's education.
			Reducing economic barriers to girls' participation in primary and lower secondary education through a programme of social transfers for education, including Education Kits and Secondary Bursaries for marginalised children.
			Reducing socio-cultural barriers to girls' education through community mobilisation campaigns and community radio programmes.
			Implementing Girls' Clubs in schools supporting girls safety, development, participation and self-esteem.

Country	Project	Education Focus	Project Description
Nepal			Training school councils in gender issues and providing funds for school improvement.
			Improving access of marginalised girls to enhanced teaching methodologies for reading, leading to improved learning outcomes.
			Offering additional learning opportunities to marginalised girls during crucial transition years through 'Transition Classes'.
			Building the capacity and commitment of government and other education stakeholders to embed PAGE-M methodologies in the education system.
	The Business of Girls' Education	Upper primary	<p>Training marginalised girls and boys on peer education, life and vocational skills. These girls will become 'Lead Girls', promoting self-empowerment and dialogue about home and school environments. Training teachers and School Council members on gender responsive curriculums and methods. Teaching parents about literacy and gender awareness (70 per cent are female).</p> <p>Engaging local community radio stations and the private sector to promote gender responsive programming and messaging and to engage girls, boys, teachers, parents and communities.</p>
	Sisters for Sisters' Education in Nepal	Lower, upper primary and lower secondary	<p>Training marginalised girls to complete a full cycle of education. These girls provide academic and emotional support to some of the most marginalised girls, or 'Little Sisters', by 'Big Sisters', who mentor the girls through their schooling and act as positive role models. International volunteers train and support the Big Sisters, and work with them to mobilise commitment in communities and resources for the continuation of the scheme.</p> <p>Providing nine-month "Bridge Courses" (preparatory classes and school enrolment support) to girls who have never been to school or who dropped out</p>

Country	Project	Education Focus	Project Description
			<p>in Grades 1 to 3 and learning support classes to low performing girls to help keep them in school</p> <p>Mentoring the Big Sisters through male and female ‘adult champions’ from the local community facilitates negotiations with parents, adding credibility to the scheme.</p> <p>Establishing gender-friendly school environments in schools that the projects’ marginalised girls attend.</p>
	Supporting the Education of Marginalised Girls in Kailali	Upper primary and lower and upper secondary	<p>Conducting an enrolment drive to decrease the information and financial barriers to girls’ education, through working with School Management Committees and Parent Teacher Associations.</p> <p>Establishing after school and out-of-school Girls’ Clubs. The curriculum will cover English, maths, science, life skills and sexual health education for after school clubs and basic numeracy, literacy and sexual health education for out-of-school girls’ clubs.</p> <p>Supporting a small number of female entrepreneurs who sell solar lamps in the community to also act as role models for the girl pupils, promote clean energy in the school and later train some school leavers to become solar lamp entrepreneurs themselves. The Empower Generation organisation will promote clean energy events and train the entrepreneurs in business skills.</p> <p>Setting up a Kailali Girls Transition Fund - a large, sustainable revolving fund, through Saving and Credit Cooperative Societies (SACCOs) - for post education support (vocational training needs and concrete business plans) as girls transition into adulthood.</p> <p>Introducing ‘Clubs of excellence’ awards.</p> <p>Providing ‘Educate Girls: Alleviate Poverty’ Upgrade Award (typically infrastructure) of the community’s choice which may include water serviced</p>

Country	Project	Education Focus	Project Description
Nigeria			female sanitation blocks, tube wells or drinking water provisions, classroom or playground upgrades, boundary walls and gates, and inverters to manage load-shedding.
			Providing training to girls in financial literacy and entrepreneurship. Matching girls to private enterprise service providers and low interest financing, to enable them to access vocational training schools, apprenticeships and business start-up support. Partners include financial institutions and, for example, the Micro Enterprise Development Fund.
	Educating Nigerian Girls in New Enterprises	Senior secondary and vocational training	ENGINE is establishing over 170 learning spaces where girls and young women between the ages of 16 and 19 will meet for academic support and training sessions over a nine-month period. Approximately 5,400 girls who are still in school will receive after-school tutoring, as well as training to advance their leadership and entrepreneurship skills. Additionally, a vocational training programme focused on business and employment readiness will be offered to approximately 12,600 young women who are currently out of school.
Rwanda			Young women who complete the vocational training programme will have the opportunity to choose from a variety of employment opportunities, including receiving assistance to set up their own businesses as micro-retailers of Coca-Cola and d.light products.
	Rwandan Girls' Education and Advancement Programme	Upper and lower primary, secondary	Schools are adopting the "Education that Pays for Itself" self-financing education model, with business and practical skills classes added to the current curriculum, and through setting up income generating activities. Profits generated through the school businesses will pay for costs families cannot afford (e.g. school uniforms, school fees and books).
			Setting up Mother-Daughter Clubs (MDCs) that target the most marginalised girls in the schools and their mothers to run various activities, including

Country	Project	Education Focus	Project Description
Sierra Leone	Supporting Marginalised Girls in Sierra Leone to Complete Basic Education with Improved Learning Outcomes	Upper primary and lower secondary	community outreach on the importance of girls' education and establishing IGAs.
			Installing separate lockable girls' sanitation facilities using ECOSAN composting toilets, with a focus on improving the school environment for girls. The compost from the ECOSAN toilets is being used for the income generating school gardens.
			Broadcasting an educational radio soap opera nationally on Radio Rwanda and the BBC Great Lakes Service, following the success of the radio soap opera "Urunana" in transmitting health messages.
			Improving the access of marginalised girls, allowing them to complete nine years of basic education.
Somalia	Educate Girls, End Poverty	Lower primary, upper primary and secondary	Increasing learning outcomes for girls and building the skills needed for life.
			Improving girl-friendly and inclusive learning environments.
			Ensuring girls' voices and needs are listened and responded to and ensuring their participation in educational decision-making.
			Increasing the number of marginalised girls who enrol and stay in school, supported by their communities, families, schools and mentors
Somalia			Increasing the number of primary and lower secondary schools across Somalia that provide a more gender sensitive environment for learning, and a more relevant quality of teaching for girls.
			Developing the capacity of the Ministry of Education across all zones and regions of Somalia, to provide leadership in promoting girls' education and undertake routine monitoring of gender equality in education.

Country	Project	Education Focus	Project Description
South Sudan			Mobilising communities, mothers and girls to participate routinely and more forcefully in education policy, and the planning, monitoring and budgeting processes for their schools.
	Kobcinta Waxbarashada Gabdhaha – Somali Girls Education Promotion Programme	Primary, lower and upper secondary	<p>Mobilising 173 rural communities to support girls’ education.</p> <p>Recruiting, training and supporting 270 teachers, including 90 females, to provide a relevant, quality education for primary and secondary school rural girls.</p> <p>Constructing culturally appropriate child/girl-friendly learning facilities (or refurbished) and equipping 150 rural primary schools, 20 secondary schools and three secondary school boarding facilities for rural girls.</p> <p>Strengthening Ministry of Education policies and the Quality Assurance function to support the delivery of a relevant, quality education for rural girls in primary and secondary school.</p>
	What’s Up Girls?!	Lower, and upper primary	<p>Addressing key stakeholders (girls, teachers and fathers and other key male stakeholders) combining three innovative methods: School Mothers, the What’s Up?! packages and use of Digital Audio Players (DAPs).</p> <p>Implementing the ‘School Mother’ method that has been successful in the Rumbek East County for the last three years. The ‘School Mother’ method allows women who are respected in the community and who support girls’ education to become advocates that work with communities and parents.</p> <p>Addressing cultural beliefs and rites which are underlying issues preventing girls’ education through the What’s Up?! packages.</p> <p>Providing training using solar-powered DAPs.</p>

Country	Project	Education Focus	Project Description
Tanzania	A Community Based Approach: Supporting Retention, Re-entry and Improving Learning	Upper primary and lower secondary	Setting up girls' study clubs to reach girls who have dropped out of the last grade of primary or early in lower secondary. These girls will receive three hours of learning sessions five days a week. Local women will be trained to provide tutoring/facilitation support to the girls, remaining a resource to the community. Being registered under the government's Institute of Adult Education (IAE), these girls are expected to complete their lower secondary education through the study clubs.
			Providing support for the girls who are at risk of dropping out in the government primary schools through additional subject based tutoring support and support through peer mentors. These girls will receive one hour tutoring sessions, three days a week in mathematics and English. Subject based teachers will be selected from within the school and be provided with additional training on subject matters and pedagogy. Furthermore, girls will be selected and trained as peer mentors who will support other girls and boys in the upper primary grades (6 and 7) in learning, improving attendance and developing understanding of life skills issues.
			Providing training in life skills, covering health, hygiene, reproductive health, pregnancy and marriage, sexual abuse and negotiation skills. The clubs will offer a safe, supportive environment, and peer-to-peer support.
Tanzania and Zimbabwe	A New 'Equilibrium' for Girls	Lower secondary	Targeting the wider community through awareness-raising activities, involving workshops with leaders, radio and theatre campaigns, and collaboration with head teachers and teachers of 100 government schools.
			Increasing the retention and progression of marginalised girls through secondary school. Improving learning outcomes of female and male students.

Country	Project	Education Focus	Project Description
Uganda	Supporting Slum and Homeless Street Girls with Disabilities in Kampala City to access quality Primary Education	Upper and lower primary	<p>Increasing uptake and use of a mobile technology platform that supports education planning and extends learning and networking among young people in rural areas.</p> <p>Empowering secondary graduates to reinvest in the local education system.</p> <p>Developing robust, engaged local capacity and collaboration in support of vulnerable children's education.</p> <p>Informing GEC dialogue, practice and policies in the education sector.</p>
			<p>Identifying, mapping, assessing and enrolling disabled girls into school.</p> <p>Developing an inclusive education teacher training manual and capacity building module for teachers.</p> <p>Training families on disability and income generation, and providing support with business start-up activities.</p> <p>Adapting 10 schools' infrastructure so that disabled girls can have easier access. Providing schools with accessible materials (braille, sign language charts etc.) and assistive devices (wheelchairs, glasses, hearing aids etc.).</p> <p>Providing transport for disabled girls and paying school fees. Individual Education Plans will be developed for each girl and sign language interpreters will be available in classrooms.</p> <p>Engaging school students and parents of non-disabled children in these schools in discussion on disability and inclusive education through Child-to-Child clubs.</p> <p>Establishing an Inclusive Education Resource Centre in every school that will focus on: education and medical assessment, remedial teaching, therapeutic services, counselling, etc.</p>

Country	Project	Education Focus	Project Description
	Good School Toolkit: Creating a Violence-Free and Gender Equitable learning Environment at School	Upper and lower primary	<p>Extending the Good School Toolkit rollout to additional schools in Kampala, Luwero, Lira and Kabaloro. The Toolkit will emphasise the supportive learning environments that are needed to retain and teach marginalised girls. A ‘Good School’ consists of good teachers, a good learning environment and a responsive and progressive school administration.</p> <p>Establishing activism centres in the four implementing districts of Uganda in collaboration with eight partner organisations. These will support schools and communities, engage with toolkit ideas and increase support for marginalised girls’ education in these districts.</p> <p>Launching a community activism and multimedia campaign that will engage the communities around the schools in on-going dialogues about girls’ education. The campaign will reinforce the community based discussions and school-based interventions using local and national TV, newspapers, magazines, radio and other non-traditional communication media.</p>
	Creative Learning Centres (CLCs) for Girls aged 10-18 in Greater Kampala	Upper primary, lower secondary	<p>Linking community based organisations to schools through a network approach that provides non-formal education in a wide range of subjects, in order to encourage girls to re-engage with their education.</p> <p>Setting up 20 Creative Learning Centres to deliver education that addresses girls’ needs in Greater Kampala. The most marginalised girls are identified and each girl creates an individual learning action plan with the help of dedicated and trained female teachers.</p> <p>Training teachers in this accelerated learning programme and offering mentoring throughout the programme. Voluntary classroom assistants are engaged to support these teachers</p> <p>Offering support to the families of marginalised girls enrolled in the CLCs through mentors that encourage and support girls’ education.</p>

Country	Project	Education Focus	Project Description
			Linking the CLCs to enable girls to engage in an inter-school league and an annual sports competition through a mobile resource unit (with books, media and sports equipment).
	Keeping Marginalised Girls in School by Economically Empowering their Parents	Lower, upper, primary and lower secondary	<p>Economically empowering marginalised mothers in Uganda by turning them into micro-retailers of their clean burning fuel briquettes. Each of these mothers will earn at least \$152/month from retailing EFA's briquettes. They will be contractually obliged to spend the income they generate from selling the fuel briquettes on sending and keeping their daughters in school.</p> <p>Providing transportation services for girls who are either disabled or who live over 4km from the schools. Recruiting a female mentor/role model to support each family. These trained female counsellors ensure that marginalised girls benefit from this approach, visiting girls in school and mothers in their homes.</p> <p>Conducting community and school-based sensitisation campaigns to enlighten parents, teachers and community leaders about the importance of educating girls and to also inform them about government laws that prohibit early marriages.</p> <p>Training teachers to improve learning in schools.</p> <p>Providing guidance and counselling to girls through awareness raising on sexual abuse issues, sensitisation and the promotion of codes of conduct for schools as well as by encouraging girls to report abuse. A range of activities are included in the project, e.g. professional counsellors, talking compounds in schools, girls clubs, peer learning / debates, and advocacy for girls representation on school leadership committees.</p>
	Innovating in Uganda to Support Educational Continuation	Primary and Secondary	Delivering project deliverables centred around micro-finance activities - savings, loans, insurance and financial education, over two to three years to low (and medium) cost private schools and to households.

Country	Project	Education Focus	Project Description
	by Marginalised Girls in relevant Primary and Secondary Education		<p>Providing school improvement loans and training to school proprietors in order to build infrastructure and improve their educational services.</p> <p>Providing parents with school fee loans at all grades, in particular to support the attendance of girls at upper primary and lower secondary levels. School fee loans are intended to address cash flow issues to economically active poor households. Average loan size is three to six months, based on school terms and repaid in weekly or monthly instalments.</p> <p>Adapting and delivering a financial education programme ('Aflatoun') to girls in schools. By improving the quality of education provided by low-cost private schools, this project aims to demonstrate the potential of low-cost private schooling in providing accessible, affordable, relevant and quality education.</p> <p>Opening child savings accounts for girls to enable families to save for school materials and fees.</p> <p>Encouraging parents to save and qualify for 'EduSave' - an insurance-linked savings product to protect children's schooling against the death or permanent disability of a parent.</p>
	Girls Enrolment, Access, Retention and Results	Secondary	<p>Providing low-cost, quality and sustainable secondary education. The project focuses on four key areas: enrolment, attendance, retention and results. It provides a relevant and partly vocational education to girls in schools with improved gender-appropriate facilities and practices.</p> <p>Improving attendance by using a mobile-phone based school information management system to understand the barriers to girls' access and identify girls at risk of dropping out.</p> <p>Focusing on the safety of girls in school, including building sanitation facilities and water points at new schools to make them more girl-friendly. These are accompanied by lessons on hygiene and safety.</p>

Country	Project	Education Focus	Project Description
Zambia	Child Centred Schooling: Innovation for the Improvement of Learning Outcomes for Marginalised Girls in Zambia	Upper primary	Conducting research into the issue of harassment and teasing of girls while in school. This will be new research and will improve understanding of this issue and lead to programming of more effective responses to this issue.
			Adapting the PEAS curriculum to become more relevant to the lives of girls, and including gender sensitive health messages into in the community engagement plan; and including supplementary curriculum material to have a targeted focus on literacy and numeracy.
			Introducing some elements of vocational training and training teachers in the implementation of gender-responsive pedagogy.
			Supporting the retention and progression of vulnerable girls through primary school by providing Safety Net Fund cash transfers, psychosocial support from trained Teacher Mentors and zero tolerance Child Protection initiatives (key pillars of the Camfed Model).
Zimbabwe	Improving Girls' Access through Transforming Education	Primary and lower secondary	Improving learning outcomes for marginalised girls by training and supporting teachers to integrate into their teaching practices the Fundacion Escuela Nueva (FEN) child-centred pedagogy and learning resources, designed for children to lead and assess their own learning, facilitated by teachers.
			Mobilising members of Cama – the Camfed network of educated rural young women – to monitor the project's progress, act as role models in schools and build data literacy and ownership of the project.
			Increasing household economic capacity to support and prioritise girls' education through the Village Savings and Loans model.
			Mobilising target communities to support equal education and tackling barriers to girls' attendance through Mothers' Groups, School Development

Country	Project	Education Focus	Project Description
			Committees, religious bodies local traditional leaders, male champions and partnership with the girls themselves.
			Developing the capacity of School Development Committees to lead participatory management of schools, gender sensitive programming and initiatives such as mechanisms for reporting abuse and support connected to menstrual hygiene and WASH.
			Mobilising target communities through social accountability activities, school score-carding and action maps, in partnership with all stakeholders including the Zimbabwe Government.
			Supporting schools, communities and Mothers through the Power Within model.
			Improving male involvement and “men’s voice for change” through Male Champion support.
			Training parents on menstrual hygiene and the creation of Reusable Menstrual Pads.
			Reducing the barrier of distance through provision of bicycles for both boys and girls.
			Increasing children’s capacity in reading fluency and comprehension targeting improved literacy through the roll-out of Literacy Improvement Reading Camps.

The learnings from the GEC collection of projects provided valuable insights around what inputs proved effective in improving outcomes for girls. A combination of five factors were found to be most effective for focusing efforts moving forward in girls education: regular in-school coaching for teachers to improve their practice, along with structured teaching and learning materials for use in the classroom; extracurricular activities such as girls and boys clubs aimed at improving girls' motivation and self-esteem; regular collation of data on girls' learning and their participation in education and extra-curricular activities to be used to make programmatic decisions; the recognition of the need to work with boys and men – especially girls, not only girls; and, engagement at three levels – with communities, school governance, and national policymakers – to promote change (UK Aid, 2018).

In 2016, GEC Phase II was implemented with a budget of approximately 450 million GBP. This phase of GEC will enable up to 1 million marginalized girls (currently supported through Phase 1) to continue to learn, complete primary school and transition on to secondary education. A further 500,000 highly marginalized adolescent girls, who are out of school, are targeted to gain literacy, numeracy and other skills. It is estimated that at least 400,000 girls will complete junior secondary school in the first four years of the extension. The extension will build on what we have learnt so far in Phase 1 and further deepen global understanding of what works for girls' education, particularly during adolescence and in the transition from education to work.

An additional 108 million GBP were allocated to under the Leave No Girls Behind funding structure specifically allocated for highly marginalized adolescent girls. This initiative supports interventions providing literacy, numeracy and skills relevant for life and work to highly marginalized, adolescent girls who have never attended or have already dropped out of school. These are girls who experience complex marginalization because of their circumstances. These

include orphans, married or young mothers, girls with a disability, nomadic girls, refugees, those from the poorest communities and those with no access to education.

Finally, the Girls Education Challenge Transition Phase (GEC-T) was announced for implementation 2017-2024 with a budget of 272 million GBP. Projects under GEC-T are intended to support girls in their transitions from primary to secondary, post-primary and secondary, or skills training institutes, and through to employment to provide the much needed support to keep girls in school and provide opportunities for employment and advancement previously unavailable to them.

VAS Y Fille! Program

In 2012, the International Rescue Committee (IRC) received \$41.3 million USD in funding from DFID's GEC program in order to implement a program to increase access and quality of education for girls in the DRC. In conjunction with Save the Children UK and Catholic Relief Services, the IRC implemented the Valorisation de la Scolarisation de la Fille (VAS-Y Fille!) program in 400 schools covering five provinces (Bandundu, Equateur, Katanga, Kasai, and Province Orientale) in the DRC, and the program was implemented over three years (2013-2017).

VAS-Y Fille! was designed with four key outputs intended to complement one another to remove barriers to marginalized girls' access to education and to improve learning outcomes:

Increased parental financial capacity to support girls to succeed in primary education. The project hypothesized that when families have increased financial capacity and value education for girls, families will choose to allocate resources to girls education increasing enrollment, re-enrollment, and attendance for girls, leading to improved learning outcomes.

Increased quality and quantity of reading and math opportunities. The project hypothesized that when girls receive better quality or increased hours of instruction, learning outcomes will improve.

Increased community involvement to ensure girls' access to quality education in a safe environment. The project hypothesized that increased parental involvement in school management, thereby gaining more control over decisions, and insight into the value of education, parents will view the school more favorably and therefore be more willing to enroll their girls in school, and support their regular attendance and completion.

Increased civil society engagement in providing alternative learning opportunities for out-of-school girls to catch up and complete primary school. The project hypothesized that more available accelerated learning program (ALP) classes at no cost to young girls will encourage greater numbers of out of school girls to enroll and complete their primary school experience.

Figure 3 shows a visual representation of the program's hypothesized theory of change, indicating the effects, outcomes, and impact of the VAS-Y Fille! Program. Interventions implemented in the program included the following: 1) scholarships to pay for education fees and school supplies paid monthly and dependent on attendance, 2) saving and credit groups working within communities to increase the financial capacity of families, 3) tutoring in reading, writing, and mathematics, 4) teacher training in advanced methods for teaching reading, writing, and mathematics, 5) teacher training on increasing girl-friendliness in the classroom, 6) awareness raising activities in the community such as text messages and other communications media regarding the program, and female leader advocacy in favor of education of girls, and 7) recruitment and training of Accelerated Learning Program (ALP) teachers. ALP is a three-year

program offered to older girls who have no completed primary school, in order to allow them to move on to secondary school classes.

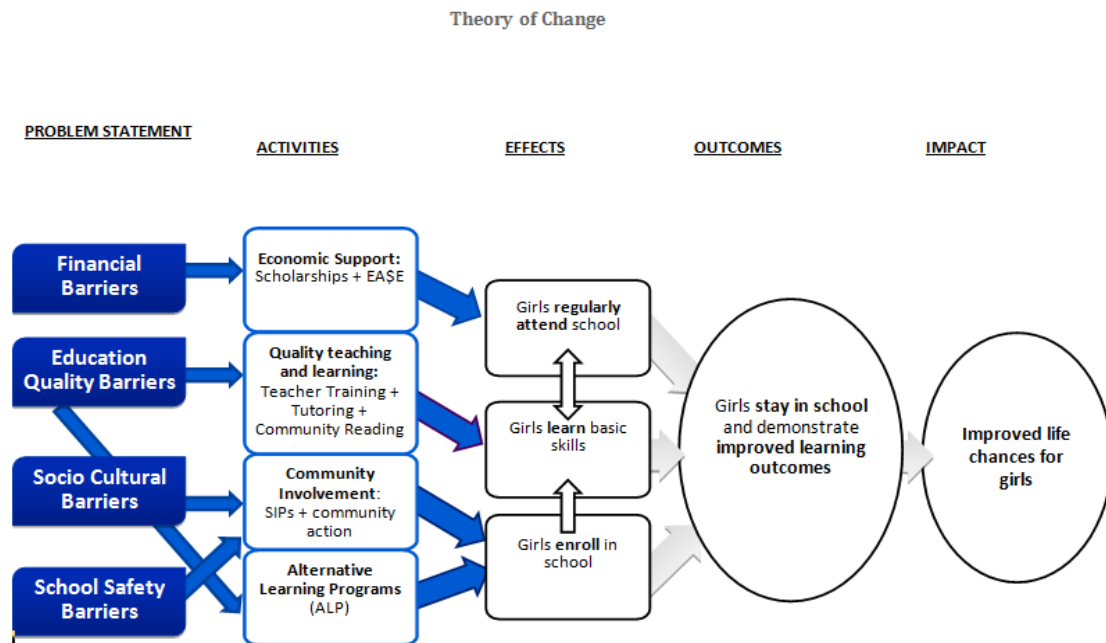


Figure 3. VAS-Y Fille! Program theory of change (UK Aid, 2013)

While VAS Y Fille! programming was designed to overcome barriers to education in the DRC holistically, the interventions were discrete, and beneficiaries did not necessarily see all aspects of the program. Rather, interventions were administered to groups based on needs of a student or family using targeted criteria. Table 2 shows the planned program activities by population, with details around the specific exposure to the program components outlined (table from the VAS-Y Fille Monitoring and Evaluation Framework; UK Aid, 2013).

Table 2. VAS-Y Fille! populations and intended program component exposure (UK Aid, 2013)

	\$	S	T	T/S	TT	LB	CFP	ALP	Exposure to Interventions
Community						X	X		Literacy Boost (LB) has no targeting criteria for service provision but a limited number of spots. Community focused programs (CFP) have no targeting criteria except for those that focus on school improvements.
Parents of scholarship recipients	X					X	X		Parents of scholarship recipients (grade 5&6) will be encouraged to join EA\$E groups. They could also benefit from LB or CFP. This direct benefit may lead to indirect benefits for girls from these families, specifically girls in Grade 5 and 6.
Out-of-school girls (ages 6-18)						X	X	X	ALPs have only age as a targeting criterion (9 to 15 years old). Out-of-school girls could benefit directly from any of the three interventions.
Grade 1 – Girls					X	X	X		Teacher training (TT) is a blanket intervention at the school level, so all students will benefit from TT. No targeted interventions are provided for grades 1&2. However, these students may benefit from certain pieces of the LB and CFP.
Grade 1 – Boys					X	X	X		
Grade 2 – Girls					X	X	X		
Grade 2 - Boys					X	X	X		
Grade 3 – Girls			X		X		X		

	\$	S	T	T/S	TT	LB	CFP	ALP	Exposure to Interventions
Grade 3 – Boys			X		X		X		In grade 3, 15 academically vulnerable girls and 10 academically vulnerable boys per class per year will benefit from tutoring (T). All children, even who do not benefit from T, will benefit directly from TT. Children who do not directly benefit from T may benefit indirectly through increased motivation or a smaller standard deviation in class skill level. All grade 3 students may benefit from CFP.
Grade 4- Girls			X		X		X		In grade 4, 15 academically vulnerable girls and 10 academically vulnerable boys per class per year will benefit from tutoring (T). All children, even who do not benefit from T, will benefit directly from TT. Children who do not directly benefit from T may benefit indirectly through increased motivation or a smaller standard deviation in class skill level. All grade 4 students may benefit from CFP.
Grade 4- Boys			X		X		X		
Grade 5 - Girls		X	X	X	X		X		In grade 5, 15 academically vulnerable girls per class per year will benefit from T. Approximately 20 economically vulnerable girls will receive a scholarship (S). These two groups of girls may overlap up to

	\$	S	T	T/S	TT	LB	CFP	ALP	Exposure to Interventions
									100%. All girls, even those who do not benefit from T, S or T/S, will benefit directly from TT. Non-recipient girls may benefit indirectly from T, S or T/S through increased motivation or a smaller standard deviation in class skill level. All grade 5 students may benefit from CFP.
Grade 5 – Boys					X		X		In grade 5, 10 academically vulnerable boys per class per year will benefit from T. However Grade 5 boys are not eligible for S, thus nor T/S. They will continue to benefit directly from TT and may benefit indirectly from interventions aimed specifically at girls. All grade 5 boys may benefit from CFP.
Grade 6 – Girls		X			X		X		In grade 6, approximately 20 economically vulnerable girls will receive a scholarship (S). All girls, even those who do not benefit from S, will benefit directly from TT. Non-recipient girls may benefit indirectly from S through increased motivation or a smaller standard deviation in class skill level. All grade 6 students may benefit from CFP.

	\$	S	T	T/S	TT	LB	CFP	ALP	Exposure to Interventions
Grade 6 - Boys					X		X		Grade 6 boys are not eligible for S. They will benefit directly from TT and may benefit indirectly from interventions aimed at girls. All grade 6 students may benefit CFP.

In fall of 2013, VAS-Y Fille program staff coordinated a baseline study for the purposes of identifying the control and intervention cohorts to be tracked over time for the evaluation. The baseline data collection included quantitative (i.e., household and school survey data) and qualitative (i.e., qualitative interviews around marginalization of girls and barriers to education) data. Table 3 outlines the sample size, sampling strategy, and data sources for the quantitative survey data collection. Using the quantitative data collection to identify a sample of informants (planned as $n = 162$) to further comment on marginalization of girls, barriers and opportunities for education, and to allow for a better understanding of household decision making on education.

Table 3. Baseline study quantitative data collection plan (UK Aid, 2013)

	Household Survey (HH)	School-Survey
Sample Size	approximately 1440 households across 86 intervention and non- intervention communities	86 schools in intervention and non- intervention communities (40 girls per school from grades 3-6)
Sampling Strategy & Eligibility Criteria	Random selection; HH must be in catchment area of school sampled for school survey.	Random, stratified selection of 40 girls from grades 3-6 per school. Girl may or may not directly benefit from an intervention in Year 1.
Data sources	caregiver of child per HH or Head of household, out of school girl if applicable	Girl students, school director, administrative school data

VAS-Y Fille! Program Evaluation

The ultimate goal of the evaluation of the VAS-Y Fille! program is to test the hypothesized theory of change (see Figure 3). Out of the theory of change, six evaluation questions were identified across three program facets: program impact, program effectiveness, and program financial efficiency (evaluation questions are outlined in Table 4). At the outset, VAS-Y Fille! program staff also identified evaluation logframe indicators along with yearly targets for the program across the four outputs (see Table 5).

Table 4. VAS-Y Fille! program evaluation questions (UK Aid, 2013)

Program Impact	What is the impact of the VAS Y Fille package of support (scholarships, EASE groups, teacher training, tutoring, community reading activities, parent participation and ALPs) on enrolment, learning, attendance and retention of girls?
	What effect did VAS Y Fille have on community attitudes and behaviors towards girls' education?
Program Effectiveness	Which components of VAS-Y Fille appeared to impact learning the most?
	Which components of VAS-Y Fille appeared to impact attendance and retention the most?
	Which components of VAS-Y Fille appeared to impact learning the least?
	Which components of VAS-Y Fille appears to impact attendance and retention the least?
Program Financial Efficiency	What is the cost-effectiveness of VAS-Y Fille?
	What is the most cost-effective combination of VAS-Y Fille components?

A longitudinal design was chosen to allow for greater reliability and insight into changes at the individual level, and the study involved choosing both a control and intervention group by way of the baseline study outlined above.

Table 6 shows the original cohort design of the project, and the design intended to follow all cohorts through the Midline and Endline data collections as feasible. In order to show evidence of the program's intended outcomes around enrollment, attendance, and learning outcomes in mathematics and reading, several metrics and tools were selected for data collection and are outlined in Table 7.

Table 5. VAS-Y Fille! project outputs with targets by year (UK Aid, 2013)

Output	Indicator	Year 1 Target	Year 2 Target	Year 3 Target
Output 1: Increased parental funding capacity to support girls to success in and complete primary education.	Indicator 1.1: Percentage of girls receiving scholarships regularly attend school	100% of girls receiving scholarships regularly attend school (16,000 girls)	100% of girls receiving scholarships regularly attend school (40,000 girls)	100% of girls receiving scholarships regularly attend school (56,000)
	Indicator 1.2: Percentage of 12,000 parents participating in EA\$E groups have increased financial assets to afford girls' education	0% of 12,000 parents participating in EA\$E groups have increased financial assets to afford girls' education	80% of 4,000 parents participating in EA\$E groups have increased financial assets to afford girls' education	90% of 12,000 parents participating in EA\$E groups have increased financial assets to afford girls' education
	Indicator 1.3: Average percent increase in spending on education-related expenses for girls by EA\$E participants	0% increase in spending on education-related expenses for girls by EA\$E participants	10% increase in spending on education-related expenses for girls by EA\$E participants	15% increase in spending on education-related expenses for girls by EA\$E participants
Output 2: Increased quality and quantity of reading and math instruction.	Indicator 2.1: Percentage of teachers applying improved teaching practices in the classroom	40% of teachers trained apply improved teaching practices (1,120 teachers)	60% of teachers trained apply improved teaching practices (1,680 teachers)	80% of teachers trained apply improved teaching practices (2,240 of 2,800 teachers)
	Indicator 2.2: Monthly average number of	Average of 6 additional instructional hours per	Average of 8 additional instructional hours per	Average of 10 additional instructional hours per

Output	Indicator	Year 1 Target	Year 2 Target	Year 3 Target
	additional instructional hours per child enrolled reading & math tutoring	child enrolled reading & math tutoring	child enrolled reading & math tutoring	child enrolled reading & math tutoring
	Indicator 2.3: Percentage of community members that participate in literacy activities with their children	0% of community members that participate in literacy activities with their children	7% of community members that participate in literacy activities with their children	10% of community members that participate in literacy activities with their children
Output 3: Increased community/ COPA involvement ensures girls' access to quality education in a safe environment	Indicator 3.1: Percentage of community members participating in COPA-led awareness raising activities (disaggregated by sex and age group)	4% more community members participating in COPA-led awareness raising activities (disaggregated by sex and age group)	7% more community members participating in COPA-led awareness raising activities (disaggregated by sex and age group)	10% more community members participating in COPA-led awareness raising activities (disaggregated by sex and age group)
	Indicator 3.2: Percentage of community members who report their comprehension on the importance of girls education has improved	0% community members who report their comprehension on the importance of girls education has improved	3% community members who report their comprehension on the importance of girls education has improved	6% community members who report their comprehension on the importance of girls education has improved
	Indicator 3.3: Percentage of gender-	0% of gender-enhanced SIPs completed by	50% of gender-enhanced SIPs completed by	90% of gender-enhanced SIPs completed by

Output	Indicator	Year 1 Target	Year 2 Target	Year 3 Target
	enhanced SIPs completed by COPAs to create safe learning environments	COPAs to create safe learning environments	COPAs to create safe learning environments (200 of a total 400 projects financed)	COPAs to create safe learning environments (360 of a total 400 projects financed)
	Indicator 3.5: Percentage of girls and parents who report the school environment as being more girl-friendly	0% of girls and parents who report the school environment as being more girl-friendly	3% of girls and parents who are report the school environment as being more girl-friendly	5% of girls and parents who are report the school environment as being more girl-friendly
Output 4: Increased civil society engagement in providing alternative learning programs for out-of-school girls to catch up and complete primary school	Indicator 4.1: Number of students enrolled in Accelerated Learning Programmes (ALP) (disaggregated by sex)	4,320 children (2160 girls) enrolled in Accelerated Learning Programmes (ALP)	5,520 children (2760 girls) enrolled in Accelerated Learning Programmes (ALP)	6,652 children (3,326 girls) enrolled in Accelerated Learning Programmes (ALP)
	Indicator 4.2: Number of ALP students who remain in the ALP program during the project cycle (measured by attendance & retention)	3888 (1944 girls) of ALP students who remain in the ALP program during the project cycle (measured by attendance & retention)	4968 (2484 girls) ALP students who remain in the ALP program during the project cycle (measured by attendance & retention)	5,988 of ALP students who remain in the ALP program during the project cycle (measured by attendance & retention)

Table 6. VAS-Y Fille! cohort design (UK Aid, 2013)

	Cohort A	Cohort B	Cohort C	Cohort D
Survey	Household	Student	Student	Student
Grade level or age at baseline	Out-of-school girls 6-15 years of age	Girls in grade 5 and 6	Girls in grades 3 and 4	Girls in new ALP centers
Potential Treatment exposure at baseline⁶	Community educational outreach activities	Scholarships, tutoring, both scholarships and tutoring, or blanket interventions	Tutoring, or blanket interventions	Accelerated learning programme
Data Collection Timing	Baseline, Midline, Endline	Baseline, Midline, Endline (original 5th graders only)	Baseline, Midline, Endline	Baseline, Midline, Endline

Table 7. VAS-Y Fille! indicator matrix (UK Aid, 2013)

Indicator	Measure	Data Source	Data Collection	Expected Impact
Student Enrollment	Total number of students enrolled in a school disaggregated by grade and sex.	School records and self-report for household data.	Beginning of school year.	Increased enrollment for girls and improved gender parity.
Primary School Completion	Total number of students successfully completing or graduating from the final grade of primary school in a year, disaggregated by sex.	Certification and exam records.	Beginning of school year for previous year completion.	Triangulation for self-reported data; comparison data.
Retention	The total number of students belonging to a school-cohort who reached each successive grade, disaggregated by sex.	School records.	Beginning of school year.	Triangulation for self-report data.
Drop-Out	The number of students who fail to complete a given level of schooling, disaggregated by grade and sex.	School records.	Beginning of school year for previous year.	Triangulation for self-report data.
School Attendance	Average number of school days attended by students.	School records.	Random collection of records.	Triangulation for self-report data.

Indicator	Measure	Data Source	Data Collection	Expected Impact
Learning Outcome	Mean score on each subtest of Early Grades Reading Assessment (EGRA) and Early Grades Mathematics Assessment (EGMA), disaggregated by sex.	EGRA and EGMA	Assessments administered at Baseline, Midline, and Endline for all girls in evaluation.	Comparison data over time and across groups.

Issues in International Development Evaluation

Hundreds of millions of dollars are put forth each year by major international aid organizations in sectors such as health, agriculture, and education. The British Department for International Development (DFID) reports their 2020/21 FY funding for educational programs in the amount of 572.4 million GBP (<https://devtracker.dfid.gov.uk/>), with nearly 30% of the budget allocated specifically to primary education programs. And the United States Agency for International Development (USAID) reported a budget for the 2021 fiscal year with 430.5 million USD pledged for primary education programs (<https://www.usaid.gov/sites/default/files/documents/9276/FY-2021-CBJ-Final.pdf>). With the total international aid budget from just two organizations surpassing 30 billion USD, the importance of accountability of spending cannot be understated, and the ability to demonstrate positive outcomes of funded initiatives becomes more and more important.

In recent history, these organizations demonstrated success through an adequate accounting of inputs (e.g., dollars spent on professional development for teachers or on scholarships) and immediate outputs (e.g., test score changes, enrollment rates, attendance, etc.). However, due to the introduction of initiatives such as the Millennium Development Goals (<http://www.un.org/millenniumgoals/>), Sustainable Development Goals (<https://www.un.org/sustainabledevelopment/sustainable-development-goals/>), and pay-for-performance incentives (<http://go.worldbank.org/FVDDBVIZD0>), development aid organizations have shifted their focus to the desired long-term intended outcomes of a program (e.g., quality of life increases, employment rates, etc.; Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2011).

Impact evaluation seeks to establish a link between the desired outcomes of a program or intervention to said program or intervention. Establishing a link between a program and outcomes requires the collection and evaluation of both factual and counterfactual evidence (Gertler et al., 2011; Howard White & Raitzer, 2017). Factual evidence includes information regarding the measurement of outcomes for those receiving the benefit of the intervention or program such as achievement tests for an educational program. Counterfactual evidence includes the same information regarding outcomes for those individuals not receiving the treatment or intervention, allowing for a direct comparison between the two groups.

Adequately evaluating counterfactual evidence required in an impact evaluation, randomization of the assigned intervention to equivalent groups is necessary; this is referred to as a randomized controlled trial (RCT) design (Jones, Jones, Steer, & Datta, 2009). RCT designs require collection of quantitative outcome data from comparison groups in an attempt to estimate the impact of the

intervention through calculations of the differences between the groups (Torgerson, Torgerson, & Taylor, 2015).

International development contexts provide unique challenges for evaluators and researchers in collecting reliable and trustworthy quantitative data as well as in carrying out a strict RCT design. Possible challenges to such projects cited by White (2005) White and others include:

1. changes to the staff managing or implementing the project, or to the evaluation or research team, throughout a long-term project,
2. spillover effects, or indirect treatment effects can cause over- or under-estimates of treatment effects, leading to problematic policy decisions, and, of particular interest to the current study,
3. inadequate funding for the development and adaptation of standardized instruments or well-trained data collection staff.

Summary

Given the impetus of impact evaluation in international development contexts, the rigor of outcome measures becomes even more important. The challenges surrounding the development and implementation of education programs in developing countries experiencing conflict are numerous and wide-reaching. Barriers such as a dearth of supplies and infrastructure, underqualified teachers, and overall attitudes toward education present unique and serious challenges to development work (GCPEA, 2018; UNESCO, 2018). These barriers coupled with the noted difficulties with tight timelines and a lack of program staff trained in the process of

adaptation of measures, make not only implementation but evaluation of programs extremely difficult. Therefore, a focus on the reliability and sensitivity of the measures used to evaluate program outcomes is required.

This study uses Generalizability Theory (GTheory; Brennan, 1992; Shavelson & Webb, 1991) and Rasch Measurement Theory (RMT; Andrich, 1978; Rasch, 1980; Wright & Masters, 1982) to assess possible sources of unreliability in data taken from an international evaluation to be used as evidence of success in outcomes of an educational initiative. In both a Generalizability study (GStudy) and a Many-Facet (MF) model (Linacre, 1989), the researcher can identify and select pertinent facets (factors that may be sources of variance) and look at them in relation to one another, allowing us to attribute smaller or larger sources of variability to a particular facet. For example, in the case of cross-cultural research, these possible facets may include country or region, language of instrument, first language of the participant, and enumerator or rater. The lower the error variance in the data, the higher the quality, or reliability (Bayerl & Paul, 2007). By parsing apart the sources of variability, we can see the impact of a particular facet on the quality of the data, and make requisite changes in order to improve it, making the results particularly useful in pilot or longitudinal studies wherein there is a possibility of adapting or editing the instrument.

Therefore, the primary research question guiding this research is: How can Generalizability Theory and Rasch Measurement Theory be used to assess the reliability of cognitive and non-cognitive outcome measures used in an international development education evaluation? The current study will use GTheory and the MF model to analyze a baseline dataset from an international development education evaluation particularly when coupled with

inadequate adaptation of non-cognitive measures. Two types of measures will be assessed: a set of subjective, or affective, survey items, and an objective achievement measure of reading.

Conducting analyses on both types of measures will allow a more comprehensive discussion on the usefulness of the two analytic methods in evaluations such as this. The results of the analyses will also inform the results of an informal translation process used with non-cognitive measures as well as informing the validity and reliability of a commonly used early grades reading assessment.

CHAPTER II. LITERATURE REVIEW

Educational Interventions in Developing Nations

As noted in Chapter I, the stated second Millennium Development Goal (MDG) was to achieve universal primary education, to “ensure that, by 2015, children everywhere, boys and girls alike, will be able to complete a full course of primary schooling” (U.N., 2015, pp.24).

Three indicators were used to provide evidence of meeting the said goal: 1) a net enrollment ratio in primary education, 2) an increased proportion of students reaching grade 5, and 3) an increase in the literacy rate of 15-24 year-olds. While significant progress was made toward achieving this goal, girls, children in particular geographic regions (i.e., sub-Saharan Africa), and children in conflict zones, continue to fall behind (U.N., 2015, pp.25).

Building on the MDGs, in 2016, 17 Sustainable Development Goals (SDGs) were implemented as a part of the 2030 Agenda for Sustainable Development (U.N., 2015b). Goal 4 of the SDGs, to “ensure inclusive and quality education for all and promote lifelong learning” (U.N., 2015b, pp.14), builds upon and expands the successes of Goal 2 of the MDGs. There are many ways in which developmental aid agencies have provided assistance and intervention to developing nations in order to help them meet their goals. This section will serve to discuss the types of educational interventions most commonly used, and the results of these interventions having been implemented.

Educational interventions being implemented in developing nations can be grouped into two general types: demand interventions and supply interventions (Krishnaratne, White, & Carpenter, 2013). As shown in Figure 4, demand-type interventions include programs designed to

reduce educational costs, provide information to students and their families, and increasing preparedness of students to attend and succeed in school. Supply-type interventions, on the other hand, include providing infrastructure, people, professional development, and management strategies. Traditionally, interventions have focused on the supply-side, providing building, teachers, and classroom materials. However, there have been recent increases in providing assistance on the demand-side, focusing more on education quality than simply the quantity of schools or classrooms.

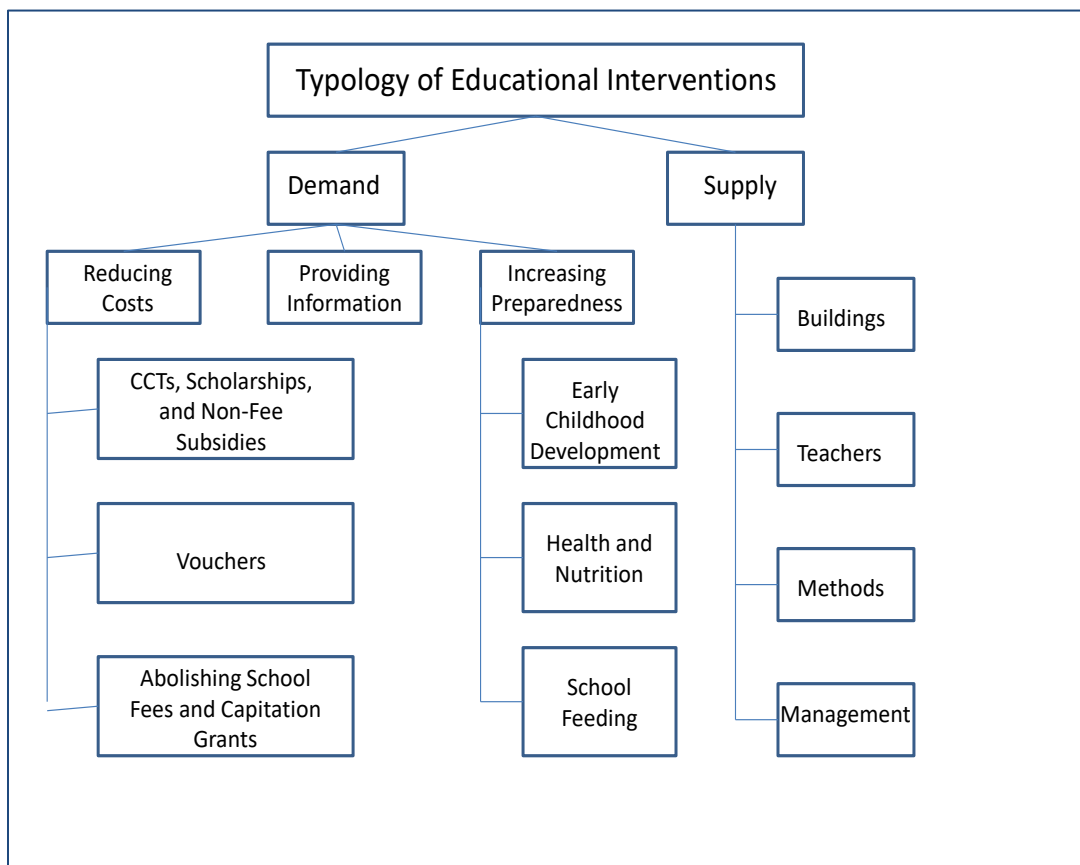


Figure 4. Approaches to Educational Interventions in Developing Nations (Krishnaratne et al., 2013)

New Schools and Infrastructure

On the supply-side of educational interventions, the first, and most common, is providing new schools and infrastructure to an area or country. School infrastructure projects generally supply the actual materials for building a structure, textbooks, and even teachers such that either new schools may be built in communities where there are not enough, or no schools currently, or they may serve to improve current facilities. In Afghanistan, a randomized evaluation investigated the effect of distance to school on a child's enrollment and achievement. The Partnership for Advancing Community-Based Education in Afghanistan is a five-year initiative to provide educational opportunities to children where there is a lack of formal schools. Community-based schools were randomly assigned to be built in subsets of villages within a particular region of the country. Schools were created in either 2007 or 2008, and as such, villages to receive their school in 2008 were used as the control group, and the treatment group consisted of the villages receiving a school in 2007. Results showed large increases in enrollment and test scores among all children with the intervention. Girls were shown to be more sensitive to distance effects than boys, and introducing a community-based school all but eliminated the gender gaps for enrollment and achievement after one year (Burde & Linden, 2013).

The Burkinabe Response to Improve Girl's Chances to Succeed (BRIGHT) program initiated in Burkina Faso was implemented in 2005 through 2008, and provided for the building of 132 primary schools along with a set of additional interventions to be provided. Data collected included household surveys administered a year after implementation, language and math tests administered immediately after the household survey, school surveys including school information and enrollment and attendance data, and application data from all villages who

applied to be a part of the BRIGHT program (this provided comparison data). Results indicated positive impacts on school enrollment, with a slightly larger impact for girls, as well as positive impacts on math and language scores, with equal impacts for boys and girls. These results are preliminary, and as such, it has yet to be determined if these increases have been sustained over time (Levy, Sloan, Linden, & Kazianga, 2009).

Materials

The second type of supply-side interventions include providing schools and teachers with materials needed in schools. These materials may include chalkboards, books, computers, teaching guides, and even specialized teachers. These interventions are thought to affect learning outcomes by improving school and teaching quality. In China, a randomized experiment compared 26 treatment schools with 31 control schools after one semester of computer assisted language learning provided at the third grade. Students in the treatment group were expected to participate in two 40-minute sessions per week during lunch break or after school that emphasized the language curriculum. Students in the control group participated in the traditional language curriculum in class. Pre- and post-implementation surveys were completed with both groups which consisted of a language and math component, as well as a demographics component. Results indicated that those students who received the treatment had significantly improved language and math scores over the control group students. In addition, those students in the treatment group reported higher levels of self-efficacy in school as well as higher levels of self-confidence (Lai et al., 2012) While results are encouraging, they could be due to the fact that students in the intervention group are receiving a full 80 minutes of extra instruction each week.

Banerjee, Cole, Duflo, and Linden (2007) evaluate two interventions providing remedial help to children in urban schools in India that cater to poor families. The first intervention provides a teacher's aide for children in grades three or four who have not mastered basic math concepts. The aide takes the children out of the classroom for two hours each school day (half of the day). The second intervention is provided to all children in grade four but is adapted to the current achievement level of the individual student. The second intervention is a computer assisted learning program that is offered two hours per week during which students play games involving solving math problems that adapt to their ability as they play. Results indicated that the first intervention increased average test scores even in the first year, and by a greater amount the second year with the weakest students showing the highest gains. Results for the computer assisted learning intervention showed a similar increasing pattern of test score increases over time, but this was consistent across all groups of students. Both interventions, however, show nearly a return to baseline only one year after the program ends.

A randomized intervention providing computers to be integrated into the teaching of language in public schools in Columbia over a period of two years showed little effect on learning and other outcomes. The program Computers for Education was implemented randomly across a sample of 97 schools, wherein the number of available computers was increased in treatment schools, and training was provided to teachers regarding how to incorporate the technology into their language instruction. Results showed no increase in achievement or any other outcomes. When probed, this failure appears to be due to the fact that the program was not implemented as intended. That is, teachers did not incorporate the technology into their classrooms and curriculum (Barrera-Osorio & Linden, 2009). However, it is possible that even with implementation the results may have been similar.

Teachers

The third type of supply-side interventions is the addition of teaching resources such as extra teachers, financial incentives for teachers based on student achievement, and providing teachers with resources or aids for the classroom. In Pakistan, the Community Support Process (CSP) program was an experiment to create community support for primary schools for girls, thereby increasing girls' enrollment in segregated schools to be taught by local female teachers. In order to allow for the lack of qualified female teachers in rural areas, the qualifications for teaching were lowered slightly from the government standard. Women who met the new qualifications were provided with a short introductory course on teaching methods and in-service teacher training to make up for the lack of formal education. The treatment (CSP) schools were compared to a sample of schools not participating in the CSP program with similar demographics. Results indicated that both girls' and boys' enrollment increased significantly (Kim, Alderman, & Orazem, 1998) .

In India, 120 informal education centers were randomly assigned to a treatment or control group. In the treatment schools, each teacher was provided with a camera to take a photo of the teacher with students at the beginning and the end of the school day to track the exact time of the school's opening and closing. These teachers were then provided with a financial incentive to teach more days of the month. Teachers in the control schools were paid their same base salary and reminded that excessive absence could lead to dismissal. In addition to attendance information, children were also assessed academically pre- and post-intervention. Results showed a significant decrease in teacher absence in the treatment schools. In addition, student achievement at treatment schools was significantly higher than that at the control schools (Duflo,

Hanna, & Ryan, 2012). While these results are encouraging, the process of monitoring teachers daily can be expensive and arduous. And, increased teacher presence does not indicate increased teacher quality.

A program in Kenya provided a significant financial incentive to teachers for those teachers with the highest achieving classes as well as those with the most improved classes. Schools were randomly assigned to a treatment or comparison group, with teachers in the treatment group being provided incentives for achievement increases. Each year the program provided prizes valued at up to 43% of typical monthly salary to teachers in grades 4 to 8 based on the performance of the school as a whole on the Kenyan government's district-wide exams. Results of the comparison did show a significant increase in student achievement for the treatment schools, but no increase in teacher attendance or homework assigned, and no changes in pedagogy were reported. However, teachers in treatment schools did report more test preparation. While there were gains for treatment schools during the program's implementation, they were not sustained long term (Glewwe, Ilias, & Kremer, 2010).

Management

The fourth, and final, type of supply-side educational interventions is that of school-based management programs. These programs allow for authority and accountability at the school-level rather than the city, county, or country level. The authority includes the allocation of funds for school improvement through a local school management committee (Skoufias & Shapiro, 2006; Yamauchi & Liu, 2013) providing comparative information on student performance with nearby schools (Barrera-Orsorio & Linden, 2009), and allowing these committees to directly monitor teacher performance (Barrera-Orsorio & Linden, 2009; Duflo et al., 2012; Kim et al., 1998).

In Mexico, parents whose children's schools were managed within their community spent more time with their children helping with homework and, thus, improving achievement (Skoufias & Shapiro, 2006). And in a study in India, schools were randomly assigned to treatment and control groups where teachers in the treatment groups were monitored for attendance daily by students and local villagers, and were provided financial incentives based on their attendance. Results showed a decrease in absenteeism and a marginal increase in student achievement (Duflo et al., 2012).

Supply-Side Summary

Krishnaratne et al. (2013) note that the building of new structures, or the improvement of current facilities, has shown the strongest effect on enrollment, attendance, dropout rates, and student achievement than any other intervention used, supply- or demand-side. In general, providing additional materials or resources to teachers was found to have at least a short-term impact on all outcomes, including student achievement. However, teachers must be willing to incorporate said materials into their classroom and teaching. Incentives for teacher attendance have also shown increases in achievement. However, teacher incentives based on student achievement have not shown long-term success. School-based management programs in general have shown an improvement in student achievement, as well as a feeling of ownership of the schools by parents and the community at large.

Reduced Educational Costs

The first type of demand-side interventions attempts to reduce educational costs for a family is a popular intervention strategy, thought to increase learning outcomes via increasing

attendance by increasing the family's overall income having removed education costs. The success of these programs relies heavily on targeting the appropriate groups for aid, and that the aid is providing subsidy that is actually needed by the family (Krishnaratne et al., 2013). Educational costs may be reduced in a variety of ways, discussed below.

For example, conditional cash transfers (CCTs) are regular payments to a person or family, and are contingent on some behavior (i.e., attending school, school performance, etc.). In a review by Krishnaratne et al. (2013), results suggested that overall, transfers increase attendance and reduce dropout rates, but that the transfers must be large enough to offset costs to the family. In Columbia, a one-year pilot program was implemented to determine the most appropriate of three possible intervention programs: 1) a standard design incentivizing attendance with bi-monthly transfers ($n = 3,427$), 2) a modified design wherein the bi-monthly payments are slightly lowered with one-third of the total funding set aside to be provided at re-enrollment ($n = 3,424$), and 3) a design incentivizing graduation and college enrollment with similar bi-monthly payments as the second design, but a large lump sum payment provided at graduation, or sooner, should they move into higher education ($n = 1,133$). Randomization of treatment was at the individual level, and designs one and two were administered at all possible grades six through eleven, where the third design was administered only for grades nine through eleven.

Data for the study in Columbia came from several sources: demographic data from both the Colombian government and the program registration process, enrollment information from the government, self-report surveys, and the schools, attendance records from direct observation school records, and self-report surveys, and re-enrollment and graduation intentions from self-report surveys and school records. Data were analyzed using simple linear regression and results

indicated that the third design, incentivizing graduation, was the most effective of the three, followed by the second design. However, the third design group was the smallest of the three by a significant amount, and as most students drop out of school prior to grade 9, this may be a group of students who were on track for regular attendance and graduation regardless of the program. In addition, the authors note an unintended negative sibling effect which was strongest for girls wherein children in the control group who had a sibling in a treatment group attended less school than those with a sibling also untreated. That is, though the financial strain was lessened for one child, this did not result in re-direction of spending to a child who was not selected for treatment as was expected (Barrera-Osorio, Bertrand, Linden, & Perez-Calle, 2011).

In a medium-term outcome evaluation of the Oportunidades program in Mexico, Behrman, Parker, and Todd (2009) showed overall increases in the number of grades completed and on-time progression through grades by students in the treatment group, particularly those who began receiving treatment at ages 6 through 8 years. The Oportunidades program is a long-term initiative with many components, including the use of bi-monthly conditional cash transfers dependent on regular school and health clinic attendance. Authors investigated impact via two approaches: 1) differences between the two original treatment (immediate program benefits) and control (18 month delayed program benefits) groups, and 2) differences between the original treatment group and a newly selected control group with no program benefits. As noted, results were generally positive, with no significant differences between the original treatment and control groups due to the 18 month delay in program beginning. Most of the participants at time of evaluation were still in primary school, and as such, there is insufficient data to show consistent and compelling results. In addition, as the Oportunidades program is complex with several

facets and initiatives to which all participants were exposed, it is difficult to attribute changes to a single part of the program.

Progresa, another large-scale program initiated in Mexico, was evaluated in its early stages, and showed initial decreases in dropout rates, particularly in transitions between primary and secondary school. Like Oportunidades, Progresa is a large, many-faceted program, with a main focus of ensuring households have funds to have their children complete basic education. Cash transfer amounts are conditional on school attendance, and are larger for females, and increase with grade like the Columbia program, in an attempt to offset the possible increased household income associated with older children entering the workforce rather than completing their education. The initial program was designed as a randomized experiment with communities assigned to either a treatment or control group, and data were collected at three time points via large-scale surveys. As noted, results are generally positive, indicating that program participants show earlier ages of enrollment, lowered dropout rates, less repetition, and better progression through grades. Initial findings for PROGRESA (Jere R. Behrman, Sengupta, & Todd, 2005; Skoufias & Parker, 2001) educational outcomes suggested increases in enrollment and attendance, and decreases in child labor. However, this effect cannot be separated from other program inputs.

In 2002, Bangladesh introduced an experimental program, Challenging the Frontiers of Poverty Reduction: Targeting the Ultra Poor (CFPR) in an attempt to provide income opportunities for poor families that were sustainable over time, and managed by the program participants themselves (Ahmed, Rabbani, Sulaiman, & Das, 2009). The CFPR program emphasized a more involved targeting process, entrepreneurship development, and the creation of

an environment with social supports to enable a path out of poverty. As with the *Oportunidades* program, CFPR included many different initiatives, including a weekly stipend to be invested into their enterprise of choice, and continue until said enterprise begins to supply an income.

Individuals were selected for the program by way of ranking districts from poorest to richest and the poorest 10% of households were chosen. Results indicated no significant effect on school enrollment rates between treatment and control groups. However, the duration of the program at time of evaluation was two years, possibly too short a time span to show real gains.

In Ecuador, researchers evaluated the efficacy of a cash transfer program on school enrollment and child labor in Ecuador. The *Bono de Desarrollo Humano* program was initiated in 2004, with no explicit conditions on monthly cash transfers made only to women. The evaluation was carried out in a randomized fashion using a lottery system with a baseline survey administered prior to program initiation and a follow-up survey approximately 18 months later. Results of the evaluation indicated a large increase in enrollment rates and a large decrease in child work. While there was no enforced condition of the cash transfers, the government did television spots stressing the role of parents in ensuring school enrollment and good health status of their children, which may have been taken as conditions of the transfer and helped to increase the effects seen. As such, with time, parents may come to see that this is not a monitored condition and enrollment rates may be seen to decrease (Macours, Schady, & Vakis, 2012; Schady & Araujo, 2006).

Another way in which development organizations have reduced educational costs is through the use of vouchers. Voucher programs may be restricted (awarded by lottery, with renewal contingent on progress) or unrestricted (available to all). These types of programs are

thought to increase enrollment through reduction of family costs, and learning outcomes through the possible availability of better quality schools previously unaffordable to some families (Krishnaratne et al., 2013). In Columbia, the program Programa de Ampliacion de Cobertura de la Educacion Decundaria (PACES) was initiated which used a lottery system to distribute vouchers to cover partial costs of private secondary schools for students maintaining adequate progress. Initial results showed that, of the applicants for PACES, voucher recipients and non-recipients were both as likely to be enrolled in school, with non-recipients more likely to be enrolled in public rather than private institutions. In addition, recipients completed slightly more education, and were more likely to complete grade 8 than non-recipients. Achievement test results indicated that voucher recipients and girls in particular, had higher scores than non-recipients (Angrist, Bettinger, Bloom, King, & Kremer, 2002). Longer-term results assessed seven years after program implementation showed higher graduation rates and test scores for the voucher recipients than non-recipients (Bettinger, Kremer, & Saavedra, 2009).

In Chile, however, a voucher program in place since 1981 has provided controversial results. Chile's government instituted said voucher program such that students wishing to enroll in private schools would be provided vouchers to cover the tuition, resulting in a mass exodus of students from public institutions. However, results found by some show that test scores are not greater, grade repetition is not decreased, and student progress is not better in communities where the number of private institutions has increased due to the interest in private schools by residents. Other results show quite the opposite, citing positive impacts on test scores and college entrance exams (Contreras, Sepúlveda, & Bustos, 2010; Gallego, 2006; Sapelli & Vial, 2005).

Finally, development organizations have also investigated the effects of school fee reduction interventions. School fee reduction interventions may include the provision of free education or of meeting some, but not all, educational costs such as school uniforms or textbooks but not fees (or vice versa). In Uganda, universal primary education (UPE) was introduced to all primary grades in 1997, resulting in a nearly 60% increase in enrollment. Grogan (2009) used census survey data collected three years after program implementation to show that while the age at enrollment decreased due to the program, there were several negative consequences of such an increase in students to schools. These consequences included textbook and teacher shortages as well as severe classroom overcrowding which resulted in some districts requiring several school “shifts” throughout the day or week to keep up with the demand resulting in a negative effect on retention. Investment in infrastructure over time, however, appears to have improved the resources available and allowed the retention rates to stabilize (Bategeka & Okurut, 2005; Deininger, 2003).

The results of a randomized evaluation in rural primary schools in Kenya (Kremer, Miguel, & Thornton, 2009) showed that the use of a merit-based scholarship program increased student achievement and attendance for both girls and boys, and increased attendance for teachers. The Girls Scholarship Program was implemented in two rural districts in Kenya, randomized at the administrative division level (with eight divisions in each district) with half of the primary schools in each division receiving the treatment. Girls who excelled on their exams in grade 6 were awarded with a scholarship covering her school fees and supplies for the following two years, as well as public recognition of her award. Data were collected regarding student and teacher attendance, achievement scores, school supply purchases, time use by students, and

student attitudes. Results showed large achievement increases for the recipients of the scholarships, as well as for girls and boys ineligible for the award.

Another initiative in Kenya which provided school uniforms to students in poor communities by way of a lottery was evaluated by Evans, Kremer, and Ngatia (2008). Results showed a drop in absenteeism of nearly half, and significant increases in average test scores in communities where the program was implemented. However, there were several additional benefits to schools which were chosen to be a part of the initiative, which could be responsible for some of the positive outcomes. Schools were visited several times a year by a nurse and provided care to any local child or adult who requested it, an agricultural representative visited schools and organized students to grow crops on school grounds, and during one year of the uniform initiative, schools received a large grant for classroom construction and supplies. Thus, any combination of these four inputs may be responsible for the positive outcomes.

To summarize, there are many ways in which organizations may attempt to reduce educational costs for students and their families, and each of these has shown differential effectiveness. Conditional cash transfer programs (CCTs) have shown that conditions on the transfers, even if they were simply implied or unmonitored, were more effective in increasing attendance. To date, however, there is no evidence of an impact of CCTs on learning outcomes. Overall, voucher programs have shown no significant effect on enrollment, dropout rates, or learning outcomes. The opposing results for the Chilean program show some evidence as to how difficult it can be to evaluate programs such as these with such complex inputs and wide-reaching outcomes. Overall, it can be shown that the effects of reducing educational costs on attendance and dropout rates are positive, though somewhat dependent on context. However, the link

between these types of programs and positive learning outcomes is tenuous and requires further investigation.

Increasing Preparedness

The second demand-type intervention strategy involves increasing family and student preparedness to attend school reliably and succeed in the classroom. Types of interventions include early childhood development (ECD) programs, school feeding programs, and health-based programs. The focus of these interventions is, for the most part, situated outside the classroom.

ECD programs are designed to enhance a child's cognitive and social skills to increase school preparedness. These types of programs include: building and equipping preschool classrooms and providing trained teachers, at home daycare programs, and increasing parental engagement. In Uruguay, the government built new or refurbished pre-school classrooms and increased the number of teachers significantly. Enrollment rates increased by 76% over 9 years, with rates for the poorest families increasing by 60%. In addition, attendance rates of treated children were significantly higher, as was the average number of years attended, even within a family. The differences between the treated and untreated groups only increased as time passed. However, as with most of the initiatives discussed in this chapter, there were several other programs implemented across Uruguay, and as such, discussing only the results of a single program is difficult (Berlinski, Galiani, & Manacorda, 2008).

In Turkey, the Turkish Early Enrichment Project was a four year study designed to investigate the effects of an educational preschool environment and a mother training program

that included a cognitive program and a discussion group that covered topics such as nutrition, discipline, child health, etc., on children's cognitive performance and school success. Children were assessed using several cognitive assessments, as well as personality and social development measures. Mothers were observed interacting with their children, as well as interviewed regarding a host of demographic characteristics. Initial results indicated that children of mothers exposed to training trended toward higher IQs, and school grades, as well as showed decreased aggression. Trained mothers also reported higher educational expectations for their children, greater availability for homework help, and a greater amount of interaction with their children in the home. Six years after the program ended, a full 10 years after initial implementation, another set of measures was administered, and fathers were also interviewed. At this time, a larger proportion of children whose mothers were trained remained in school, but no significant achievement differences remained past primary school. Parents of the children from the treated mothers group reported higher educational expectations, and fathers perceived these children as more motivated than their counterparts. In sum, the clear success of the program was that of educational retention, which may be attained by more cost-effective means. The marginal cognitive advantage that children showed at the four year mark was not present in terms of educational achievement after 10 years (Kagitcibasi, Sunar, & Bekman, 2001).

In rural Mozambique, a center-based community driven preschool model was implemented in 30 of 76 total communities. Surveys collecting demographic information were carried out for a baseline as well as a battery of cognitive, motor, language, socio-emotional development, and health assessments were administered to children in both the intervention and control communities, and endline measures were administered after two years. Results indicated that in treatment communities there was an increase in preschool enrollment, an increase in

school enrollment at the appropriate age, an increase on time spent on school activities, as well as improved cognitive, motor, and socio-emotional characteristics. In addition, there were positive sibling effects in the treatment communities such that enrollment of older sibling increased in said communities. While initial results are encouraging, it should be noted that the impacts on language development and health were marginal, and as there was effort put into infrastructure as well, the results cannot be specifically tied to the implementation of the program alone (Martinez, Naudeau, & Pereira, 2013).

Another way in which development organizations may increase preparedness is through school feeding programs which either provide meals for children at school or with food parcels to take home to be shared with the family. Providing meals to students in school has been commonly used across the world to increase attendance and enrollment rates (Krishnaratne et al., 2013). In Kenya, 50 schools were randomly sorted into a treatment and control group of 25 each, and treatment schools were provided breakfast to be served to preschool classes only over a period of two years. Prior to program implementation, a baseline survey and school attendance checks were completed. Each year, attendance checks were completed, and in the third year, cognitive assessments were completed, anthropomorphic measurements were taken, and endline surveys were administered. Results indicated that the program improved test scores, but only in those schools where the teacher was motivated and experienced prior to implementation. The program in this case, though designed to do otherwise, took a large amount of classroom time from the teacher, which may explain the results (Vermeersch & Kremer, 2005).

In Burkina Faso, two school feeding programs were evaluated for impact on student health and educational achievement. One feeding program includes providing students with lunch

each school day, and the other provides girls with take home rations every month, conditional on 90% attendance. Results after one year indicate a small enrollment increase for girls, and a slight improvement in math scores, also for girls. While there were some enrollment increases, results showed that for those households with a large number of children (generally agricultural households), there were no improvements in achievement or attendance. That is, when children were needed at home to work, the incentive was not enough (Kazianga, De Walque, & Alderman, 2009).

The Food for Education program was introduced in Bangladesh in 1993, and its main feature provided a monthly food ration to families judged as poor with at least one child attending primary school that month. The ration amount can increase to a maximum by sending more than one child to school, and the children must maintain an 85% attendance rating to receive it. Recipients are selected through a lengthy process moving from district to household selection where randomization occurs. Results indicated large increases in both attendance rates as well as duration of schooling (Meng & Ryan, 2010).

The third way in which organizations may attempt to improve preparedness is through a variety of health-based interventions which may include prevention, treatment, provision of meals, first aid kits, or even counselling. Health problems of children in the developing world are highly related to their ability to get to or from school, as well as their ability to actively participate in their own learning (Krishnaratne et al., 2013). Students in grades four through six in an elementary school in Indonesia were randomly divided into a treatment and control group. Students in the treatment group were provided fish oil supplements for three months, and questionnaires regarding aggressiveness and impulsiveness were administered pre- and post-

study. Attendance was also closely monitored and blood was drawn pre- and post- study. Results indicated that students provided fish oil supplements had better attendance in school, but no other differences between the treatment and control groups were found (Hamazaki et al., 2008).

In Sri Lanka, a double-blind study was carried out over nine months to investigate the impact of malaria and its prevention on educational attainment. Children attending grades one through five in four different schools were randomly assigned to treatment and control groups. Language and mathematics test scores were used to show achievement, and attendance records were monitored. Results indicated that children who received the anti-malaria medication scored higher in both mathematics and language, and showed significantly lower absenteeism rates. In addition, during the intervention, the incidence of malaria decreased by over half (Fernando, De Silva, Carter, Mendis, & Wickremasinghe, 2006).

Educational training for treatment and management of asthma and epilepsy in Argentina was shown to have significant effect on attendance. The program included five weeks of meetings of 8 – 10 families with parents and children's groups held separately and activities include games, drawings, stories, videos, and role-playing. Children were shown how to manage their own health, and parents were shown how to facilitate this management. Interviews were carried out before the program, six months after the end of the program, and one year after the end of the program. Results indicated that parents and children in the treatment group had significant improvements in knowledge, beliefs, attitudes, and behaviors related to their respective illness. In addition, children in the treatment group had fewer health crises, visits to the doctor, and visits to the emergency room than those in the control group. However, sample sizes were small, with only a total of 202 participants (Tieffenberg, Wood, Alonso, Tossutti, & Vicente, 2000).

Increasing preparedness for families and children to enter school may be done in many ways. In general, Early Childhood Development programs have shown significant positive effects on children's school achievement, if the program implemented is of good quality. However, the results of these programs have not been shown to have long-term positive educational outcomes, and depend heavily on experienced and reliable teachers. School feeding programs have shown a positive effect on attendance and enrolment, but gains in achievement depend heavily on high-quality teaching, indicating that the feeding may not be the relevant factor. Results of health-based interventions vary heavily in terms of their significance. It appears that some programs (malaria treatment) are more effective in that they are targeted to an illness with severe cognitive impairment, causing more problems with learning than some other less serious illnesses.

Providing Information

The third, and final, demand-type intervention strategy is providing students and families with information regarding educational quality or the economic benefits of higher education. This type of intervention is thought to affect change by way of empowering students and parents to make evidence-based decisions about education. These interventions have not been shown to have any significant effect on learning outcomes, enrollment, attendance, or dropout rates (Krishnaratne et al., 2013). In Madagascar, schools were randomly assigned to one of three interventions: 1) where teachers inform parents and children of the average projected earnings at each level of education, 2) a role model shared with parents and children their family background, educational experience, and current achievements, and 3) a combination of both 1) and 2). Surveys were completed after parents and children were exposed to their intervention that included information about perceived educational returns, student attendance, and student

achievement. Results showed that children of those parents who had under-estimated the relationship increased their attendance, but the opposite effect was found for the children of those parents who over-estimated the relationship. Families who spoke to role models who were also poor, showed larger increases in achievement, and those families who received the third intervention showed the smallest changes overall (Nguyen, 2008). In general, the impact of these types of interventions is small, and there is little information regarding the circumstances under which there is significant gain.

Summary

In summary, there are five outcomes to consider when evaluating an educational initiative: enrollment, attendance, progression, repetition, dropout rates, and student achievement in the form of test scores. The most promising interventions discussed for increasing enrollment include the creation or improvement of school buildings, early childhood development programs, and school feeding programs. In terms of attendance, those interventions that are most effective include conditional cash transfers, health-related interventions, school feeding, and providing teachers with resources. Promising programs that have shown effective rates of student progress are conditional cash transfers and school-based management programs, and the most effective programs for showing decreases in dropout rates are providing teachers with more resources, including professional development and additional help in the classroom. Finally, proven programs that have shown increases in student achievement are those that include additional resources to teachers (especially computers), additional teachers or classroom help, school feeding, and school-based management programs. However, the effects of all of these programs have only been seen in student achievement in math.

Challenges Surrounding Translation and Adaptation of Measures

To date, international evaluation work has focused little on the development and adaptation of valid quantitative outcome measures. The field of educational measurement has investigated the issues surrounding development and adaptation for decades, culminating in a set of standards around translation and adaptation of measures in the Standards of Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) and more specifically, a second edition of The International Test Commission (ITC) Guidelines for Translating and Adapting Tests (International Test Commission, 2017). Both of these handbooks provide useful frameworks to inform this type of evaluation work.

One of the most studied issues in cross-cultural or cross-national research that is directly applicable to international development evaluation is that of translation or adaptation of data collection instruments (Beaton, Bombardier, Guillemin, & Ferraz, 2000; Hambleton, Merenda, & Spielberger, 2004; J. A. Harkness, Villar, & Edwards, 2010; van de Vijver & Matsumoto, 2011). Recommended methods for adaptation include expert judge review or committee-translation (Carlson, 2000; McGorry, 2000), and a lengthy back-translation process (Harkness, 1999) to be completed for each language in which the instrument will be administered. However, as noted previously, this process may not always be adhered to in the fast-paced and under-funded context of international development evaluation. Quality adaptation of a measure to be used across languages, regions, or cultures is an integral step in establishing measurement invariance, or equivalence, so evaluators can be sure that the scores on a measure in one context (i.e., language,

nation, or culture) have the same meaning (Byrne & van de Vijver, 2010; Kline, 2015; Milfont & Fischer, 2010).

Though it is not always feasible to carry out a quality adaptation of a measure before collecting data, there are many ways in which a researcher can determine the quality of the outcome of the efforts. The most common analytic procedure recommended to establish measurement invariance is an analysis of the covariance structure of the data (i.e., factor analysis (FA) or structural equation modeling (SEM) (Byrne & van de Vijver, 2010; Kline, 2015; Milfont & Fischer, 2010; Schaffer & Riordan, 2003; Sharma & Weathers, 2003). Other methods can include the use of Modern or Classical Test Theory (e.g. Differential Item Functioning, item analyses) which allow both statistical and graphical inspection of parameter invariance across populations (Maydeu-Olivares, Morera, & D’Zurilla, 1998).

When using an approach that analyzes the covariance structure of the data, results can provide us with evidence of two particular types of evidence of invariance (Kline, 2015). Equal form (or configural) invariance is the most basic wherein we can say that the basic structure of the data is the same across groups but not the weighting of the particular items. Metric invariance, however, allows us to say that the actual item factor loadings (i.e., the relative weighting of each item in the factor model) are equal across groups. Partial metric invariance allows for some item loadings to vary in the model, whereas construct-level measurement invariance requires all loadings to be fixed across groups as equal.

While the results of analyses of covariance structure can provide the researcher with a particular kind of information regarding invariance, researchers must move beyond basic structural analysis to gain insight into the possible causes of invariance present. In addition, the

use of SEM and FA require large sample sizes which are not always feasible, particularly when using pilot data to recommend changes to a measure for full implementation. In response to the lack of more specific diagnostic information regarding the possible sources of invariance in the results of FA or SEM analyses, GTheory (Cronbach, Rajaratnam, & Gleser, 1963) and the Many-Facet (MF) model (Engelhard & Wind, 2018; Linacre, 1989) may be used to further shed light on the issue. GTheory and the MF model allow the partitioning out of multiple sources of error variance in a single analysis (Linacre, 1989; Shavelson & Webb, 1991).

In a review of impact evaluation reports from the Poverty Action Lab (<http://www.povertyactionlab.org/evaluations>), Innovations for Poverty Action (<http://www.poverty-action.org/work/publications>), DFID (<http://r4d.dfid.gov.uk/>), and the International Initiative for Impact Evaluation, or 3ie (<http://www.3ieimpact.org/en/evidence/impact-evaluations/>), none discussed the adaptation or translation of quantitative measures used, nor the validation of said measures. It is unclear at this juncture which, if any, standardization measures are being routinely undertaken, and how this might affect evaluation findings. Given the apparent lack of standardization of quantitative measures, and the increased importance of said measures as required by impact evaluation and the use of pay-for-performance models, evaluators require a more complete picture of the possible effects on reliability and validity of measures used.

Validity and Reliability

Validity Evidence

The Standards for Educational and Psychological Testing defines validity as “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (American Educational Research Association et al., 2014). Five types of validity

evidence are outlined in the Standards and are useful in evaluating the proposed use of a particular measure: 1) evidence based on test content is obtained from an analysis of test content and the intended construct to be measured. 2) evidence based on response processes is generally obtained from an analysis of individual responses to show the fit between the intended construct to be measured and the nature of the performance or response of the individual interacting with the item or activity, 3) evidence based on internal structure is obtained through an analysis of the relationships among test items and components and comparison to the construct being measured, 4) evidence based on relations to other variables is obtained through the analysis of test items or components to other variables known to correlate with the construct of interest, or, conversely, known not to correlate with the construct of interest, 5) evidence based on consequences of testing is the most complex and difficult evidence to obtain, requiring the consideration of intended and possible unintended consequences of testing. Much of this evidence is collected over large spans of time, and particularly when a measure is being used in a novel way, to ensure that the use is sanctioned.

Reliability

Reliability is broadly defined as “the desired consistency (or reproducibility) of scores” (Crocker & Algina, 2008), and depends heavily “on characteristics of the test, the conditions of administration, and the group of examinees” (Traub & Rowley, 1991). Reliability may be assessed in many ways, depending on the type of the assessment and whether we want to compare individuals to one another (norm-referenced assessment) or to some external criterion or cut score (criterion-referenced assessment). In general, we want to ensure that an individual’s score on a particular assessment is consistent across administrations.

In the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014), eight standards, or aspects of reliability evidence are outlined (shown in Figure 5). As the basis of reliability is in the consistency of a score, there are basic requirements around the replication of an assessment in an independent administration such that the construct of interest would not be expected to change from administration to administration. This may involve parallel (or alternate) forms which requires two consistently designed forms of an assessment be administered at the same time or at different times. The decisions made around replication will directly affect the way in which the reliability coefficient is calculated, which are the second and third aspects of reliability. There are three general types of reliability coefficients: parallel (or alternate) forms as noted above, test-re-test reliability which requires an assessment to be administered to the same individuals with a short time span between administrations, and internal consistency measures which require only one version of an assessment be administered to a group of individuals, and results in a lower bound estimate of reliability.

The fourth standard of reliability involves an examination of possible factors that may affect the reliability coefficient or the precision of measurement. These factors include the administration procedure, the use of raters in assessment, and differences in intended vs. assessed populations of interest. Fifth, errors of measurement should be calculated around the resulting test scores allowing for confidence bands to be created for a fuller picture of the precision of the measurement. As with reliability coefficients themselves, the ways in which standard errors of measurement are calculated, interpreted, and communicated depends on the way in which the replication was designed as well as on the score interpretation being either norm- referenced (i.e., intended to allow for comparisons between test takers), or criterion -referenced (i.e., test takers are compared to a criteria, such as a pass score).

Decision consistency, the sixth consideration in reliability, is particularly relevant when test takers are to be classified based on their assessment score. In these situations, there is specific interest in reliability of measurement at the cut score(s), resulting in a particular evaluation of the conditional standard errors at and around these scores. When the interest is in the reliability of mean scores of groups of individuals, the seventh consideration comes into play, the reliability and precision of group means. This may be a consideration in evaluations of program effectiveness or educational accountability systems in evaluating the effectiveness of some intervention or other factor. In these cases, the investigation should focus heavily on possible variation due to sampling errors, and ensuring that the sample size is sufficient and representative. Finally, the last consideration is in the documentation of reliability or precision coefficients and research. Test developers will often have a test manual for a commercial assessment that can be referenced, and the source documentation may provide the level of information needed for an individual to have confidence in the assessment's use in their specific circumstances. However, it is also important for test users to document their circumstances and test use, and their own investigations into the reliability of an assessment for use with their population of interest.

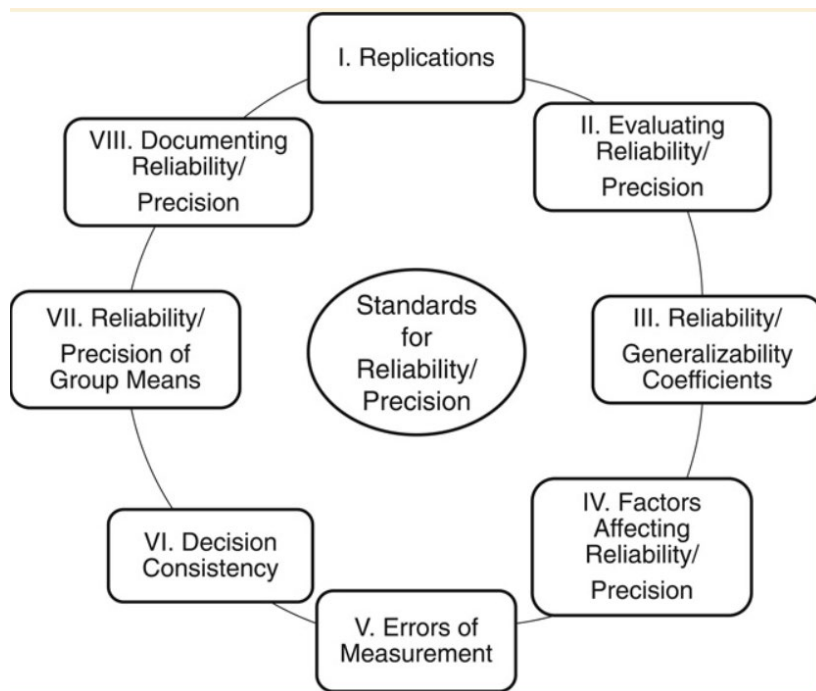


Figure 5. Reliability standards from Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014)

We can see then, that reliability of outcome measures is an integral component of evaluation, and requires in-depth investigations and consideration. The less reliable the outcome measure(s) being used in an evaluation, the more likely program decisions will be incorrect. This could result in successful programs being penalized, or unsuccessful programs being continued, with serious financial ramifications.

Using Generalizability Theory to Establish Scale Reliability

One of the ways in which reliability may be assessed is through the use of Generalizability Theory (Lee J. Cronbach et al., 1963; Gleser, Cronbach, & Rajaratnam, 1965). In a Generalizability study (GStudy), the researcher can identify and select pertinent facets (factors

that may be sources of error variance) and look at them in relation to one another, allowing us to attribute smaller or larger sources of variability to a particular facet. For example, in the case of cross-cultural research, these possible facets may include country or region, language of instrument, first language of the participant, and enumerator or rater. The lower the error variance in the data, the higher the quality, or reliability (Bayerl & Paul, 2007). By parsing apart the sources of variability, we can see the impact of a particular facet on the quality of the data, and make requisite changes in order to improve it, making the results particularly useful in pilot or longitudinal studies wherein there is a possibility of adapting or editing the instrument.

In a study of measurement invariance (a statistical property indicating that the same construct is being measured over some specified variable) of an empirical measure used in manufacturing, both FA and GTheory were used to fully assess invariance across three industries in an attempt to provide evidence for use of the scale in benchmarking one industry against another (Malhotra & Sharma, 2008). The measure included six flexibility scales and a total of 104 items; 147 responses were used in the analysis. Results showed the benefit of using both FA and GTheory to fully assess the invariance of the scale across three groups. However, the authors recommend the use of GTheory alone in cases of small sample size as FA methods require large samples (Kline (2015) recommends 20 cases to each item analyzed).

Sharma and Weathers (2003) also used both GTheory and factor analysis to assess the measurement invariance of a self-report scale of consumer ethnocentrism across four countries. The scale included 17 items and three GTheory analyses were completed: one with an artificially balanced design of 70 cases per group (280 total cases), and second using the full data set with slightly unequal sample sizes across the groups (71, 70, 76, and 73; 290 total cases), and a third

using unbalanced designs with all combinations of sample sizes of 40, 50, 60, and 70 cases per group. The authors found similar results across all three sets of analyses and concluded that the procedure was robust against the effects of even extreme unequal sample sizes.

Recommendations from Sharma and Weathers (2003) include the use of factor analysis where reasonable sample sizes are possible due to the advantage of statistical criteria and the ability to assess problematic items in cases where measurement equivalence has not been validated, that is, in early stages of scale development. However, the authors note that GTheory provides valuable insight regarding the sources of variability (i.e., across countries or items) such that researchers can further investigate the phenomena for cause, as well as providing the researcher with the number of levels of a facet or the number of subjects required for a particular desired generalizability. In sum, the authors suggest that, where possible, the two analyses should be used in concert to assess measurement invariance.

Solano-Flores and Li (2006) used GTheory in a study investigating the error variance associated with testing linguistic minorities. The authors verbally administered a set of 12 open-ended mathematics items selected from the National Assessment of Educational Progress to a group of 170 grade 4 and 5 English learners with a common first language, Haitian-Creole. Items were translated from English into three Haitian-Creole dialects, and then back-translated into English. Using only GTheory to analyze the data, the authors were able to isolate dialect as a significant source of measurement error, contributing to differences in achievement across examinees.

Durvasula, Netemeyer, Andrews, and Lysonski (2006) used both an empirical and simulated cross-national data set in order to investigate the appropriateness of GTheory in

assessing measurement invariance across countries. A measure of advertising attitudes consisting of three scales and a total of 10 items was administered in English to four groups, and translated (and back-translated) into Greek for administration to one group with sample sizes across groups ranging from 87 to 179. A GStudy allowed the authors to note that the majority of the variability was within-country and across-subjects, not across countries. This result provided insight into the previous research using FA that showed simply that the measure was invariant across countries, but not why, which is imperative in attempting to correct the problem.

Dzhambov and Dimitrova (2014) used GTheory to develop a shorter version of a noise sensitivity assessment while maintaining adequate reliability of the measure. The measure, the Noise Sensitivity Questionnaire was developed to quantify noise sensitivity as related to different daily situations. The scale consists of 35 items across 5 subscales, and was developed in English. The authors used a back-translation method to translate the scale into Bulgarian, and a short form of 15 items was administered. GStudy results indicated that the shorter form of the survey showed adequate reliability and predictive validity.

Cor and Peeters (2015) used GTheory in the development of a new assessment program in Pharmacy by varying the number of items and testing occasions. The authors use a specific exam as a case study to show how GTheory can be used to achieve desired reliability while also meeting content specifications. Authors recommend the use of GTheory throughout the test development process, particularly when using rater-scoring.

Oh, Osgood, and Smith (2015) used GTheory to study the extent to which the Caregiver Interaction Scale, and the Promising Practices Rating Scales could serve as reliable and valid measures of the quality of an afterschool program. The authors noted the added benefit of using

GTheory in providing evidence that such scales are sensitive to day-to-day fluctuations which would not necessarily be noted if other psychometric analyses were used.

Kang, Bjornson, Barreira, Ragan, and Song (2014) used GTheory to investigate the minimum number of days needed to establish reliable physical activity estimates in children. A GStudy was used to initially quantify the proportions of error variance attributable to all facets in the study, and followed up with a decision study in order to estimate the minimum amount of data needed in order to achieve adequate reliability of the measure. And Gadbury-Amyot, McCracken, Woldt, and Brennan (2014) used GTheory to validate a new assessment in a dental school involving the use of portfolios. The authors provide suggestions regarding the type of scoring and number of raters needed to ensure reliable scores.

GTheory provides a straightforward and understandable framework in which to investigate possible sources of error that contribute to lower reliability estimates. The method also provides information useful in the modification of experimental designs that allow for maximization of reliability. The GStudies and DStudies outlined in the above section show the utility of the method in providing evidence of measurement invariance in cross-cultural research as well as in assessment development and validation.

Using the Rasch Measurement Theory to Establish Scale Reliability

One other way in which reliability may be assessed is through the use of Rasch Measurement Theory (RMT; Rasch, 1980). Similar to GTheory, the researcher can identify and select pertinent facets (factors that may be sources of error variance) and look at them in relation to one another using the Many-Facet (MF) model (Linacre, 1989). One of the major benefits of

the RMT tradition is the visual representation of the outcome data on a continuum called a Wright map (Wright & Masters, 1982).

Using a university-wide student evaluation of teaching survey, Van Zile-Tamsen (2017) moved through the process of using RMT in the form of the Rating Scale Model (RSM) and another Rasch model, the Partial Credit Model (PCM; Wright & Masters, 1982) to assess the psychometric properties of the scale in terms of reliability, validity, and item difficulty. Results indicated that the RSM provided strong diagnostic indicators at the item level, useful in assessing if each item is functioning optimally for precision of measurement of the construct in question. This information allows for the scale designer to make decisions at the item level about changes to increase the precision of the measure (i.e., reliability).

Ölmez and Ölmez (2019) used the RSM to provide validity evidence for the use of a math anxiety scale with undergraduate students. Results indicated that the scale included several items that did not fit the model well, requiring either deletion or revision, and a lack of items allowing for differentiation of low or very high levels of math anxiety. This level of detailed analysis allowed researchers to further revise the scale for more widespread use.

In another validation study, Tabatabaee-Yazdi, Motallebzadeh, Ashraf, and Baghaei (2018) developed a teacher success questionnaire that would provide students' perspectives on what makes a Persian language teacher successful. The questionnaire was administered to a sample of students and the data were analyzed using the RSM to examine the psychometric qualities of the scale in terms of dimensionality, use of response category, sample appropriateness, and reliability. Results indicated a set of items should be flagged for revision or

deletion due to a poor fit within the model, and reasonable reliability supporting future use in the specified setting.

In another language study, ParahitaAnandi and Zailaini (2019) assessed the quality of a self-assessment speaking rubric originally developed in English to be used with English language learners. The rubric was translated and modified by the authors for use with Indonesian students learning Arabic as a foreign language. Data from a small sample of students was collected and the authors completed a rating scale analysis allowing for the review of summary statistics, item fit, principal component analysis, and Wright map. Findings indicated the scale was appropriate in length, all items provided good fit to the model, and the scale showed adequate reliability with the sample allowing authors to conclude that the translated rubric for use with Indonesian students of Arabic was reasonable.

In a study by Randall and Engelhard (2010), both Confirmatory Factor Analysis (CFA) and RMT were used to investigate the psychometric properties and multigroup measurement invariance of scores across subgroups, items, and persons on the Reading for Meaning items from the Georgia Criterion Referenced Competency Test (CRCT). Authors sought to determine measurement invariance across both accommodation provided and disability status for a high-stakes state assessment. Results of the CFA showed evidence of invariance across disability status, but not accommodation type, and the results of the item-level Rasch analysis showed similar results. The authors outlined the differential, but meaningful distinctions across the two analyses, and the importance of both in providing a full picture of the measurement invariance of the assessment.

Studies Using Both Rasch Measurement Theory and Generalizability Theory to Establish Scale Reliability

In their book on rater-mediated assessments, Engelhard and Wind (2018) outline the distinct differences in approaches of GTheory and RMT. Rater-mediated assessments are those assessments where a rater assigns a score to an individual's responses. The authors outline the theory and underpinnings of each method of analysis and move through a comparative analysis using the same dataset and outlining the advantages and disadvantages of using GTheory and the MF model to inform the assessment of rating scales in practice. The authors show the utility of both analytic procedures in determining scale quality and measurement invariance and conclude with the guidance that each of these approaches both comes from a different view of the assessment process, therefore the interpretation of results around psychometric quality, and the assessment system itself will differ.

In a study by Iramaneerat, Yudkowsky, Myford, and Downing (2008), authors used both GTheory and RMT to analyze data from an Objective Structured Clinical Examination (OSCE) as a means of approaching quality control in the assessment. Like many rater-mediated assessments, OSCE scores have several potential sources of measurement error. Authors examined a communication scale with 18 5-point items and 79 candidates and found that GTheory results provided guidance on the largest sources of error, and the MFM analyses provided a more detailed, individual-level analysis of rating consistency.

Sudweeks, Reeve, and Bradshaw (2004) used GTheory and RMT to analyze the results of essay scores of 24 undergraduate's scores on two 3-page essays (48 total essays) with 9 raters, in order to estimate potential sources of error, obtain reliability estimates, and make

recommendations for improving the essay rating process. Their conclusions agree with (Iramaneerat et al., 2008) in that the GTheory findings provide recommendations for group-level changes and the RMT results allow for a more individualized approach to making modifications to the elements in the model. Authors recommend both analyses as complementary and not adversarial approaches to this type of work.

Finally, Lynch and McNamara (1998) used GTheory and RMT in the development of a performance-based second language assessment procedure. Authors analyzed data from an English as a Second Language (ESL) assessment of speaking skills of 83 candidates including 23 items and four raters. Conclusions of the use of these two analytic procedures are consistent with Iramaneerat et al. (2008) and Sudweeks et al. (2004).

Summary

As has been noted, to date, international evaluation work has focused very little on the development and adaptation of valid quantitative outcome measures. However, the field of educational measurement has investigated the issues surrounding development and adaptation for decades, culminating in both a set of standards around translation and adaptation of measures in the Standards of Educational and Psychological Testing (American Educational Research Association et al., 2014) and more specifically, a second edition of The International Test Commission (ITC) Guidelines for Translating and Adapting Tests (International Test Commission, 2017).

Though it is not always feasible to carry out a quality adaptation of a measure before collecting data, especially given the budgets and timelines in international development evaluation projects, there are many ways in which a researcher can determine the quality of the

outcome of the efforts possible. Two of these methods, Generalizability Theory (GTheory; L. J. Cronbach, 1963), and Rasch Measurement Theory (RMT; Rasch, 1980) – more specifically, the use of the Many-Facet Model (MFM; Linacre, 1989) – allow for a complementary analysis of rater-mediated assessments including any number of facets of interest, including language.

For example, Solano-Flores and Li (2006) were able to use GTheory to isolate dialect as a significant source of error in an assessment of linguistic minorities. And, Durvasula, Netemeyer, Andrews, and Lysonski (2006) used GTheory to find that the majority of the variability for their assessment was *within*-country and across-subjects, not *across* countries. In a study using RMT, ParahitaAnandi and Zailaini (2019) assessed the quality of a self-assessment speaking rubric originally developed in English to be used with English language learners. Results allowed authors to conclude that the translated rubric for use with Indonesian students of Arabic was reasonable.

While there is evidence to support the use of GTheory and the MFM as complementary sources of evidence in terms of assessment development and validation (Engelhard & Wind, 2018; Iramaneerat et al., 2008; Smith & Kulikowich, 2004; Sudweeks et al., 2004), the methods have yet to be used either in the highly complex contexts such as in development contexts with many possible sources of error, or with the types of subjective (i.e., surveys) and objective (i.e., math or literacy assessments) measures.

CHAPTER III. METHODS

The purpose of this study is to explore the use of Generalizability Theory (GTheory; Brennan, 1992, 2001; Shavelson & Webb, 1991) and Rasch Measurement Theory (Rasch, 1980; Wright & Masters, 1982) in the form of the Many-Facet (MF) model (Linacre, 1989) assessing possible sources of unreliability in data from an international evaluation to be used as evidence of success in outcomes of an educational initiative. In both a Generalizability study (GStudy) and a Many-Facet (MF) analysis, the researcher can identify and select pertinent facets (factors that may be sources of variance) and look at them in relation to one another, allowing us to attribute smaller or larger sources of variability to a particular facet. For example, in the case of cross-cultural research, these possible facets may include country or region, language of instrument, first language of the participant, and enumerator or rater. The lower the error variance in the data, the higher the quality, or reliability (Bayerl & Paul, 2007). By parsing apart the sources of variability, we can see the impact of a particular facet on the quality of the data, and make requisite changes in order to improve it, making the results particularly useful in pilot or longitudinal studies wherein there is a possibility of adapting or editing the instrument.

Therefore, the primary research question guiding this research is: How can Generalizability Theory and the Many-Facet model be used to assess the reliability of cognitive and non-cognitive outcome measures used in an international development education evaluation? The current study will use GTheory and the MF model to analyze data from an international development education evaluation particularly when coupled with inadequate adaptation of non-cognitive measures. Two types of measures will be assessed: a set of subjective, or affective, survey items, and an objective achievement measure of reading. Conducting analyses on both

types of measures will allow a more comprehensive discussion on the usefulness of the two analytic methods in evaluations such as this. The results of the analyses will also inform the results of an informal translation process used with non-cognitive measures as well as informing the validity and reliability of a commonly used early grades reading assessment.

This chapter begins with a description of the evaluation design of the educational project used in this study, Valorisation de la Scolarisation de la Fille (VAS-Y Fille!). The design overview will then be followed by a description of the sampling methodology used in the evaluation, the instruments to be used in the proposed study, a brief description the data collection, and finally the method and analysis plan proposed.

VAS-Y Fille! Program Evaluation Design

In order to attribute aspects of the VAS-Y Fille! interventions to changes in student learning outcomes, household and community perceptions, and girl-friendliness in the classroom, a rigorous randomized controlled trial (RCT) was developed. The large representative sample for this project supports greater generalizability and precision of the results, and the longitudinal design takes into account the uniqueness of each student within the general population and allows real assessment of change.

The previous section in Chapter 1 on the

VAS Y Fille! Program outlined the intended intervention design for the project. However, over time this design was somewhat simplified with four main types of interventions, with the recipient populations outlined below:

1. **Financial Interventions** - VAS-Y Fille! awarded need-based scholarships and vouchers to an average of 40 primary girls (in 5th and 6th grade) in each of the intervention schools in order to pay for direct costs of education (school fees). The program also invited community members, specifically girls' parents, to participate in IRC's EA\$E program, which is a savings and loans association¹. The project supported an average of two EA\$E groups including 20 to 25 members each in each Vas-Y-Fille! community.
2. **Instructional Interventions** – The program provided a progressive package of support to the teachers in VAS-Y Fille! schools that included ministry-approved modules on reading and math instruction as well as gender-responsive pedagogy. An average of six teachers per school were trained every year of the intervention². After-school tutoring classes were also organized in the project schools and offered tutoring in reading and math to an average of 80 low-performing students between the 3rd and 6th grade. Each student enrolled in the tutoring program received an additional 6 hours of instruction per week.
3. **Community Involvement** – The program delivered community information campaigns with messages promoting on-time enrolment, championing the importance of education for girls

¹ The EASE model consists of a group of community members who save money together and contribute to a shared fund once a week. Individual members borrow from this common fund and pay the loan back at a modest interest rate, helping the fund grow over time. The group agrees on a pay-out date when each member will receive a share of the common fund, plus accumulated interest.

² Most schools have just one teacher per grade.

and boys, and combating socio-cultural barriers to girls' education. A minimum of two campaigns per community were offered every year. Parent-teacher associations were supported by the project to assess school safety, and develop and implement gender-focused School Improvement Plans (SIPs) that respond to girls' safety needs in and around schools, such as separated bathrooms.

4. **Alternative learning opportunities** – The program supported local civil society organizations to expand their non-formal Accelerated Learning Programmes (ALP) which provide access to education for out of school girls and boys who have never enrolled or have had to interrupt their education. The project financially supported 29 ALP centers enrolling about 200 students each year.

School Sampling

Using a randomized controlled trial (RCT) methodology, the program was evaluated across four time points (Baseline (2013), Annual (2014), Midline (2015), and Endline (2016)), with randomization occurring at the school cluster level. School clusters (212) were categorized by both province (Kasai, Province Orientale, Bandundu, Equateur, or Katanga) and subdivision (each province was composed of 2 to 9 subdivisions). To ensure equal representation across all five provinces and their subdivisions, a stratified random sampling technique was used. The data were first categorized/divided by province, then by subdivisions within each province.

For Kasai, Orientale, Bandundu, & Equateur, approximately one-half of the school clusters within each subdivision were randomly selected to receive the intervention/treatment. For Katanga, approximately 65% of the school clusters within each subdivision were randomly selected to receive the intervention/treatment. Unselected school clusters were assigned to the

control group. For the evaluation sample 43 clusters were selected randomly in both treatment and control groups (86 clusters total). For each cluster, one school was selected randomly to be surveyed, or two when the number of girls in an individual school failed to meet the threshold. As a result (i.e. the need to include seven supplemental schools when the thresholds were not met in the original 86 schools), data were collected from girls in 93 schools. All in-school and ALP girls were randomly selected for interview and/or assessment from within the 93 randomly selected evaluation schools and from 11 ALP centers. Households were also randomly selected for interview from the evaluation communities.

Although all 93 evaluation schools returned to participate in the VAS-Y Fille! project at all time points, of the evaluation, enumerators were unable to re-interview/assess specific in-school and ALP girls as well as some households for each of these follow-up data collection periods. A replacement protocol was used to ensure the sample size – for the purposes of statistical inferences-remained adequate. Because the project stakeholders were keenly interested in the impact of the project over time to determine if multiplicative effects exist, sixth grade students were not replaced via the sampling protocol. Instead, random sampling occurred within third grade classrooms so that these students could be tracked for two or more years.

Student Sampling

Though there were several survey instruments used in the evaluation, the current study will utilize data only from the Early Grades Reading Assessment (EGRA; RTI International, 2016), Early Grades Mathematics Assessment (EGMA; RTI International, 2014), and Survey for the In-School Girls. The sampling protocols used for the in-school girls is below, and Table 8 provides the sample sizes achieved for each group of in-school girls across the Baseline, Annual,

Midline, and Endline data collections. Table 9 shows the sample composition by grade for in-school girls across the Midline and Endline, with the last column showing those girls followed through the Endline.

Table 8. VAS Y Fille! sample sizes for in-school girls per group and data collection instance.

Table 8. VARS 14 line: sample sizes for in-school girls per group and data collection instance.

	Baseline (2013)		Annual (2014)		Midline (2015)		Endline (2016)	
Group	Int. ¹	Con. ²	Int.	Con.	Int.	Con.	Int.	Con.
Grade 3	407	443	415	434	473	468	35	38
Grade 4	450	412	445	429	407	437	542	567
Grade 5	451	454	506	513	562	606	606	591
Grade 6	413	394	363	339	346	299	675	623

¹ Intervention Group; ² Control Group

Table 9. VAS-Y Fille sample composition at Midline and Endline by grade.

Table 3.4.11.1 Time sample comparison at midline and Endline by grade.								
	Midline (2015)		Endline (2016)		Aggregate (Midline + Endline)		Cohort (Recontacted at Endline Only)	
Group	Int. ¹	Con. ²	Int.	Con.	Int.	Con.	Int.	Con.
Grade 3	473	468	35	38	508	506	34	34
Grade 4	407	437	542	567	949	1004	280	300
Grade 5	562	606	606	591	1168	1197	316	289
Grade 6	346	299	675	623	1021	922	373	314

¹ Intervention Group; ² Control Group

In-School Girls Sampling Protocol: Baseline (2013)

At Baseline, twenty girls were randomly selected using a Table of Random Digits from each of two cohorts: (1) class/grade 3rd/4th combined and (2) class/grade 5th /6th combined for a

total of 40 girls per school when more than 20 girls are available. When fewer than 20 girls are available in either cohort, data are collected from all girls. When fewer than 15 girls are available for interview from another randomly selected school in the school cluster, enumerators conduct supplemental interviews in a secondary school within the cluster to obtain the necessary 20 girls. As noted above, seven additional schools were selected when schools in the originally sampled 86 schools were unable to meet the required threshold. Once selected, each girl completes the Girls' Survey (oral responses recorded by enumerator) and the EGMA & EGRA assessments (oral responses recorded by enumerator).

In-School Girls Replacement Sampling Protocol

While all 93 evaluation schools continued their participation in the Vas-y-Fille! Project at both the annual evaluation and midline evaluation, enumerators were unable to re-interview/assess specific in-school and ALP girls as well as some households for each of these follow-up data collection periods. Attrition rates for the in-school girls approximated 40% over the life of the project. Given these high rates of attrition, a replacement protocol was used to ensure that the sample size remained large enough for the statistical comparisons to be completed (Annex X). It should be noted that grade 6 students were not replaced, and instead a random sampling of students in earlier grades was completed so that students in earlier grades could then be tracked over two or more years, allowing an estimate of a multiplicative effect.

Instruments

In order to provide evidence of the program's success, both quantitative and qualitative data was collected on each the intervention and control groups. The School Survey was administered to school directors during the baseline data collection. This survey included

questions surrounding student enrollment, language of instruction, teacher demographics, school's resources, and the previous year's enrollment, attendance, and average achievement in mathematics and literacy. The Girl's Survey collected overall demographic information including items around home language, family structure, health, transportation to school, parental literacy, and attitudes around education including the level of girl-friendliness in the girl's school and perception of their teacher and classmates. In addition to these surveys, the EGMA and EGRA were administered to in-school girls, ALP girls, and out-of-school girls as a measure of math and reading achievement. The EGRA, EGMA, and Girl's Survey are discussed in more detail below.

Early Grade Mathematics Assessment (EGMA)

The core EGMA was developed by RTI International (2014) to assess early mathematics skills in grades one through three. A combination of extensive research on early mathematical learning and assessment, and experts from the fields of mathematics education and cognition put together the conceptual framework and EGMA test form. The core EGMA is comprised of eight competencies that are the fundamentals of early grade mathematics to be administered orally by a trained assessor, including: 1) number identification, 2) number discrimination, 3) missing number, 4) word problems, 5) addition level 1, 6) addition level 2, 7) subtraction level 1, and 8) subtraction level 2. The EGMA has been implemented in two countries and RTI reports the coefficient alpha (Cronbach, 1951; internal consistency reliability) values for each subtest, ranging from 0.44 for word problems to 0.94 for number identification. Developers of the EGMA recommend proper adaptation, assessor training, pilot studies, and finalization of the instrument for each particular use (RTI International, 2014).

The EGMA administered in the VAS Y Fille! program consisted of 5 subtasks, and reliability results from the Baseline data collection are presented in

Table 10 below. Reliability was estimated with coefficient alpha for each subtask in two ways: the first estimate treated missing data as incorrect responses, allowing for a full dataset to be used in the estimate, and the second estimate allowed the missing data to remain missing, resulting in a reduced dataset for the reliability analysis.

Early Grade Reading Assessment (EGRA)

The EGRA was also developed by RTI International (2016) in order to provide a low-cost, valid way to measure the acquisition of reading skills in children in the early grades of primary school. The EGRA is a simple assessment of the initial steps of learning to read such as letter recognition and reading simple words, and was developed by cognitive scientists, early grade reading instruction experts, research methodologists, and assessment experts. Based upon expert feedback sought by RTI, U.S. Agency for International Development (USAID), and the World Bank, the English version of the EGRA was completed with eight subtests that are to be administered orally by trained assessors: 1) letter-name knowledge, 2) phonemic awareness, 3) letter-sound knowledge, 4) familiar word reading, 5) unfamiliar word reading, 6) oral reading fluency with comprehension, 7) listening comprehension, and 8) dictation.

While these eight components have been piloted in several languages (e.g., Arabic, French, Spanish, etc.), developers suggest that any use of the assessment, including language adaptation, or use of a portion of the full form, should be accompanied by the advice of an assessment expert. In addition, it is recommended that assessment users investigate the reliability and validity of the EGRA scores for their particular purpose (RTI International, 2016).

The EGRA administered in the VAS Y Fille! program consisted of 5 subtasks, and reliability results from the Baseline data collection are presented in

Table 10 below. As with EGMA, reliability was estimated with coefficient alpha for each subtask in two ways: the first estimate treated missing data as incorrect responses, allowing for a full dataset to be used in the estimate, and the second estimate allowed the missing data to remain missing, resulting in a reduced dataset for the reliability analysis.

Table 10. Reliability for EGRA and EGMA at Baseline

Test	Subtest	Number of Items	Reliability (Treating missing as incorrect response)	Reliability (With missing data)
<i>EGRA</i>	Recognizing/ Reading Letters Aloud	100	.980	.853
	Reading Imaginary Words	50	.967	.955
	Reading a Story	50	.987	.884
	Reading Comprehension	5	.787	.765
	Listening Comprehension	5	.727	.728
	Reading/ Recognizing Numbers	20	.940	.919
<i>EGMA</i>	Comparing Quantities	10	.861	.736
	Number Sequences Missing Values	10	.772	.715
	Addition	21	.911	.977
	Subtraction	21	.925	.970

Girl-Friendliness Survey

Thirty-eight (37 at the Annual evaluation) four-point Likert-type items were selected from the larger set of survey responses on the Girl's Survey from in-school girls. The items cover a variety of affective topics and broadly fit into three subscales: perceptions of the teacher (22 items), perceptions of school violence (7 items), and general perceptions of the school (9 items and Baseline, 8 items at the Annual evaluation). These items were chosen in particular, because of the possible difficulty in adapting these more affective concepts across cultures, and to provide a comparison of the methods used in this study between objective and subjective measures.

Survey items were developed in English by two external evaluators with expertise in scale development as well as members of the program implementation team located in the DRC in order to ensure that the local context was considered. Once English items were translated into French, a group of 15 Congolese enumerator supervisors in the DRC reviewed the items with an external evaluator to ensure quality, transparency, and consistency of meaning and interpretation. Once items were reviewed, the French and English versions of the surveys were cross-referenced and reconciled by two external evaluators.

Data Collection

All data were collected orally by a group of enumerators overseen and trained by International Rescue Committee (IRC) appointed supervisors. Enumerators administered surveys in the preferred language of the individual participant. For the EGRA, it was expected that the students completed the assessment in French, the national language of the DRC. However, for the EGMA, students were able to complete the assessment in either French or their home language, and responses in the home language were not considered incorrect as they were for the EGRA. Baseline data were collected in the fall of 2013, Annual review data in the spring of 2014,

Midline data in the spring of 2015, and Endline data in the spring of 2016. All data were collected on paper forms and entered into CSPro by an external data entry team in the DRC.

Proposed Analysis

Generalizability Theory

Generalizability Theory (GTheory; Brennan, 1992, 2011; Shavelson & Webb, 1991) is “a statistical theory about the dependability (reliability) of behavioral measurement”, wherein dependability “refers to the accuracy of generalizing from a person’s observed score on a test or other measure” (Shavelson & Webb, 1991, p.1). An individual’s score on a single occasion may be affected by many things (i.e., illness, distractions in the test space, improper administration, etc.), and GTheory allows us to estimate the sources of error variance attributable to multiple sources in one analysis.

In a Generalizability study (GStudy), the researcher can identify and select pertinent facets (factors that may be sources of error variance) and look at them in relation to one another, allowing us to attribute smaller or larger sources of variability (and thus, lower reliability estimates) to a particular facet. The lower the error variance in the data, the higher the quality, or reliability (Bayerl & Paul, 2007). By parsing apart the sources of variability, we can see the impact of a particular facet on the quality of the data, and make requisite changes in order to improve it, making the results particularly useful in pilot or longitudinal studies wherein there is a possibility of adapting or editing the instrument.

Facets in GTheory are synonymous with factors in ANOVA and can be defined as fixed or random. A facet is fixed when all possible levels, or conditions, of the facet are present in the

data set and no sampling of conditions has occurred. A facet is considered random when a sampling of levels, or conditions, has occurred in the possible universe of levels (Brennan, 1992, 2001). For example, if the purpose of an experiment is to compare test scores across three languages, language is the factor, and the three levels are the languages (i.e., French, Spanish, and German). If there is no intention on behalf of researchers to say anything about any other languages (that is, to generalize to all languages), the facet is considered fixed. If, however, there is an intention to generalize the findings to all languages, the facet is considered random. Prior to analysis, facets must be defined as either fixed or random as the estimation of variance components is carried out in different ways.

Two types of designs are possible in GTheory: a crossed design and a nested design. A crossed design is the simplest to analyze and provides the most information as a variance component is estimated for each facet individually as well as for all possible interactions. Figure 6 shows a Venn diagram of the error variance associated with two facets in a fully crossed design: item/subtest ($\hat{\sigma}_{i/s}^2$), person ($\hat{\sigma}_p^2$), and language of survey ($\hat{\sigma}_l^2$). In this example, all persons were administered all items in all languages. In the figure we can see that there are variance component estimates for each of the sections of the diagram, including all possible interactions. This allows for attribution of variability to each facet or interaction of facets. Table 11 shows the sources of error variance in the two-facet, crossed design.

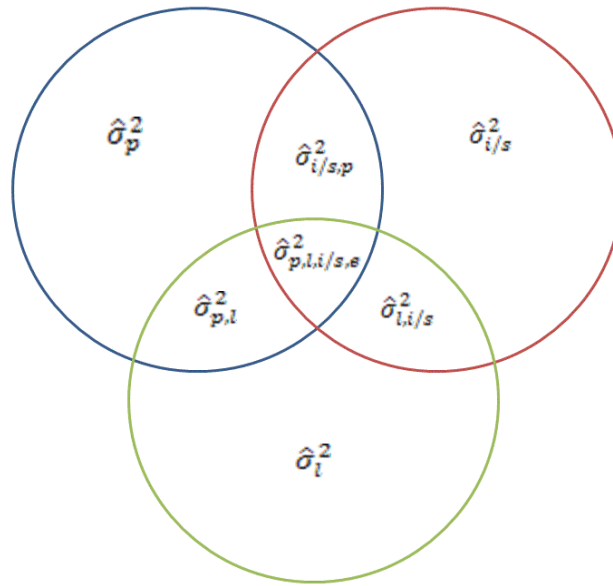


Figure 6. Sources of variability for a two-facet fully crossed design.

Variance component estimates for each of the sections of the diagram, including all possible interactions. This allows for attribution of variability to each facet or interaction of facets. Table 11 shows the sources of error variance in the two-facet, crossed design.

Table 11. Sources of variability in a two-facet fully-crossed design

Source of Variability	Type of Variation	Notation
<i>Person (p)</i>	Universe-score variance (object of measurement); systematic variance between persons responses	$\hat{\sigma}_p^2$
<i>Item (i)</i>	Constant effect for all persons due to the difficulty of items	$\hat{\sigma}_i^2$
<i>Language of Administration (l)</i>	Constant effect for all persons due to the language of administration	$\hat{\sigma}_l^2$

Source of Variability	Type of Variation	Notation
$p \times i$	Inconsistencies of item responses for particular persons	$\hat{\sigma}_{pi}^2$
$p \times l$	Inconsistencies of responses from particular persons in particular languages	$\hat{\sigma}_{pl}^2$
$i \times l$	Inconsistencies of item responses for particular languages of administration	$\hat{\sigma}_{il}^2$
$p \times i \times l, e$	Residual consisting of the unique combination of p, i, l, as well as unmeasured facets and random events	$\hat{\sigma}_{pil,e}^2$

Nested designs, however, do not allow for the same distinctions between facets as in the fully crossed design. For example, persons are nested within the language of the survey (that is, they only completed the survey in one language), and all persons completed all items/subtests. Figure 7 shows a Venn diagram of the error variance estimates for this two-facet nested design including language ($\hat{\sigma}_l^2$), items/subtest ($\hat{\sigma}_{i/s}^2$), and people nested within language ($\hat{\sigma}_{p:l}^2$). Table 12 shows the sources of error variance for this partially-nested design.

Table 12. Sources of variability in a two-facet partially-nested design

Source of Variability	Type of Variation	Notation
<i>Person (p)</i>	Universe-score variance (object of measurement); systematic variance between persons responses	$\hat{\sigma}_p^2$
<i>Item (i)</i>	Constant effect for all persons due to the difficulty of items	$\hat{\sigma}_i^2$
$l:p$	Nested component measuring the variability of persons responses across language of administration	$\hat{\sigma}_{l:p}^2$

Source of Variability	Type of Variation	Notation
$p \times i$	Inconsistencies of item responses for particular persons	$\hat{\sigma}_{pi}^2$
$(l:p) \times i, error$	Residual due to confounded and unmeasured sources of variability	$\hat{\sigma}_{pl,pli,e}^2$

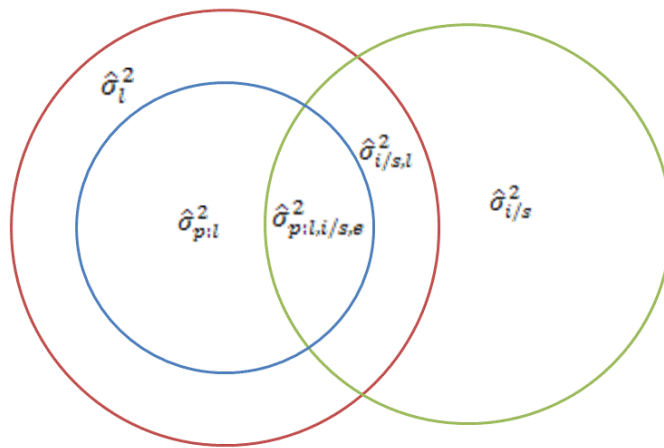


Figure 7. Sources of variability for a two-facet partially-nested design

Often, operationally, nested designs are more realistic and allow for smaller sample sizes, decreasing the time and cost of collecting data. Because of the nesting, it is not possible to disentangle the variability attributable to persons from that attributable to language of administration. It must be noted that in applied settings in which language is a facet, the design must, by necessity, be nested as we would not expect individuals to complete an interview or survey in more than one language. It may also be the case that a fully-crossed design is used using a smaller sample (perhaps a subset of the full sample) or in a pilot study in order to aid in

decision making for the full study regarding where to focus funding (i.e., more participants, less languages, less enumerators, etc. for the full study, which may well be partially or fully nested.

The results of a GStudy are estimates of the magnitude of whichever sources of error were identified in the study. These estimates may then be used to calculate reliability specific to the types of decisions one wishes to make based on the data. These interpretations may be either relative or absolute. Relative interpretations are based on the relative standing, or ranking, of scores which is often referred to as norm-referenced testing. Absolute interpretations are based on the absolute scores obtained; this is also referred to as criterion referenced testing. For example, college admissions (i.e., SAT, ACT, or GRE scores) make relative decisions by ranking student scores and taking the top applicants. Conversely, most certification decisions are made using absolute interpretations wherein there is a passing score set in advance and students must meet or exceed the score (Shavelson & Webb, 1991).

Using the results from the GStudy, the reliability coefficient appropriate for the decisions to be made, can be calculated. In general, reliability coefficients are calculated by dividing true score variance by observed score variance (which includes both true and error variance); the smaller the error variance, then, the larger the reliability coefficient. In GTheory, this calculation is done by dividing the universe score variance ($\hat{\sigma}_p^2$) by the observed score variance (including $\hat{\sigma}_p^2$ and error variance). The type of error variance included in the formula is dependent on the type of decision to be made, absolute or relative. Relative error variance (δ) includes all variance component estimates that include persons, and is used in the estimation of the Generalizability (G) coefficient ($G = \sigma_p^2 / (\sigma_p^2 + \sigma_\delta^2)$). Relative decisions, and the G coefficient are used when making norm-referenced assessment decisions, and where one wants to consider the ordering of those assessed. Absolute error variance (Δ) includes all variance component estimates, and is

used in the estimation of the Dependability (Φ) coefficient ($\Phi = \sigma_p^2 / (\sigma_p^2 + \sigma_\Delta^2)$). Absolute decisions, and the Φ coefficient are used when making criterion-referenced assessment decisions, and where one is only concerned with the level of performance of those assessed.

Many-Facet Model

The Many-Facet (MF) model is a model falling under Rasch Measurement Theory (RMT; Andrich, 1978; Rasch, 1980; Wright & Masters, 1982), which follows a different measurement tradition to GTheory. Rasch developed a model of measurement based on the cumulative distribution a set of requirements around specific objectivity which supports a view of invariant measurement that allows the conceptual separation between items and persons (Engelhard & Wind, 2018; Rasch, 1980). One of the distinct differences from GTheory, is under the MF model, persons are treated as facets, allowing for assessment of the individual in line with the rest of the data, to be treated as an object of measurement rather than the subject.

Rasch proposed a basic measurement model to represent response probabilities such that the probability of a correct response ($a_{ni} = 1$), and the probability of an incorrect response ($a_{ni} = 0$), is represented as:

$$P\{a_{ni}\} = \frac{\theta_n \sigma_i^{a_{ni}}}{1 + \theta_n \sigma_i}, \text{ where}$$

θ = the parameter for person n , representing the location of a person on the construct, and

σ = the parameter for item i , representing the location of the item.

This has more recently been expressed in exponential form as:

$$\Pr\{a_{ni} = 1\} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}, \text{ where}$$

$\Pr \{a_{ni} = 1\}$ = the probability of observing a correct response.

One of the main benefits noted previously of RMT is the value of the Wright map (Wright & Masters, 1982), which places facets on what is called a logit scale for comparison and interpretation. A logit, or log-odds, scale is a representation of the underlying scale, whether it be math ability, reading, or some behavior. In the simplest form, with persons and items only, the Wright map appears as pictured in

Logit	Person Score
	High Reading Ability
4	3
3	5 18
2	1 8
1	7 9 11
0	2 6 12 13
-1	7 15 19
-2	4 14 17
-3	16 20
-4	10
	Low Reading Ability

Figure 8. In this figure, $\widehat{\theta}_n$ represents a person's placement along the continuum (logit scale) of test items when ordered from easy to difficult, essentially showing us where the individual is placed in relation to the spectrum of difficulty for this construct. In a larger dataset, one can see all of the individual persons placed along the line, and one can assess how difficult the test was, or if there are particular groupings of persons along the line, prompting you to investigate further.

Logit	Person Score
	High Reading Ability
4	3
3	5 18
2	1 8
1	7 9 11
0	2 6 12 13
-1	7 15 19
-2	4 14 17
-3	16 20
-4	10
	Low Reading Ability

Figure 8. Simple one-facet Wright map.

When we expand the example to include more facets, for example multiple occasions of assessment and the use of a rater, the model for analyzing this type of data may be written as:

$$\phi_{nmik} = \frac{P_{nmik}}{P_{nmik} - 1 + P_{nmik}} = \frac{\exp(\theta_n - \lambda_m - \delta_i - \tau_k)}{1 + \exp(\theta_n - \lambda_m - \delta_i - \tau_k)}, \text{ where}$$

P_{nmik} = the probability of person n being rated k on occasion i by rater m ,

θ_n = the judged location of person n ,

λ_m = the severity of rater m ,

δ_i = the judged difficulty of occasion i , and

τ_k = the judged difficulty of rating category k relative to category $k-1$.

With this expansion, the Wright map becomes even more informative as pictured in

Figure 9. This example follows from the facets identified in the GTheory example, using persons,

items, and language. In this case, we use the EGRA as an example, and the three facets are persons, EGRA subtasks, and preferred home language of the girl being assessed. We can see in this example that there are 20 persons, five subtasks (LN – Letter Name, NW – Nonsense Word Reading, ORF – Oral Reading Fluency, RC – Reading Comprehension, LC – Listening Comprehension), and six languages (FR- French, TS – Tshiluba, BE – Bemba, LI – Lingala, KIL – Kilendu, KIK – Kikongo SW – Swahili). Interpretation of the map is relatively intuitive with the understanding of the study design. In the person column, we see that girl 3 had the highest level of reading ability, and girl 10 had the lowest. In the item difficulty column we can see that reading comprehension was the most difficult task for the girls, and letter naming was the least difficult. Finally, those girls who indicated that their preferred language was French has the highest level of reading ability, and the girls who indicated Swahili as their preferred language had the lowest.

Logit	Person Score	Item Difficulty	Language
	High Reading Ability	More Difficult	More
4	3		FR
3	5 18	RC	TS BE
2	1 8	NW ORF	LI
1	7 9 11	LC	
0	2 6 12 13		
-1	7 15 19		KIL
-2	4 14 17	LN	
-3	16 20		KIK
-4	10		SW
	Low Reading Ability	Less Difficult	Less

Figure 9. Wright map for Many-Facet Model.

The MF model also allows for a more detailed examination of the differences within a facet. For example, it is possible to determine if a persons' ratings are as expected by the model, which can point to further investigations of persons who do not seem to fit the model as expected. Or, one can look at whether the results of girls with different preferred languages were statistically significantly different. All of these results allow for a fulsome picture of the assessment. Taken together, the results of a GStudy and the MF model analysis will provide a more detailed picture of the EGMA, EGRA, and Girl-Friendliness survey used in the VAS-Y Fille! program evaluation.

Summary

Recall the primary purpose of this study is to explore the use of Generalizability Theory (GTheory; Brennan, 1992, 2001; Shavelson & Webb, 1991) and Rasch Measurement Theory (Rasch, 1980; Wright & Masters, 1982) in the form of the Many-Facet (MF) model (Linacre, 1989) to assess possible sources of unreliability in data from an international evaluation to be used as evidence of success in outcomes of an educational initiative. The study will respond to the following:

1. What are the largest sources measurement error in the current evaluation design, and how do they differ for subjective vs. objective measures?
2. What is the effect the of non-standard translation and adaptation procedures used on the assessments throughout the VAS-Y Fille! evaluation?
3. What facets are modifiable in a program such as VAS-Y Fille! that would allow for a decrease in the measurement error of the outcome measures?

Two types of measures will be assessed: a set of subjective, or affective, survey items, and objective achievement measures of mathematics and reading. Conducting analyses on both types of measures will allow a more comprehensive discussion on the usefulness of GTheory and the MF model in evaluation. Facets of interest across the measures include language of administration, enumerator, occasion, and person.

Subjective Measurement

Girl-Friendliness Survey

The Girl-Friendliness Survey (in Appendix A. consists of three scales with items pertaining to violence, teacher characteristics, and school characteristics from the Girl's Survey. Girls were interviewed in their preferred language, and enumerators adapted the French version of the items during the interview. As adaptation was not completed prior to administration of the surveys, language was identified as a particular interest to the researchers.

Objective Measurements

Early Grades Mathematics Assessment

The EGMA administered in this study consists of 5 activities: Number Identification (20 items), Comparison of Quantity (10 items), Sequence Completion (10 items), Addition (21 items), and Subtraction (21 items). As with the Girl-Friendliness Survey, girls were assessed in their preferred language, and enumerators adapted the French version of the items during the interview.

Early Grades Reading Assessment

The EGRA administered in this study consists of 5 activities: Letter Identification (100 items), Reading Invented Words (50 items), Reading a Story (50 items), Reading Comprehension (5 items), and Listening Comprehension (5 items). This assessment was administered only in French, as it is the national language in the DRC. However, the majority of students choose to communicate in a local language which is not the language the teachers use in the classroom.

CHAPTER IV. RESULTS

Introduction

Chapter four is comprised of three major sections: descriptive statistics for all Early Grades Reading Assessment (EGRA), Early Grades Mathematics Assessment (EGMA), and Girls' Survey results; the results from the Generalizability Theory (GTheory) analyses; and the results from the Many-Facet Model analyses. For each of these major sections, the dataset determination process is outlined, and then the results are presented in the following order:

- 1) Descriptive Statistics for the EGMA and EGRA at Baseline, and Girls' Survey results by subtest for both the Baseline and Annual evaluation points.
- 2) Baseline – Objective Measures: this section includes analyses on the Baseline evaluation data for each of the five EGRA subtasks with data at the item level, and each of the five EGMA subtasks with data at the item level.
- 3) Baseline – Subjective Measures: this section includes analyses on the Baseline evaluation data for each of the three sections of the Girl's Survey with data at the item level.
- 4) Annual – Subjective Measures: this section includes analyses on the Annual evaluation data for each of the three sections of the Girl's Survey with data at the item level. EGRA and EGMA item-level data were not available.
- 5) Longitudinal – Subjective Measures: this section includes analyses on longitudinal data from the Baseline and Annual evaluations for each of the three sections of the Girl's Survey with data at the item level. Cases where the same girls responded to the survey items at both time points were used in this analysis, with a code for administration.

It should be noted that EGRA and EGMA subtask scores were not analyzed together in either the Generalizability or Many-Facet Model analyses as their score scales are not equal. Each subtask was analyzed separately at the item level.

Descriptive Statistics

Objective Measures

Table 13 contains the descriptive statistics for each of the EGRA and EGMA subtasks administered at Baseline, including the percent of zero scores. We see that, particularly for the EGRA subtasks, there are a large proportion of zero scores, and the examinees did not perform well on this assessment in general. Results for the EGMA are slightly better, with only the Subtraction task showing a significant proportion of zero scores. Figure 10 through Figure 19 show the distributions of scores for each subtask. In all cases except for the Addition subtasks, there is, at times severe, levels of skewness. These results are not atypical for EGRA and EGMA results.

Table 13. Descriptive Statistics for Baseline EGRA and EGMA Subtasks

	n	Min.	Max.	% Zero Scores	Mean	S.D.	Skew
<i>EGMA</i>							
<i>Number Identification</i>	3434	0.00	20	0.41%	15.28	5.37	-1.03
<i>Number Discrimination</i>	3434	0.00	10	2.15%	7.44	2.77	-1.08
<i>Missing Number</i>	3434	0.00	10	6.46%	3.48	2.78	0.85
<i>Addition</i>	3434	0.00	20	1.89%	10.32	4.88	-0.09
<i>Subtraction</i>	3434	0.00	20	22.66%	6.47	5.31	0.30
<i>EGRA</i>							

	n	Min.	Max.	% Zero Scores	Mean	S.D.	Skew
<i>Letter Name</i>	3434	0.00	100	21.61%	19.41	19.77	1.10
<i>Nonword Reading</i>	3434	0.00	50	53.87%	6.25	9.61	1.78
<i>Oral Reading Fluency</i>	3434	0.00	50	63.22%	8.06	13.99	1.70
<i>Reading Comprehension</i>	3434	0.00	5	79.70%	0.45	1.04	2.48
<i>Listening Comprehension</i>	3434	0.00	5	57.66%	0.92	1.32	1.37

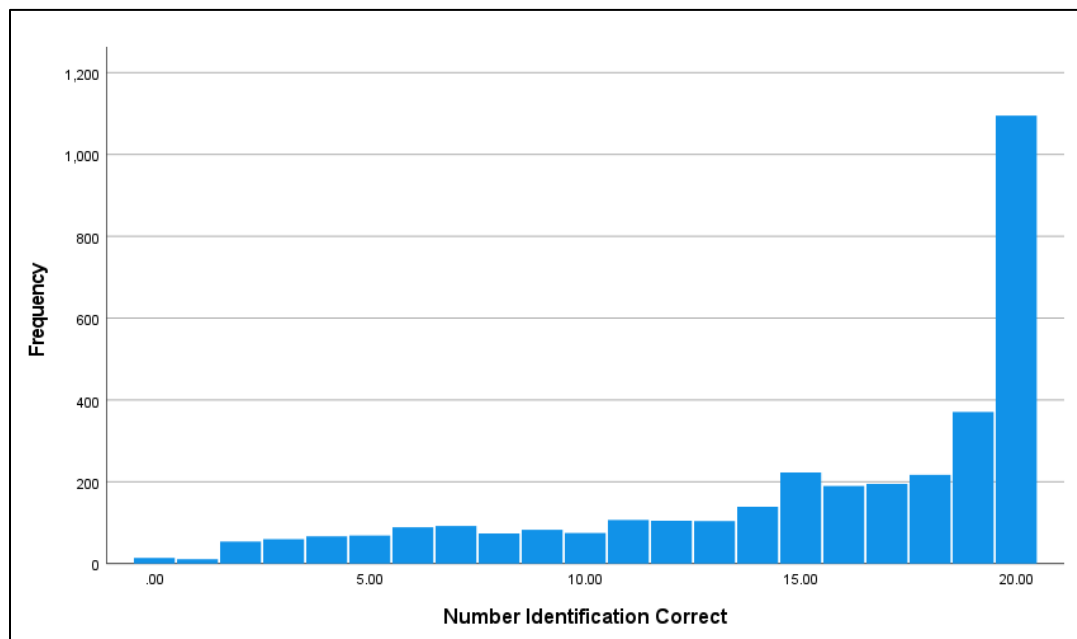


Figure 10. EGMA - Number Identification Subtask Number Correct Distribution

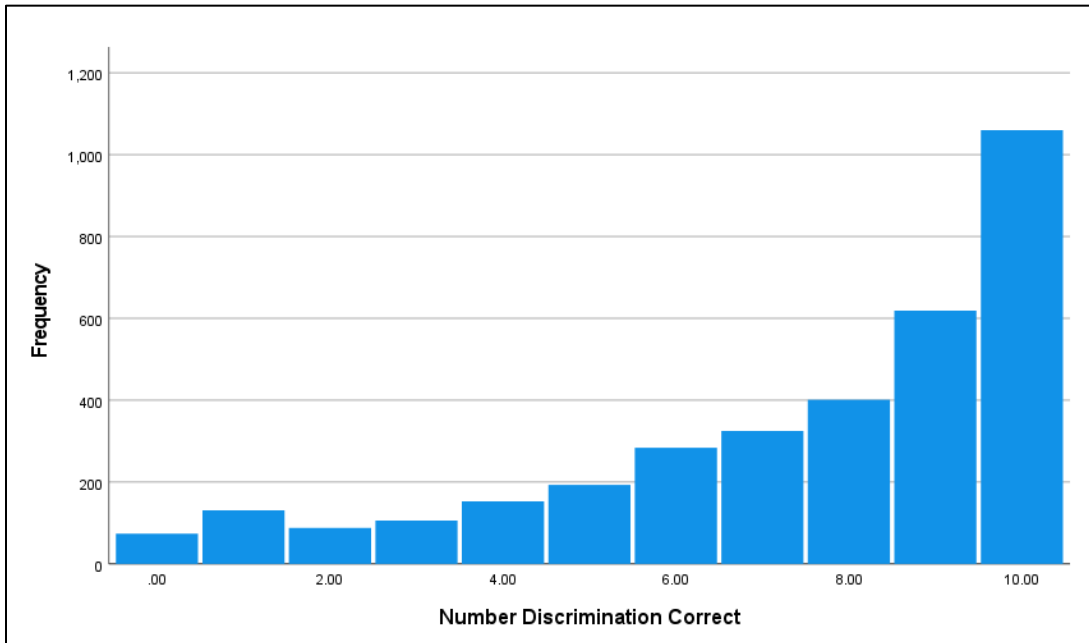


Figure 11. EGMA - Number Discrimination Subtask Number Correct Distribution

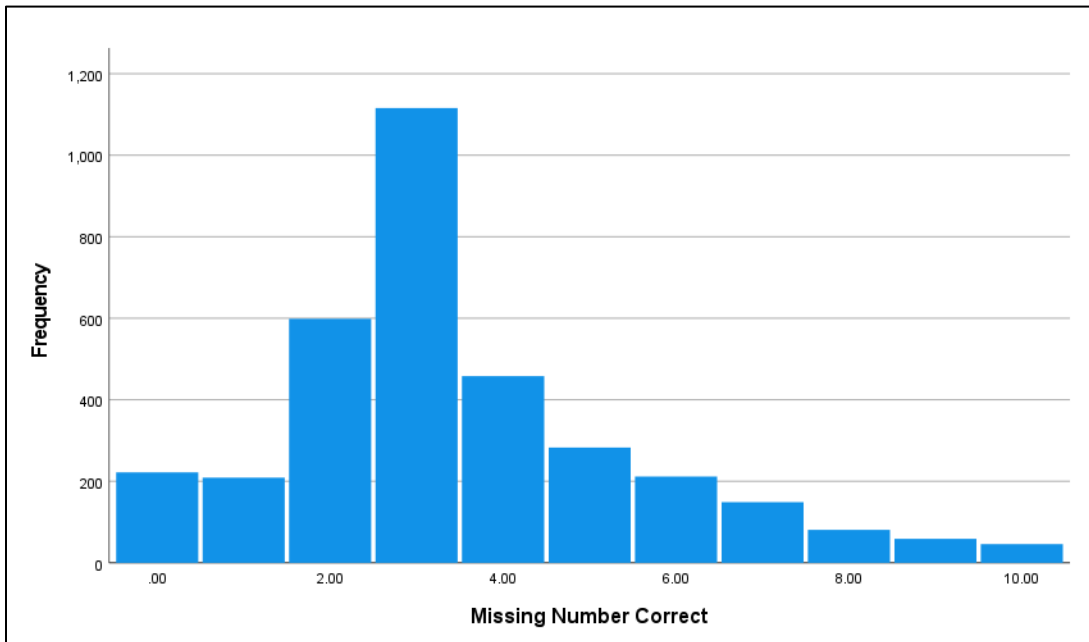


Figure 12. EGMA - Missing Number Subtask Number Correct Distribution

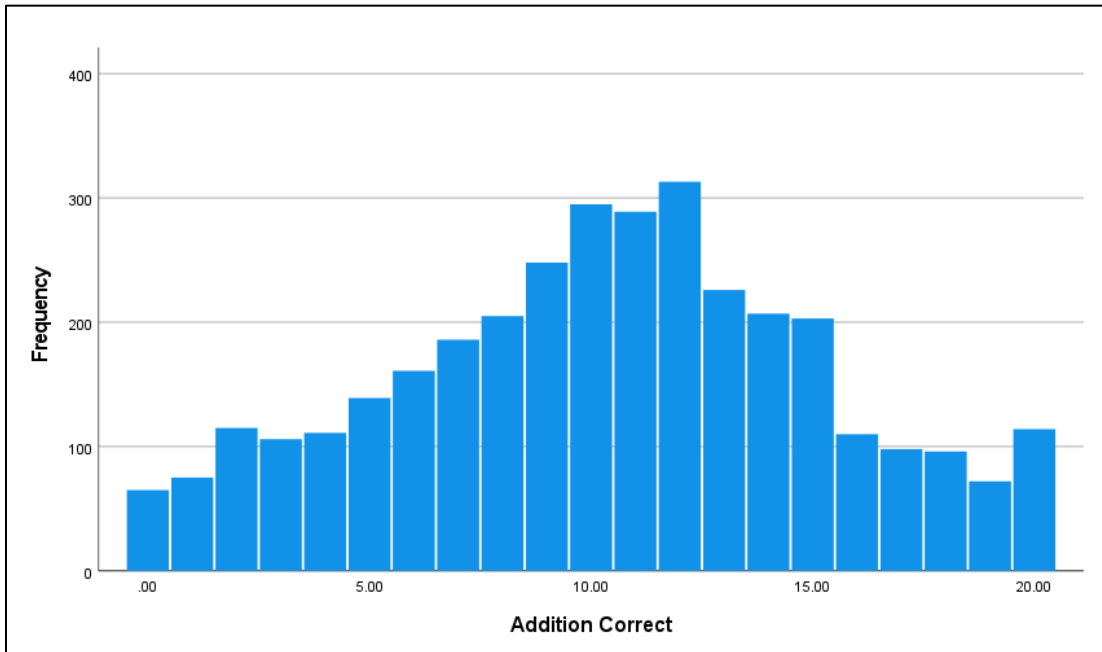


Figure 13. EGMA - Addition Subtask Number Correct Distribution

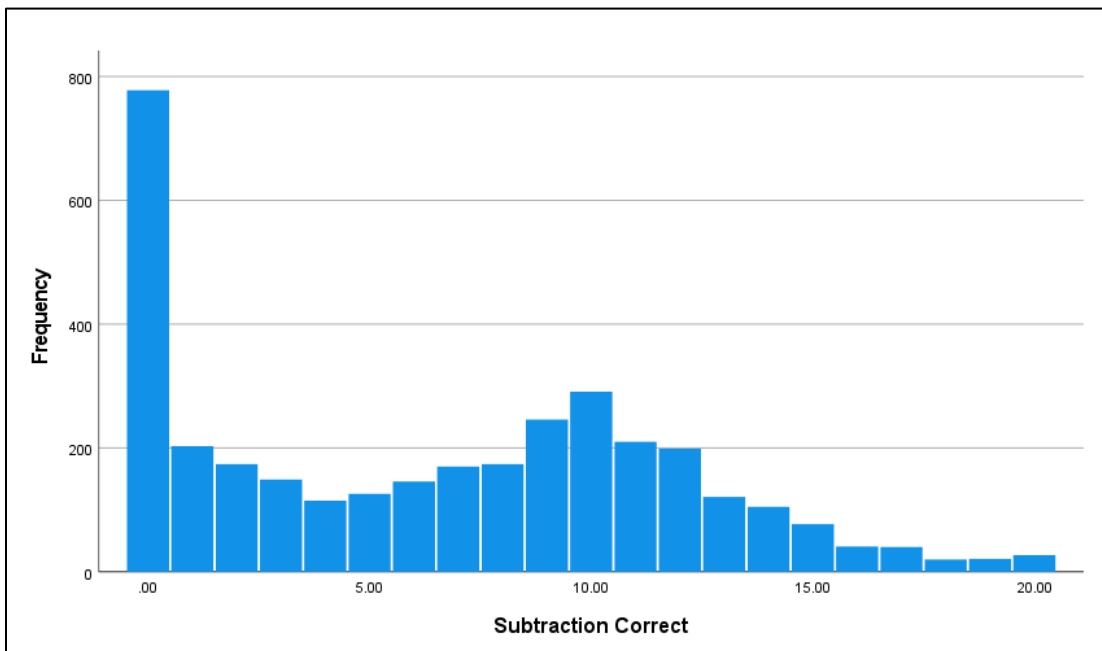


Figure 14. EGMA - Subtraction Subtask Number Correct Distribution

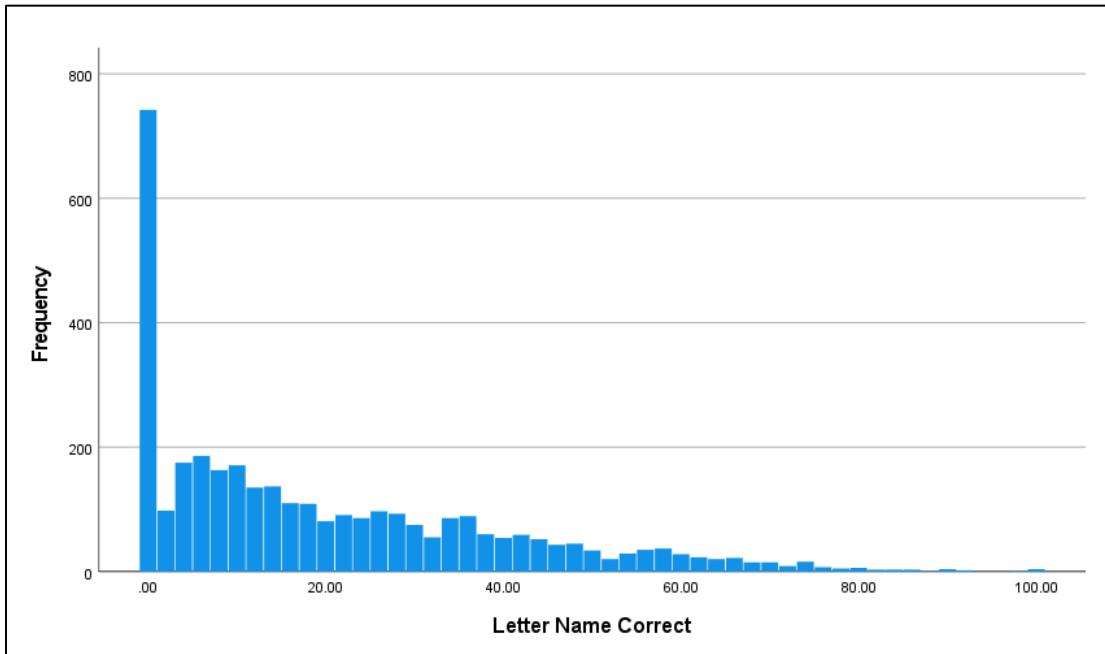


Figure 15. EGRA - Letter Name Subtask Number Correct Distribution

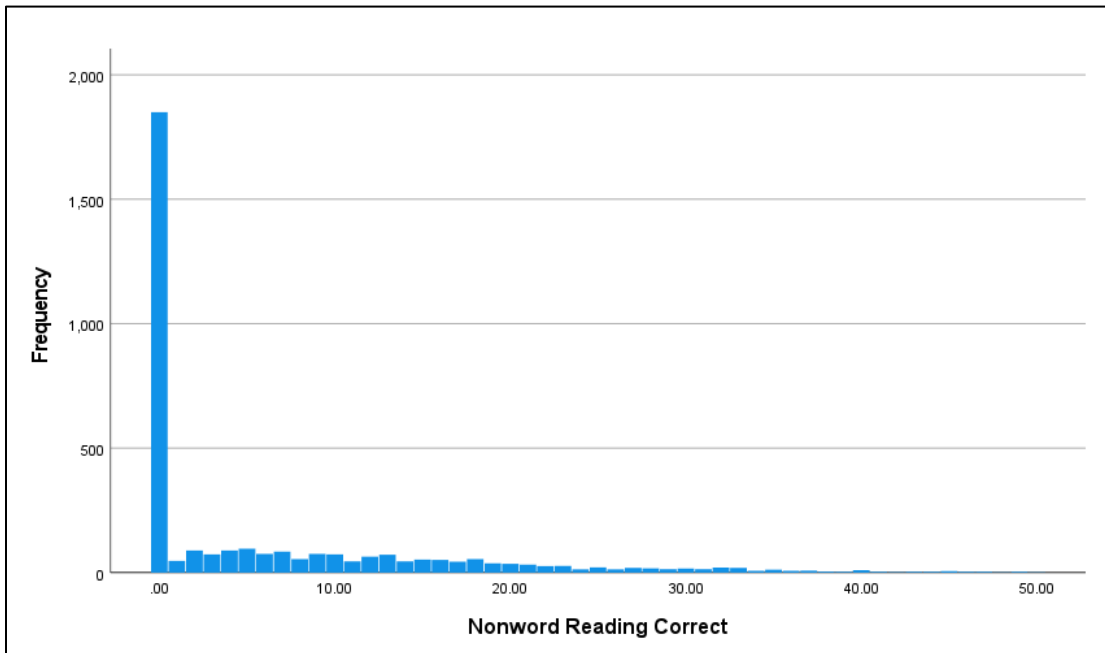


Figure 16. EGRA - Nonword Reading Subtask Number Correct Distribution

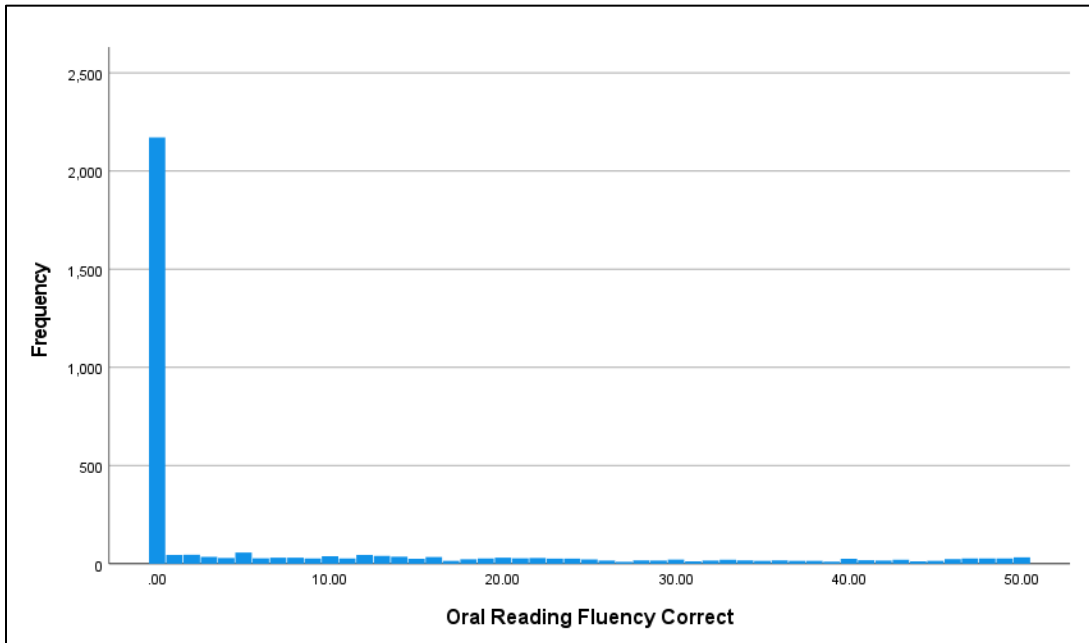


Figure 17. EGRA - Oral Reading Fluency Subtask Number Correct Distribution

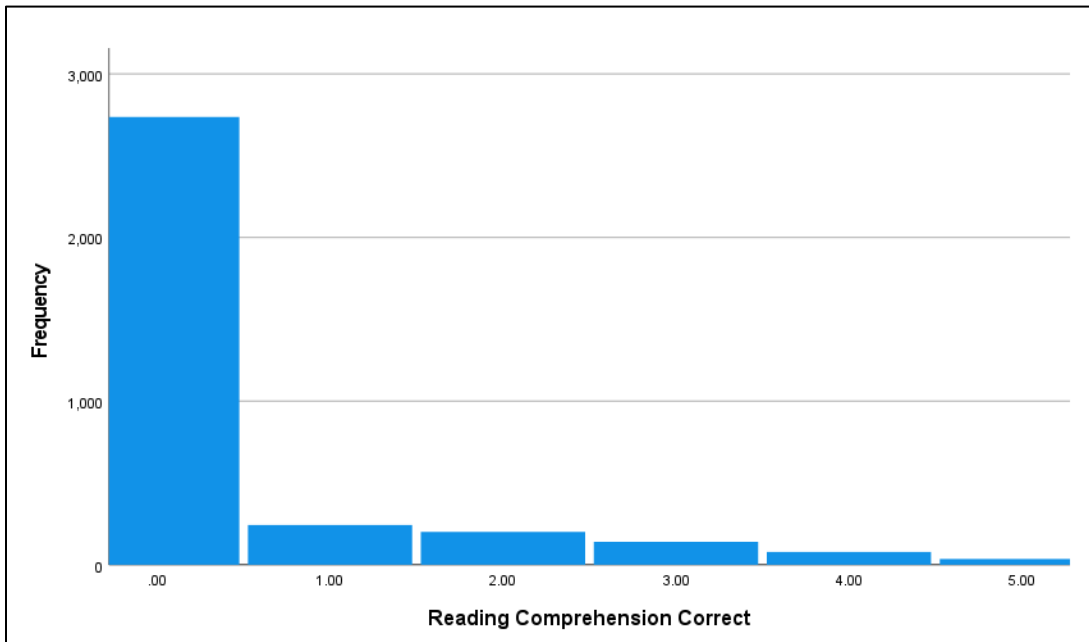


Figure 18. EGRA - Reading Comprehension Subtask Number Correct Distribution

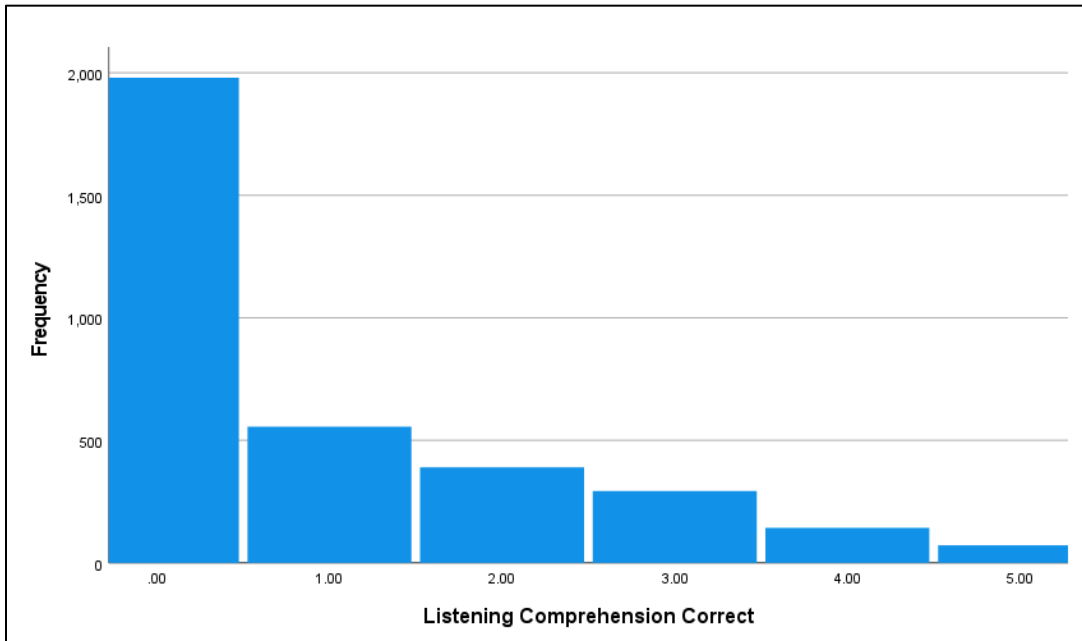


Figure 19. EGRA - Listening Comprehension Subtask Number Correct Distribution

Subjective Measures

Table 14 contains the descriptive statistics for each of the three Girls' Survey "subtests" administered at Baseline and Annual evaluation time points. In addition, it includes the statistics for the merged files with results for both Baseline and Annual represented. The merged files include data collected at the two time points only for those cases where data existed at both time points. Figure 20 through Figure 31 show the distributions of scores for each survey subtest. The survey results show, in general, less skew than the EGRA/EGMA results presented above. In addition, there is generally less variability in survey responses at the Annual evaluation than at Baseline.

Table 14. Descriptive Statistics for Baseline and Annual Girls' Survey Results

	n	Min.	Max.	Mean	S.D.	Skew
<i>General School Perceptions</i>						
<i>Baseline</i>	2663	15	36	29.84	3.82	0.05
<i>Annual*</i>	2436	16	31	25.03	2.68	-0.14
<i>Merged – Baseline</i>	1463	14	32	26.55	3.79	0.06
<i>Merged - Annual</i>	1463	16	31	25.08	2.68	-0.10
<i>Teacher Perceptions</i>						
<i>Baseline</i>	3276	44	88	71.15	7.64	-0.34
<i>Annual</i>	3413	35	67	54.50	2.77	-0.24
<i>Merged – Baseline</i>	2325	45	87	71.28	7.69	-0.35
<i>Merged - Annual</i>	2325	36	67	54.57	2.76	-0.12
<i>Perceptions of School Violence</i>						
<i>Baseline</i>	3375	7	28	23.00	3.55	0.04
<i>Annual</i>	3428	10	28	22.75	2.88	-1.02
<i>Merged – Baseline</i>	2389	7	28	22.87	3.59	-0.80
<i>Merged - Annual</i>	2389	10	28	22.73	2.87	-1.00

*Annual survey contained one fewer item than Baseline

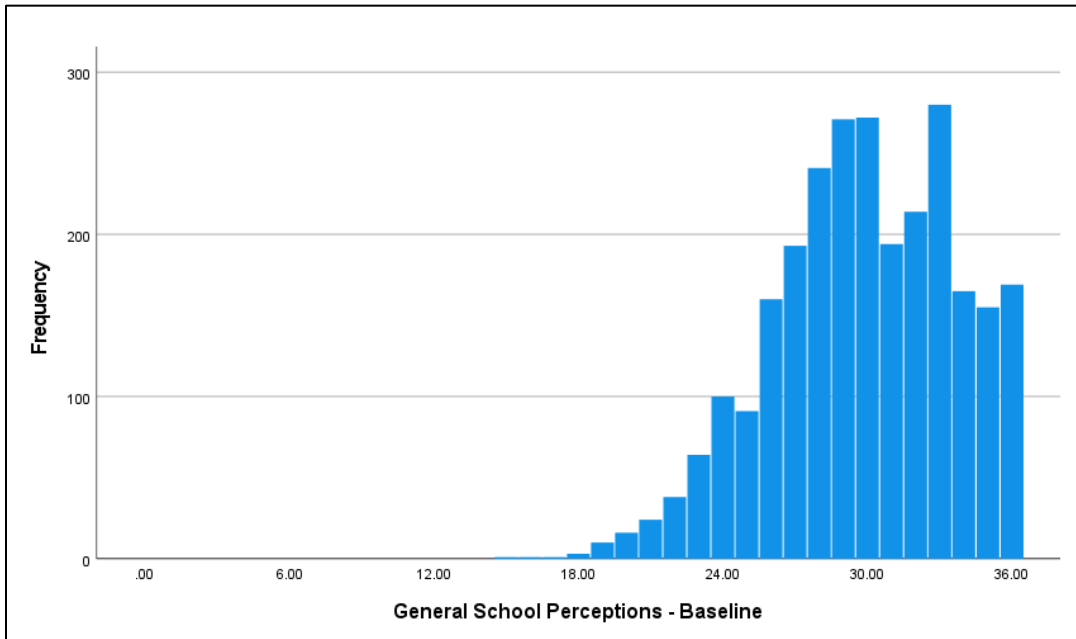


Figure 20. Girls' Survey General School Perception Baseline Total Score Distribution

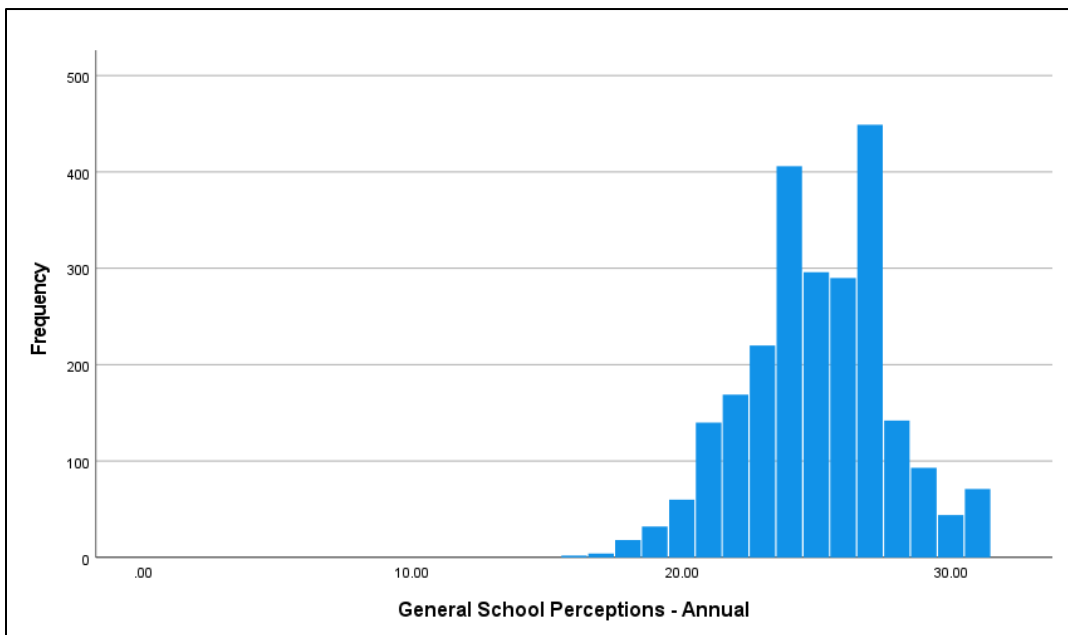


Figure 21. Girls' Survey General School Perception Annual Total Score Distribution

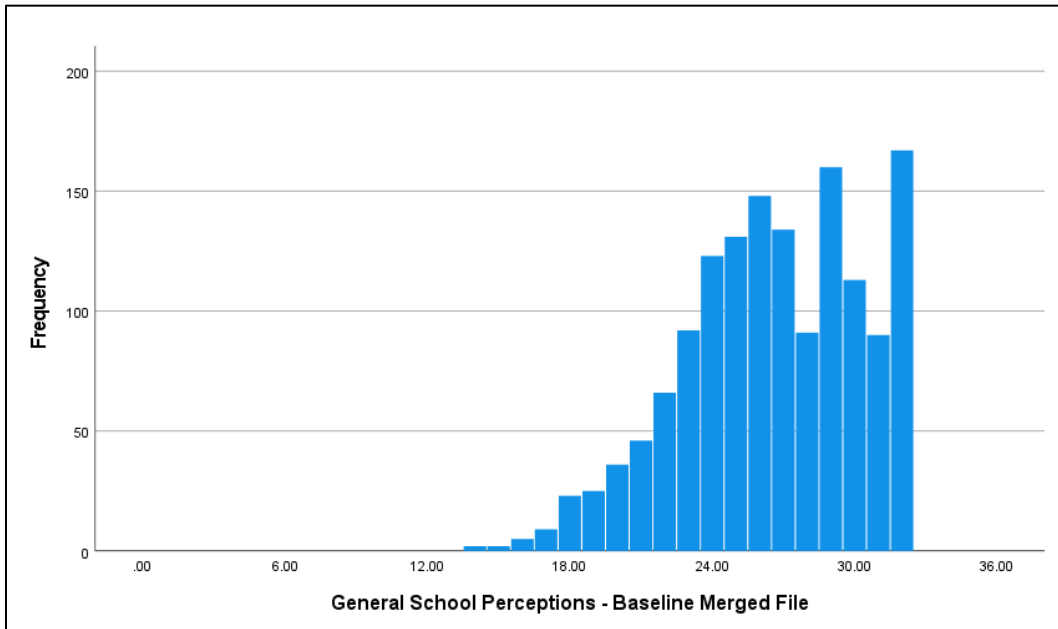


Figure 22. Girls' Survey General School Perception Baseline Merged File Total Score Distribution

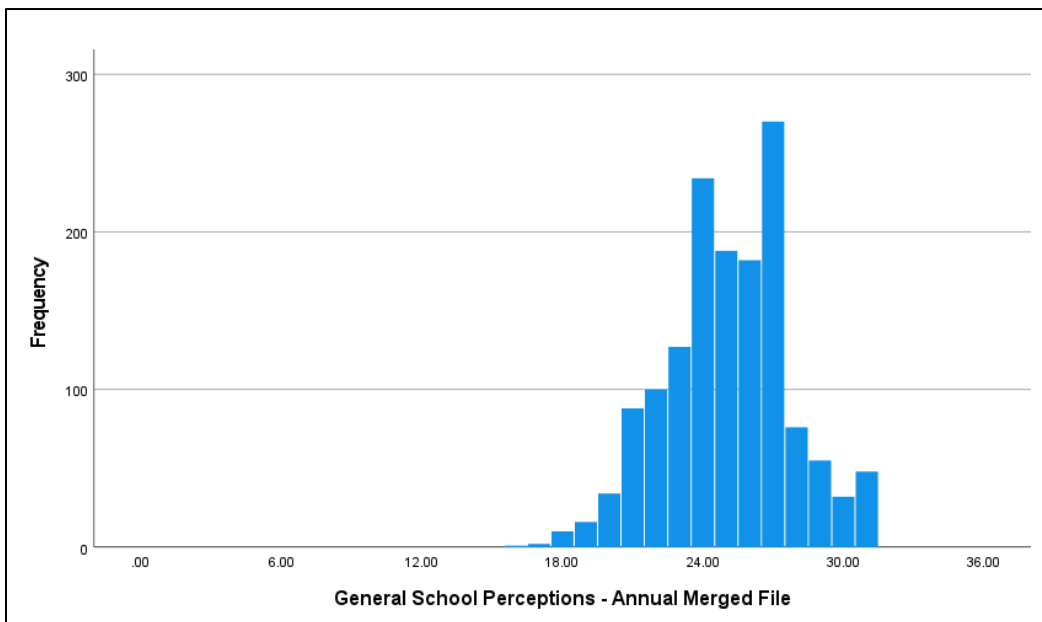


Figure 23. Girls' Survey General School Perception Annual Merged File Total Score Distribution

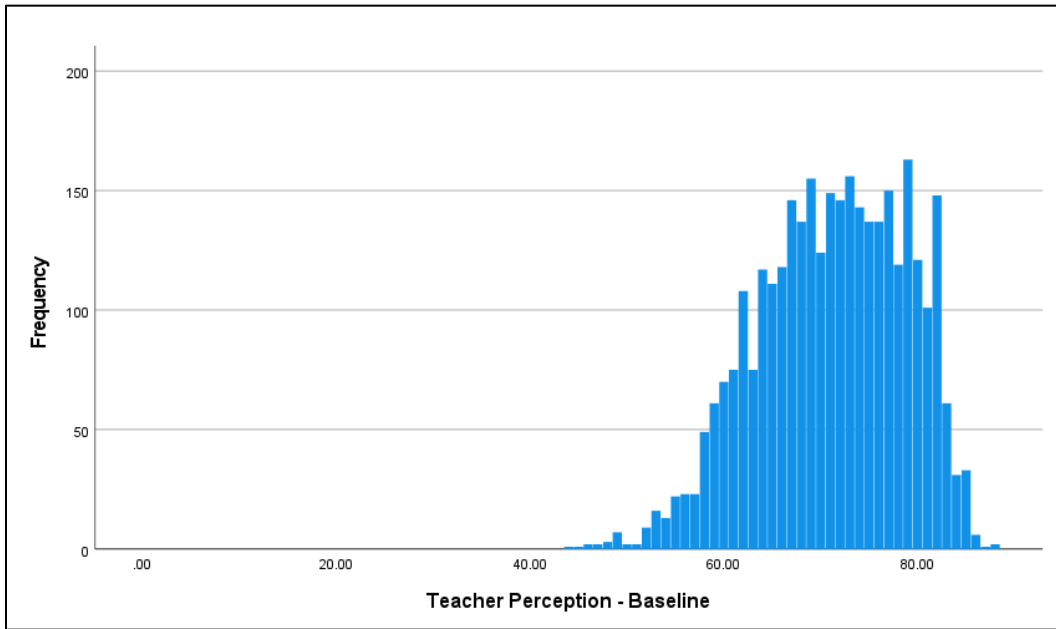


Figure 24. Girls' Survey Teacher Perception Baseline Total Score Distribution

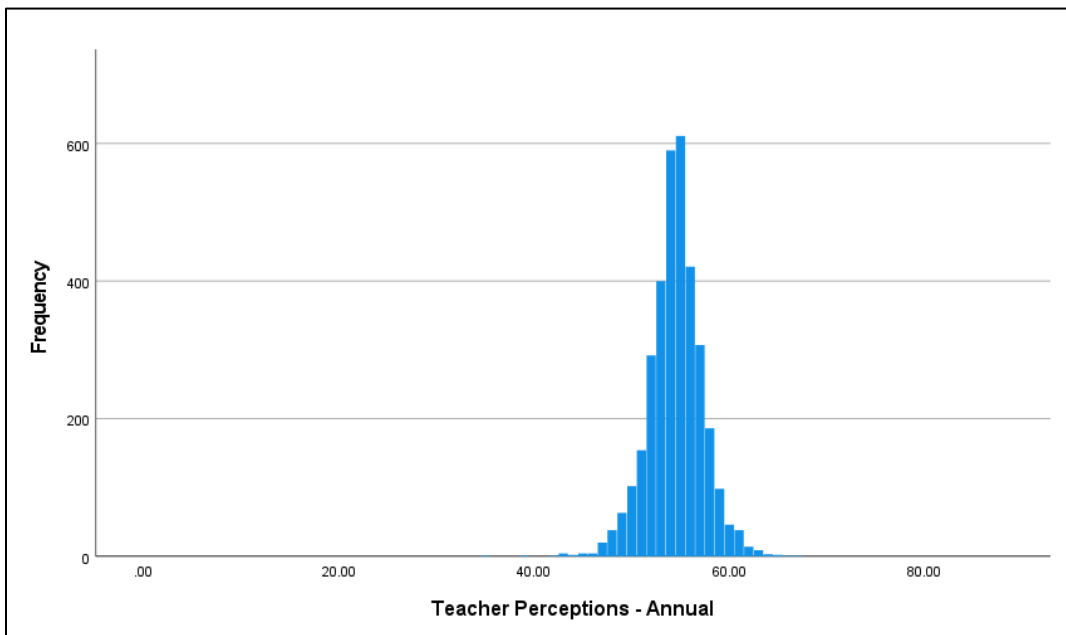


Figure 25. Girls' Survey Teacher Perception Annual Total Score Distribution

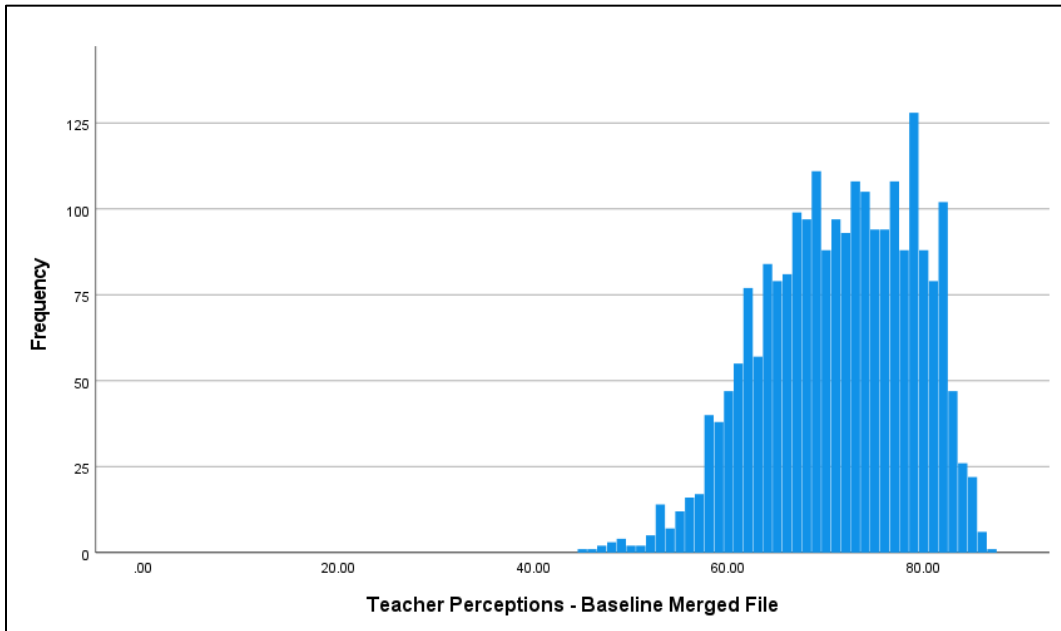


Figure 26. Girls' Survey Teacher Perception Baseline Merged File Total Score Distribution

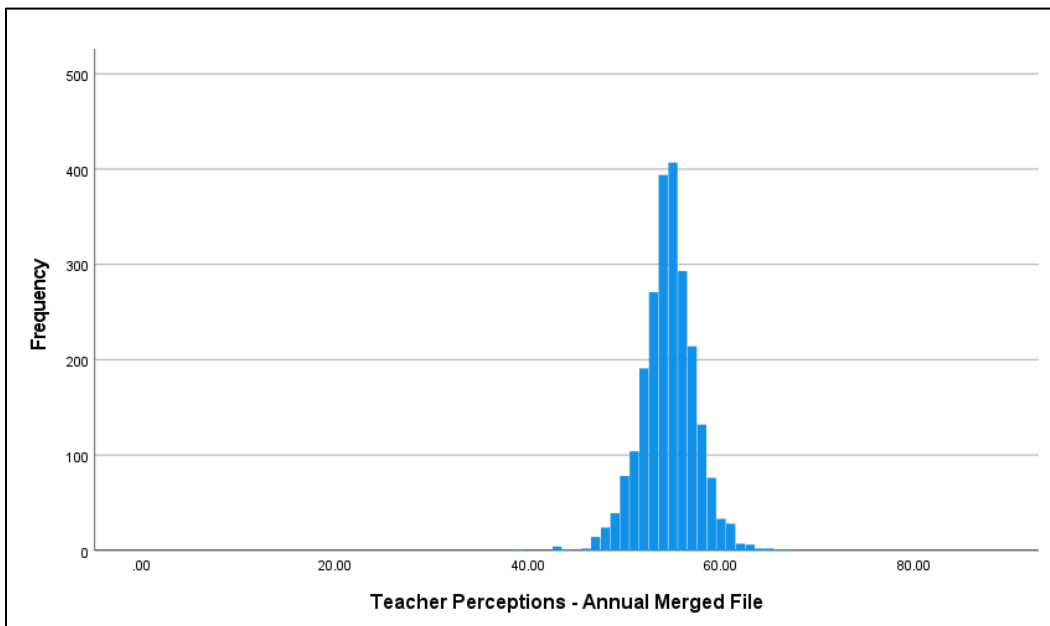


Figure 27. Girls' Survey Teacher Perception Annual Merged File Total Score Distribution

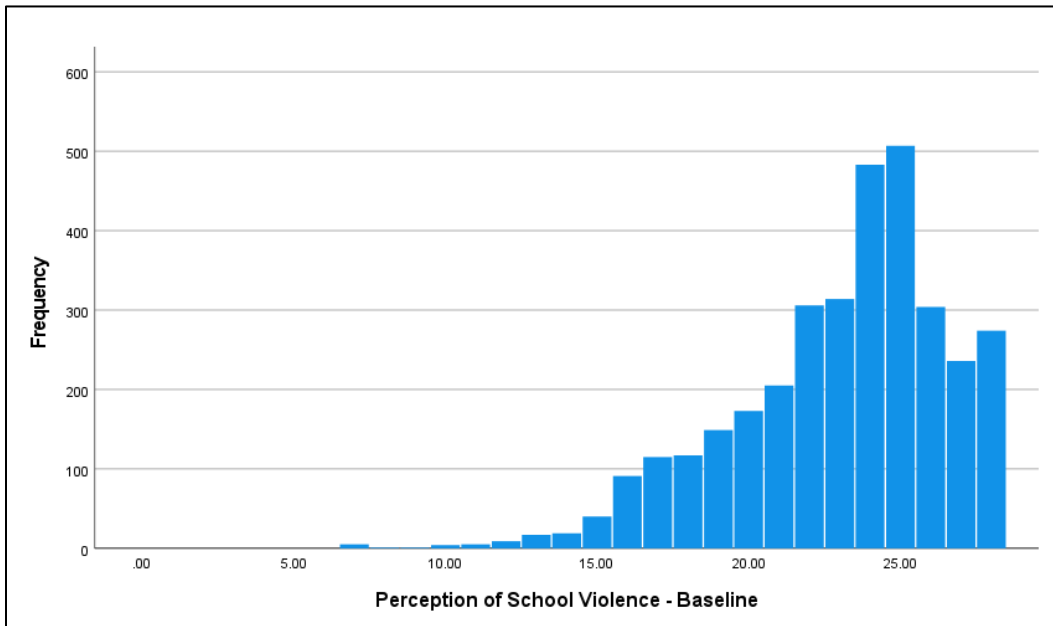


Figure 28. Girls' Survey Perception of School Violence Baseline Total Score Distribution

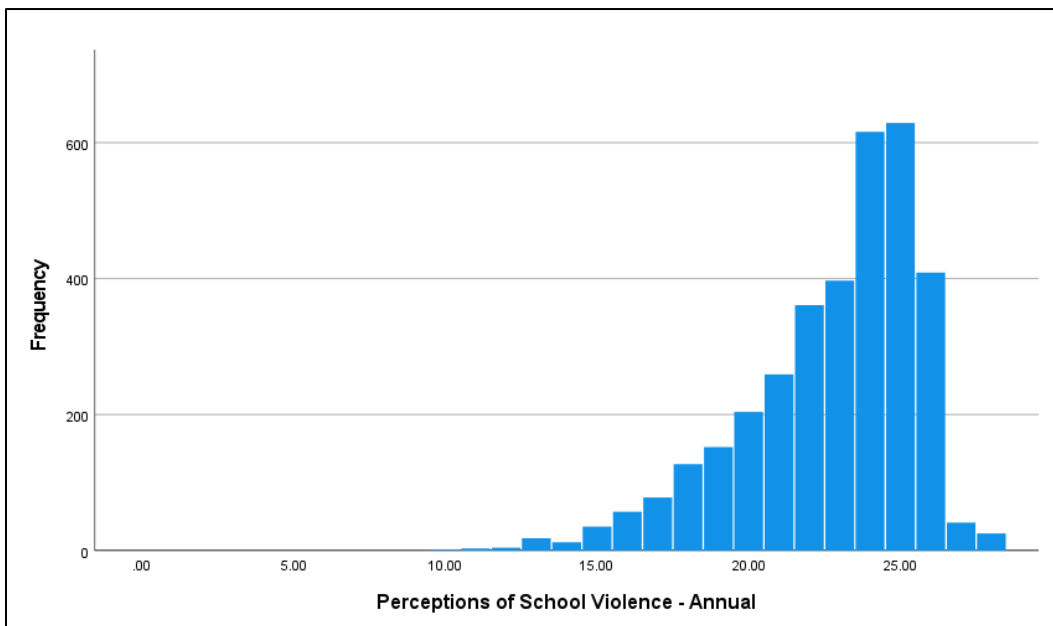


Figure 29. Girls' Survey Perception of School Violence Annual Total Score Distribution

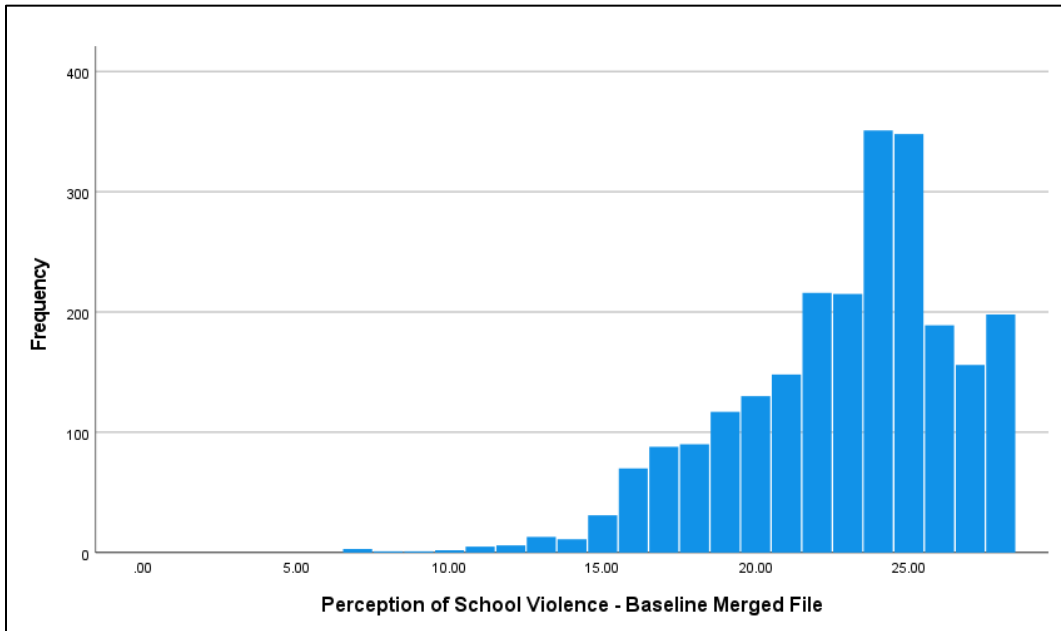


Figure 30. Girls' Survey Perception of School Violence Baseline Merged File Total Score Distribution

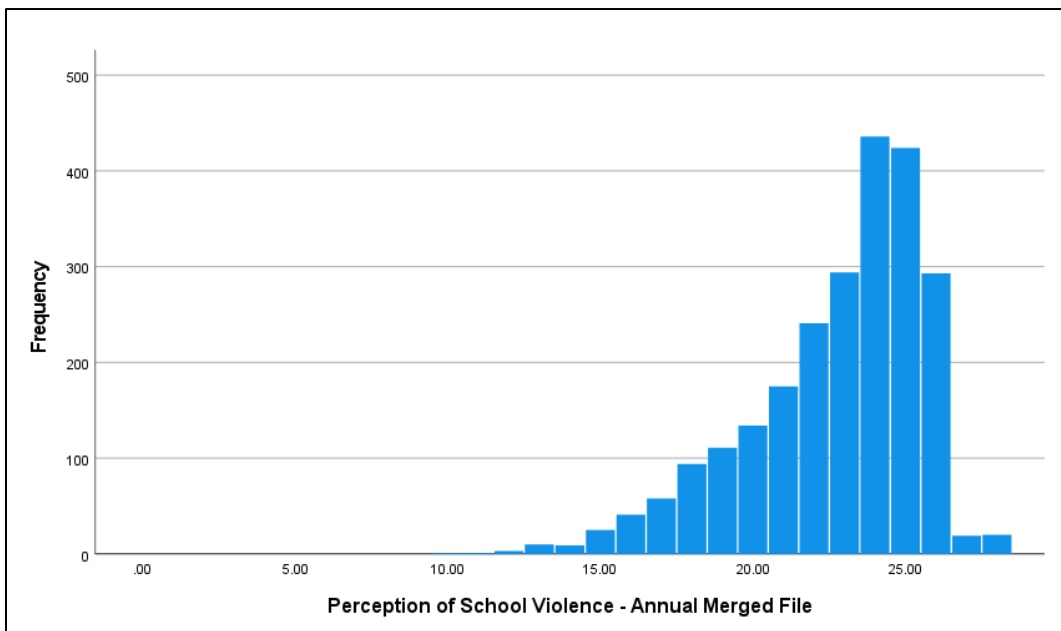


Figure 31. Girls' Survey Perception of School Violence Annual Merged File Total Score Distribution

Generalizability Theory

Determining Datasets for Analysis

Given the investigative and descriptive nature of this study, and the complexity of the sampling design used in the DRC evaluation, a fully-crossed design was created by taking samples of data from the datasets described in the previous section. For the GTheory analyses, data were selected such that there were:

- 1) at least two enumerators who administered the survey in the same language,
- 2) at least two languages administered by the same enumerator, or
- 3) a combination of 1) and 2).

The data were selected such that it was possible to use enumerator and/or language as a facet in all analyses. For the surveys and the Early Grades Mathematics Assessment (EGMA), the Language facet is interpreted as language of administration, and the enumerators administered the survey items and the EGMA in the girls' preferred languages. However, for the Early Grades Reading Assessment (EGRA), the assessment was completed in French. Therefore, the Language facet is interpreted as the girls' preferred language rather than the administered language.

As noted previously, adaptation of the surveys and EGMA was not completed prior to administration of the Baseline surveys, and thus, language was identified as a particular interest to the researchers. In addition, due to the informal nature of the survey adaptation in the field, enumerators were chosen as a facet in order to attempt to identify possible problems in training or translation. All languages and enumerators were evaluated for inclusion in the analyses, and only those who completed administrations in at least two languages were selected. Next, any

enumerators who completed fewer than 10 interviews in any of the multiple languages were removed from analysis.

It should also be noted that unlike more typical GTheory analyses, the focus here was on enumerators, language, and items included in the assessments and the survey, and the person facet was not modeled to allow for a crossed design. Including language as a facet generally necessitates a nested model, not allowing for a thorough investigation of each facet individually. The decision to exclude the person facet also means we cannot calculate the typical GTheory reliability coefficients as they require variance estimates specific to the person.

Analytic Procedure and Interpretation

The SPSS (IBM Corp., 2019) VARCOMP procedures was used to carry out all variance component estimation. Because all facets were considered to be random, the ANOVA method was implemented which is the simplest and most straightforward estimation procedure when dealing with random facets. A facet is considered to be random if we would think of the levels (enumerators, languages, or items, in this case) as a sample of a universe of possible levels, and we are attempting to generalize to said universe.

Conversely, a fixed facet is one in which data have been gathered on all levels of interest of the facet and there is no desire to generalize beyond the levels chosen. The VARCOMP ANOVA procedure produces a variance component estimate for each main effect (enumerator, language, items) and each interaction effect (enumerator \times language, enumerator \times items, language \times items, enumerator \times language \times items, error) in the model.

The estimated variance components associated with each main and interaction effect are reflective of “the magnitude of error in generalizing from a person’s score on a single item to his

or her universe score (the person's average over all items in the universe)" (Shavelson & Webb, 1991, p. 30). For example, if we consider the estimated variance component for items, this value is based on the variability of each item mean around the grand mean (the mean of means). This is intuitive if we recall that variance is the sum of all squared deviations from the mean. However, because the scale of variance is dependent on the scale used for the measure, it is not easily interpretable. Thus, in interpreting variance components, it is useful to look at the proportion of variance accounted for by a particular effect in relation to the total variance in the model.

In order to provide support of measurement invariance of the scale across enumerators and/or languages, we will assess the following relationships between the facets:

1) A small proportion of variability accounted for by enumerator provides evidence that responses are not specific to the enumerator that administered the items.

2) A small proportion of variability accounted for by the language of administration/girls' home language provides evidence that the responses to items is not language-specific.

3) A small proportion of variability accounted for by an interaction between enumerator and any other facet (i.e., item/subtest and/or language) provides evidence that the language of administration/girls' home language does not result in differential responses across items.

Generalizability Analysis Results

Baseline - Objective Measures

Early Grades Reading Assessment (EGRA)

Table 15 contains the sample size used across the five analyses for the EGRA and EGMA tasks. Table 16 contains the results of the five analyses completed on the Letter Name task of the EGRA. In the Swahili x French analysis, 32.29% of the total variability in the model was attributable to the Item facet, suggesting that the items are not redundant. While the estimates for the Enumerator main effect, Language main effect, and the interaction effects are small, they were not zero. Aside from the Error term, the Enumerator main effect accounted for the second largest amount of variance at 5.73%, indicating that there is some variation across the six enumerators and their ratings. The next largest proportion of variance comes from the Item by Enumerator interaction effect (4.17%), indicating that enumerators appear to have been interacting with the items in such a way that regardless of the language of administration, there was a larger amount of variability in responses for some enumerators and not others. The Language main effect accounted for 3.13%, indicating a small difference across the two languages. The Item by Language, and Language by Enumerator interaction effects, accounted for just over 2.00% of the total variance. The large unexplained Error variance component accounting for 52.60% of the total variance shows that there may be factors of relevance that were not included in the model that are responsible for some amount of systematic error.

Table 15. Sample Size for Baseline EGRA - Letter Name Items

	Enumerator	Language	
<i>Analysis 1</i>		Swahili	French
	1	22	11
	2	31	15
	3	38	18
	4	28	20
	5	18	31
	6	26	14
	<i>Total</i>	<i>163</i>	<i>109</i>
<i>Analysis 2</i>		Swahili	Kilendu
	1	14	21
	2	22	21
	3	14	21
	<i>Total</i>	<i>50</i>	<i>63</i>
<i>Analysis 3*</i>		Swahili	Tshiluba
	1	12	72
<i>Analysis 4*</i>		Lingala	French
	1	33	22
<i>Analysis 5*</i>		Lingala	Kikongo
	1	20	52

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

Table 16. GTheory Results for Baseline EGRA – Letter Name Items

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili x French</i>	Item	0.062	32.29%
	Language	0.006	3.13%
	Enumerator	0.011	5.73%
	Item x Language	0.003	1.56%
	Item x Enumerator	0.008	4.17%
	Language x Enumerator	0.001	0.52%
	Item x Enumerator x Language, Error	0.101	52.60%
	Total	0.192	100.00%
<i>Swahili x Kilendu</i>	Item	0.065	30.23%
	Language	0.010	4.65%
	Enumerator	0.002	0.93%
	Item x Language	0.002	0.93%

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
	Item x Enumerator	0.001	0.47%
	Language x Enumerator	0.005	2.33%
	Item x Enumerator x Language, Error	0.130	60.47%
	Total	0.215	100.00%
<i>Swahili x Tshiluba*</i>	Item	0.029	23.58%
	Language	0.002	1.63%
	Item x Language, Error	0.092	74.80%
	Total	0.123	100.00%
<i>Lingala x French*</i>	Item	0.045	24.86%
	Language	0.031	17.13%
	Item x Language, Error	0.105	58.01%
	Total	0.181	100.00%
<i>Lingala x Kikongo*</i>	Item	0.035	28.00%
	Language	0.000	0.00%
	Item x Language, Error	0.090	72.00%
	Total	0.125	100.00%

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

In the Swahili x Kilendu analysis, again the largest proportion of variance accounted for outside error was by the Item main effect (30.23%). The Language main effect was the second largest at 4.65%, followed by the Enumerator by Language interaction effect with 2.33% of the total variance. The Enumerator main effect, Item by Language interaction effect, and Language by Enumerator interaction effect accounted for the smallest non-zero variance with 0.93%, 0.93%, and 0.47% respectively. Again, the Error variance accounted for a large proportion of variability at 60.47%.

The Swahili x Tshiluba, Lingala x French, and Lingala x Kikongo analyses did not include the Enumerator facet as only one enumerator administered enough surveys in multiple languages. In the Swahili x Tshiluba analysis, as with previous analyses, the Item main effect

accounted for the largest portion of variability with 23.58% of the total. The Language main effect accounted for a small portion of the total variance at 1.63%, and Error variability in this analysis accounted for 74.80%.

In the Lingala *x* French analysis, the Item main effect accounted for 24.86% of the total variance, and the Language main effect accounted for 17.13% indicating some difference across the two languages. The Error was slightly smaller accounting for 58.01% of the total variance. In the Lingala *x* Kikongo analysis, the Item main effect (28.00%) was the only facet other than Error that accounted for any of the variance in the model with 72.00%.

Table 17. GTheory Results for Baseline EGRA – Non-Word Reading Items

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili x French</i>	Item	0.033	20.12%
	Language	0.006	3.66%
	Enumerator	0.011	6.71%
	Item x Language	0.003	1.83%
	Item x Enumerator	0.009	5.49%
	Language x Enumerator	0.001	0.61%
	Item x Enumerator x Language, Error	0.101	61.59%
	Total	0.164	100.00%
<i>Swahili x Kilendu</i>	Item	0.030	15.38%
	Language	0.003	1.54%
	Enumerator	0.000*	0.00%
	Item x Language	0.001	0.51%
	Item x Enumerator	0.000*	0.00%
	Language x Enumerator	0.014	7.18%
	Item x Enumerator x Language, Error	0.147	75.38%
	Total	0.195	100.00%
<i>Swahili x Tshiluba*</i>	Item	0.008	10.00%
	Language	0.001	1.25%
	Item x Language, Error	0.071	88.75%

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
	Total	0.080	100.00%
<i>Lingala x French*</i>	Item	0.018	15.52%
	Language	0.014	12.07%
	Item x Language, Error	0.084	72.41%
	Total	0.116	100.00%
<i>Lingala x Kikongo*</i>	Item	0.007	9.21%
	Language	0.000	0.00%
	Item x Language, Error	0.069	90.79%
	Total	0.076	100.00%

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

Table 17 contains the results of the five analyses completed on the Non-Word Reading task of the EGRA. In the Swahili x French analysis, 20.12% of the total variability in the model was attributable to the Item main effect, lower than expected, suggesting that the items may be redundant. Aside from the Error term, the Enumerator main effect accounted for the second largest amount of variance at 6.71%, indicating that there is some variation across the six enumerators. The next largest proportion is the Item by Enumerator interaction effect (5.49%) indicating that enumerators appear to have been interacting with the items in such a way that regardless of the language of administration, there was a larger amount of variability in responses for some enumerators and not others. The Language main effect accounted for 3.66%, indicating a small difference across the two languages. The Item by Language and Language by Enumerator interaction effects accounted for just over 2.00% of the total variance. The large unexplained Error variance component accounting for 61.59% of the total variance shows that there may be factors of relevance that were not included in the model that are responsible for some amount of systematic error.

In the Swahili \times Kilendu analysis, again the proportion of variance accounted by the Item main effect (15.38%) was lower than expected. The Language by Enumerator interaction effect was the second largest at 7.18%, followed by the Language main effect (1.54%) and the Item by Language interaction effect (0.51%). The Enumerator main effect and Item by Enumerator interaction effect was set to zero due to negative estimates. Again, the Error variance component accounted for a large proportion of variability at 75.38%.

In the Swahili \times Tshiluba analysis, as with previous analyses, the Item main effect accounted for a smaller than expected proportion of total variance at 10.00%. The Language main effect also accounted for a small portion of the total variance at 1.25%, and Error variability in this analysis accounted for 88.75%.

In the Lingala \times French analysis, the Item main effect accounted for only 15.52% of the total variance, and the Language main effect accounted 12.07% of the total variance, indicating some difference across the two languages. The Error accounted for 72.41% of the total variance. In the Lingala \times Kikongo analysis, the Item main effect (9.21%) was the only facet other than Error that accounted for any of the variance in the model with 90.79%.

Table 18. GTheory Results for Baseline EGRA – Oral Reading Fluency Items

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili \times French</i>	Item	0.014	5.88%
	Language	0.019	7.98%
	Enumerator	0.027	11.34%
	Item \times Language	0.000	0.00%
	Item \times Enumerator	0.003	1.26%
	Language \times Enumerator	0.004	1.68%
	Item \times Enumerator \times Language, Error	0.171	71.85%
	Total	0.238	100.00%

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili x Kilendu</i>	Item	0.015	6.76%
	Language	0.001	0.45%
	Enumerator	0.000*	0.00%
	Item x Language	0.000*	0.00%
	Item x Enumerator	0.000*	0.00%
	Language x Enumerator	0.019	8.56%
	Item x Enumerator x Language, Error	0.187	84.23%
	Total	0.222	100.00%
<i>Swahili x Tshiluba*</i>	Item	0.004	4.94%
	Language	0.001	1.23%
	Item x Language, Error	0.076	93.83%
	Total	0.081	100.00%
<i>Lingala x French*</i>	Item	0.000*	0.00%
	Language	0.010	14.71%
	Item x Language, Error	0.058	85.29%
	Total	0.068	100.00%
<i>Lingala x Kikongo*</i>	Item	0.004	5.06%
	Language	0.000	0.00%
	Item x Language, Error	0.075	94.94%
	Total	0.079	100.00%

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

Table 18 contains the results of the five analyses completed on the Oral Reading Fluency task of the EGRA. In the Swahili x French analysis, the Item main effect accounted for 5.88%, lower than expected, suggesting that the items may be redundant or too difficult. Aside from the Error term, the Enumerator main effect accounted for the second largest amount of variance at 11.31%, indicating that there is some variation across the six enumerators. The next largest proportion is the Language main effect (7.98%), indicating variation across the languages. The Item by Enumerator and Language by Enumerator interaction effects are lower at 1.26% and 1.68% respectively. The large unexplained Error variance component accounting for 71.85% of

the total variance shows that there may be factors of relevance that were not included in the model that are responsible for some amount of systematic error.

In the Swahili \times Kilendu analysis, again the proportion of variance accounted by the Item main effect (6.76%) was lower than expected. The Language by Enumerator interaction effect was the second largest at 8.56%, followed by the Language main effect (0.45%). The Enumerator main effect, Item by Language, and Item by Enumerator interaction components were set to zero due to negative estimates. Again, the error variance component accounted for a large proportion of variability at 84.23%.

In the Swahili \times Tshiluba analysis, as with previous analyses, the Item main effect accounted a smaller than expected proportion of total variance at 4.94%. The Language effect also accounted for a small portion of the total variance at 1.23%, and Error variability in this analysis accounted for 93.83%.

In the Lingala \times French analysis, the Item main effect was set to zero due to a negative estimate. The Language main effect accounted for 14.71% of the total variance, indicating some difference across the two languages. The error accounted for 72.41% of the total variance. In the Lingala \times Kikongo analysis, the Item main effect was the only facet other than error (5.06%) that accounted for any of the variance in the model with 94.94%.

Table 19. GTheory Results for Baseline EGRA –Reading Comprehension Items

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili \times French</i>	Item	0.011	6.83%
	Language	0.005	3.11%
	Enumerator	0.015	9.32%
	Item \times Language	0.000	0.00%
	Item \times Enumerator	0.004	2.48%
	Language \times Enumerator	0.003	1.86%

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
	Item x Enumerator x Language, Error	0.123	76.40%
	Total	0.161	100.00%
<i>Swahili x Kilendu</i>	Item	0.008	4.52%
	Language	0.000*	0.00%
	Enumerator	0.001	0.56%
	Item x Language	0.000*	0.00%
	Item x Enumerator	0.002	1.13%
	Language x Enumerator	0.007	3.95%
	Item x Enumerator x Language, Error	0.159	89.83%
	Total	0.177	100.00%
<i>Swahili x Tshiluba*</i>	Item	0.002	4.00%
	Language	0.000	0.00%
	Item x Language, Error	0.048	96.00%
	Total	0.050	100.00%
<i>Lingala x French*</i>	Item	0.000	0.00%
	Language	0.000	0.00%
	Item x Language, Error	0.004	100.00%
	Total	0.004	100.00%
<i>Lingala x Kikongo*</i>	Item	0.001	1.96%
	Language	0.000	0.00%
	Item x Language, Error	0.050	98.04%
	Total	0.051	100.00%

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

Table 19 contains the results of the five analyses completed on the Reading Comprehension task of the EGRA. In the Swahili x French analysis, results for the Item main effect were similarly low. Results showed that only 6.83% of the total variability in the model was attributable to the Item main effect, lower than expected, suggesting that the items may be redundant or too difficult. Aside from the error term, the Enumerator main effect accounted for the second largest amount of variance at 9.32%, indicating that there is some variation across the

six enumerators. The next largest proportion is the Language main effect (3.11%), indicating slight variation across the languages. The interaction effects between Item and Enumerator, and Language and Enumerator are lower at 2.48% and 1.86% respectively. The large unexplained error variance component accounting for 76.40% of the total variance shows that there may be factors of relevance that were not included in the model that are responsible for some amount of systematic error.

In the Swahili \times Kilendu analysis, again the proportion of variance accounted by the Item main effect (4.52%) was lower than expected. The Language by Enumerator interaction effect was the second largest at 3.95%, followed by the Item by Enumerator interaction effect (1.13%) and the Enumerator effect (0.56%). The Language main effect and Item by Language interaction component were set to zero due to negative estimates. Again, the error variance component accounted for a large proportion of variability at 89.83%.

In the Swahili \times Tshiluba analysis, the Item main effect was the only facet other than error (4.00%) that accounted for any of the variance in the model with 96.00%. In the Lingala \times French analysis, the error variance accounted for all estimated variance. In the Lingala \times Kikongo analysis, the Item main effect was the only facet other than error (1.96%) that accounted for any of the variance in the model with 98.04%.

Table 20. GTheory Results for Baseline EGRA –Listening Comprehension Items

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili \times French</i>	Item	0.017	8.85%
	Language	0.001	0.52%
	Enumerator	0.023	11.98%
	Item \times Language	0.000	0.00%
	Item \times Enumerator	0.012	6.25%
	Language \times Enumerator	0.002	1.04%

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
	Item x Enumerator x Language, Error	0.137	71.35%
	Total	0.192	100.00%
<i>Swahili x Kilendu</i>	Item	0.031	13.19%
	Language	0.013	5.53%
	Enumerator	0.004	1.70%
	Item x Language	0.001	0.43%
	Item x Enumerator	0.006	2.55%
	Language x Enumerator	0.002	0.85%
	Item x Enumerator x Language, Error	0.178	75.74%
	Total	0.235	100.00%
<i>Swahili x Tshiluba*</i>	Item	0.000	0.00%
	Language	0.012	9.09%
	Item x Language, Error	0.120	90.91%
	Total	0.132	100.00%
<i>Lingala x French*</i>	Item	0.004	6.15%
	Language	0.000	0.00%
	Item x Language, Error	0.061	93.85%
	Total	0.065	100.00%
<i>Lingala x Kikongo*</i>	Item	0.001	1.23%
	Language	0.000*	0.00%
	Item x Language, Error	0.080	98.77%
	Total	0.081	100.00%

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

Table 20 contains the results of the five analyses completed on the Listening Comprehension task of the EGRA. In the Swahili x French analysis, 8.85% of the total variability in the model was attributable to the Item main effect, lower than expected, suggesting that the items may be redundant or too difficult. Aside from the error term, the Enumerator main effect accounted for the second largest amount of variance at 11.98%, indicating that there is some variation across the six enumerators. The next largest proportion is the interaction between Item

and Enumerator (6.25%), indicating enumerators interact with items differentially. The Language effect and interaction Language and Enumerator are lower at 0.52% and 1.04% respectively. The large unexplained error variance component accounting for 71.35% of the total variance shows that there may be factors of relevance that were not included in the model that are responsible for some amount of systematic error.

In the Swahili \times Kilendu analysis, again the proportion of variance accounted by the Item main effect (13.19%) was lower than expected. The Enumerator effect was the second largest at 5.53%, followed by the Item by Enumerator (2.55%). The Enumerator main effect (1.70%), Language by Enumerator (0.85%), and Item by Language (0.43) interactions followed. Again, the error variance component accounted for a large proportion of variability at 75.74%.

In the Swahili \times Tshiluba analysis, as with previous analyses, the Item main effect was estimated to be zero, while the Language main effect accounted for 9.09% of the total variance, indicating some difference across the two languages. Error variability in this analysis accounted for 90.91% of the total.

In the Lingala \times French analysis, the Item main effect was the only main effect other than error (6.15%) that accounted for any of the variance in the model with 93.85%. Similarly, in the Lingala \times Kikongo analysis, the Item main effect accounted for a small amount of variance (1.23%), with the error accounting for the majority (98.77%).

Early Grades Mathematics Assessment

Table 21. GTheory Results for Baseline EGMA –Number Identification Items

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili x French</i>	Item	0.017	9.55%
	Language	0.010	5.62%
	Enumerator	0.011	6.18%
	Item x Language	0.002	1.12%
	Item x Enumerator	0.006	3.37%
	Language x Enumerator	0.002	1.12%
	Item x Enumerator x Language, Error	0.130	73.03%
	Total	0.178	100.00%
<i>Swahili x Kilendu</i>	Item	0.008	7.62%
	Language	0.000*	0.00%
	Enumerator	0.000*	0.00%
	Item x Language	0.001	0.95%
	Item x Enumerator	0.000*	0.00%
	Language x Enumerator	0.012	11.43%
	Item x Enumerator x Language, Error	0.084	80.00%
	Total	0.105	100.00%
<i>Swahili x Tshiluba*</i>	Item	0.015	10.64%
	Language	0.001	0.71%
	Item x Language, Error	0.125	88.65%
	Total	0.141	100.00%
<i>Lingala x French*</i>	Item	0.042	18.18%
	Language	0.039	16.88%
	Item x Language, Error	0.150	64.94%
	Total	0.231	100.00%
<i>Lingala x Kikongo*</i>	Item	0.037	20.33%
	Language	0.005	2.75%
	Item x Language, Error	0.140	76.92%
	Total	0.182	100.00%

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

Table 21 **Error! Reference source not found.** contains the results of the five analyses completed on the Number Identification task of the EGMA. In the Swahili \times French analysis, 9.55% of the total variability in the model was attributable to the Item main effect, lower than expected, suggesting that the items may be redundant or too difficult. Aside from the error term, the Enumerator main effect accounted for the second largest amount of variance at 6.18%, indicating that there is some variation across the six enumerators. The next largest proportion is the Language main effect (5.62%), indicating variation across the languages. The interaction effects between Item and Language, Item and Enumerator, and Language and Enumerator are lower at 1.12%, 3.37%, and 1.12% respectively. The large unexplained error variance component accounting for 73.03% of the total variance shows that there may be factors of relevance that were not included in the model that are responsible for some amount of systematic error. In the Swahili \times Kilendu analysis, aside from error, only the Item main effect (7.62%) and the Item by Language interaction effect (0.95%) had non-zero estimates. Again, the error variance component accounted for a large proportion of variability at 80.00%.

In the Swahili \times Tshiluba analysis, as with previous analyses, the Item main effect accounted a smaller than expected proportion of total variance at 4.94%. The Language effect also accounted for a small portion of the total variance at 1.23%, and Error variability in this analysis accounted for 93.83%.

In the Lingala \times French analysis, the Item main effect was larger than previous analyses at 18.18%, and Language accounted for 16.88% indicating some difference across the languages. The Error accounted for 64.94% of the total variance. In the Lingala \times Kikongo analysis, the Item main effect was the largest for this analysis set at 20.33%, and Language accounted for 2.75%. The Error accounted for 76.92% of the total variance.

Table 22. GTheory Results for Baseline EGMA –Number Discrimination Items

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili x French</i>	Item	0.007	3.66%
	Language	0.006	3.14%
	Enumerator	0.023	12.04%
	Item x Language	0.002	1.05%
	Item x Enumerator	0.008	4.19%
	Language x Enumerator	0.006	3.14%
	Item x Enumerator x Language, Error	0.139	72.77%
	Total	0.191	100.00%
<i>Swahili x Kilendu</i>	Item	0.004	6.78%
	Language	0.000	0.00%
	Enumerator	0.000	0.00%
	Item x Language	0.000	0.00%
	Item x Enumerator	0.000	0.00%
	Language x Enumerator	0.001	1.69%
	Item x Enumerator x Language, Error	0.054	91.53%
	Total	0.059	100.00%
<i>Swahili x Tshiluba*</i>	Item	0.024	13.79%
	Language	0.004	2.30%
	Item x Language, Error	0.146	83.91%
	Total	0.174	100.00%
<i>Lingala x French*</i>	Item	0.029	11.15%
	Language	0.033	12.69%
	Item x Language, Error	0.198	76.15%
	Total	0.260	100.00%
<i>Lingala x Kikongo*</i>	Item	0.032	15.46%
	Language	0.002	0.97%
	Item x Language, Error	0.173	83.57%
	Total	0.207	100.00%

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

Error! Reference source not found. Table 22 contains the results of the five analyses completed on the Number Discrimination task of the EGMA. In the Swahili x French analysis,

only 3.66% of the total variability in the model was attributable to the Item main effect. Aside from the error term, the Enumerator main effect accounted for the largest amount of variance at 12.04%, indicating that there is some variation across the six enumerators. The next largest proportion is the from the Item by Enumerator main effect (4.19%), indicating that enumerators appear to have been interacting with the items in such a way that regardless of the language of administration, there was a larger amount of variability in responses for some enumerators and not others. The Language main effect and Language by Enumerator interaction facets both account for 3.14% of the total variance, and finally, the Item by Language main effect accounted for 1.05%. The large unexplained error variance component accounting for 72.77% of the total variance shows that there may be factors of relevance that were not included in the model that are responsible for some amount of systematic error.

In the Swahili \times Kilendu analysis, only the Item main effect (6.78%) and the Language by Enumerator interaction effect (1.69%) had non-zero estimates. The majority of the variance was attributable to the Error term (91.53%).

In the Swahili \times Tshiluba analysis, the Item main effect accounted for only 13.79% of total variance. The Language effect also accounted for a small portion of the total variance at 2.30%, and Error variance in this analysis accounted for 83.91%.

In the Lingala \times French analysis, the Item main effect accounted for 11.15% of total variance, and the Language main effect accounted for 12.69% of the total variance, indicating some difference across the two languages. The error accounted for 76.15% of the total variance. In the Lingala \times Kikongo analysis, the Item main effect accounted for 15.46% of total variance, and the Language main effect accounted for 0.97%. The Error accounted for 83.57% of the total variance.

Table 23. GTheory Results for Baseline EGMA –Missing Number Items

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili x French</i>	Item	0.048	18.11%
	Language	0.001	0.38%
	Enumerator	0.041	15.47%
	Item x Language	0.000	0.00%
	Item x Enumerator	0.024	9.06%
	Language x Enumerator	0.001	0.38%
	Item x Enumerator x Language, Error	0.150	56.60%
	Total	0.265	100.00%
<i>Swahili x Kilendu</i>	Item	0.079	30.38%
	Language	0.000*	0.00%
	Enumerator	0.016	6.15%
	Item x Language	0.003	1.15%
	Item x Enumerator	0.025	9.62%
	Language x Enumerator	0.013	5.00%
	Item x Enumerator x Language, Error	0.124	47.69%
	Total	0.260	100.00%
<i>Swahili x Tshiluba*</i>	Item	0.133	58.59%
	Language	0.003	1.32%
	Item x Language, Error	0.091	40.09%
	Total	0.227	100.00%
<i>Lingala x French*</i>	Item	0.113	64.57%
	Language	0.003	1.71%
	Item x Language, Error	0.059	33.71%
	Total	0.175	100.00%
<i>Lingala x Kikongo*</i>	Item	0.134	64.42%
	Language	0.001	0.48%
	Item x Language, Error	0.073	35.10%
	Total	0.208	100.00%

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

Table 23Error! Reference source not found. contains the results of the five analyses completed on the Missing Number task of the EGMA. In the Swahili x French analysis, 18.11%

of the total variability in the model was attributable to the Item main effect. Aside from the error term, the Enumerator main effect accounted for the second largest amount of variance at 15.47%, indicating that there is some variation across the six enumerators. The next largest proportion is the Item by Enumerator interaction effect (9.06%). The Language effect and Language by Enumerator interaction effect each accounted for a small proportion of total variance at 0.38%. The large unexplained error variance component accounting for 56.60% of the total variance shows that there may be factors of relevance that were not included in the model that are responsible for some amount of systematic error.

In the Swahili \times Kilendu analysis, the proportion of variance accounted by the Item main effect (30.38%) indicated a lack of redundancy for the task. The Item by Enumerator interaction effect was the second largest at 9.62%, followed by the Enumerator main effect (6.15%) and the Language by Enumerator interaction effect (5.00%). The Item by Language interaction effect accounted for a small portion of variance at 1.15% of the total, and the Language effect estimate was set to zero due to negative estimates. The Error variance component accounted for a large proportion of variability at 47.69%.

In the Swahili \times Tshiluba analysis, as with previous analyses, the Item main effect accounted for a large proportion of total variance at 64.57%. The Language effect accounted for a small portion of the total variance at 1.71%, and Error variability in this analysis accounted for 40.09%.

In the Lingala \times French analysis, the Item main effect accounted for the majority of total variance at 64.57%. The Language main effect accounted for only 1.71% of the total variance, and the Error accounted for 33.71% of the total variance. Similar to the Lingala \times French analysis, the Lingala \times Kikongo analysis showed that the Item main effect accounted for the majority of

total variance at 64.42%. The Language main effect accounted for only 0.48% of the total variance, and the Error accounted for 35.10% of the total variance.

Table 24. GTheory Results for Baseline EGMA –Addition Items

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili x French</i>	Item	0.072	29.51%
	Language	0.006	2.46%
	Enumerator	0.006	2.46%
	Item x Language	0.000	0.00%
	Item x Enumerator	0.011	4.51%
	Language x Enumerator	0.003	1.23%
	Item x Enumerator x Language, Error	0.146	59.84%
	Total	0.244	100.00%
<i>Swahili x Kilendu</i>	Item	0.099	38.98%
	Language	0.000*	0.00%
	Enumerator	0.000*	0.00%
	Item x Language	0.000*	0.00%
	Item x Enumerator	0.001	0.39%
	Language x Enumerator	0.010	3.94%
	Item x Enumerator x Language, Error	0.144	56.69%
	Total	0.254	100.00%
<i>Swahili x Tshiluba*</i>	Item	0.095	37.70%
	Language	0.002	0.79%
	Item x Language, Error	0.155	61.51%
	Total	0.252	100.00%
<i>Lingala x French*</i>	Item	0.102	38.64%
	Language	0.025	9.47%
	Item x Language, Error	0.137	51.89%
	Total	0.264	100.00%
<i>Lingala x Kikongo*</i>	Item	0.110	46.03%
	Language	0.001	0.42%
	Item x Language, Error	0.128	53.56%
	Total	0.239	100.00%

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

Table 24 contains the results of the five analyses completed on the Addition task of the EGMA. In the Swahili \times French analysis, 29.51% of the total variability in the model was attributable to the Item main effect, indicating a lack of redundancy in the item set. Aside from the error term, the Item by Enumerator interaction effect accounted for the second largest amount of variance at 4.51%. The next largest proportion was the Language and Enumerator main effects, each accounting for 2.46% of the total variance, indicating there may be a small difference across both languages and enumerators. The interaction effect between Language and Enumerator was lower at 1.23%. The large unexplained error variance component accounting for 59.84% of the total variance shows that there may be factors of relevance that were not included in the model that are responsible for some amount of systematic error.

In the Swahili \times Kilendu analysis, again the proportion of variance accounted by the Item main effect (38.98%) was as expected. The Language by Enumerator interaction effect was the second largest at 0.39%. The Language and Enumerator main effects and the Item by Language interaction components were set to zero due to negative estimates. The error variance component accounted for a large proportion of variability at 56.69%.

In the Swahili \times Tshiluba analysis, as with previous analyses, the Item main effect accounted for 37.70% of total variance. The Language effect accounted for a small portion of the total variance at 0.79%, and Error variability in this analysis accounted for 61.51%.

In the Lingala \times French analysis, the Item main effect accounted for 38.65% of the total variance. The Language main effect accounted for 9.47% of the total variance, indicating some difference across the two languages. Error accounted for 51.89% of the total variance. In the Lingala \times Kikongo analysis, the Item main effect accounted for nearly half (46.03%) of the total

variance. The Language main effect accounted for only 0.42% of the total variance, and Error accounted for 53.56% of the total variance.

Table 25. GTheory Results for Baseline EGMA –Subtraction Items

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili x French</i>	Item	0.068	26.05%
	Language	0.007	2.68%
	Enumerator	0.006	2.30%
	Item x Language	0.000	0.00%
	Item x Enumerator	0.012	4.60%
	Language x Enumerator	0.003	1.15%
	Item x Enumerator x Language, Error	0.165	63.22%
	Total	0.261	100.00%
<i>Swahili x Kilendu</i>	Item	0.081	31.64%
	Language	0.002	0.78%
	Enumerator	0.000*	0.00%
	Item x Language	0.000*	0.00%
	Item x Enumerator	0.004	1.56%
	Language x Enumerator	0.015	5.86%
	Item x Enumerator x Language, Error	0.154	60.16%
	Total	0.256	100.00%
<i>Swahili x Tshiluba*</i>	Item	0.034	17.17%
	Language	0.001	0.51%
	Item x Language, Error	0.163	82.32%
	Total	0.198	100.00%
<i>Lingala x French*</i>	Item	0.036	17.82%
	Language	0.034	16.83%
	Item x Language, Error	0.132	65.35%
	Total	0.202	100.00%
<i>Lingala x Kikongo*</i>	Item	0.025	16.13%
	Language	0.000	0.00%
	Item x Language, Error	0.130	83.87%
	Total	0.155	100.00%

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

Table 25 contains the results of the five analyses completed on the Subtraction task of the EGMA. In the Swahili \times French analysis, 26.05% of the total variability in the model was attributable to the Item main effect, indicating a lack of redundancy in the item set. Aside from the error term, the Item by Enumerator interaction effect accounted for the second largest amount of variance at 4.60%. The next largest proportion was the Language main effect (2.68%), then the Enumerator main effect (2.30%), and finally the Language by Enumerator interaction effect (1.15%). The large unexplained error variance component accounting for 63.22% of the total variance shows that there may be factors of relevance that were not included in the model that are responsible for some amount of systematic error.

In the Swahili \times Kilendu analysis, again the proportion of variance accounted by the Item main effect (31.64%) was as expected. The Language by Enumerator interaction effect was the second largest at 5.86%, followed by the Item by Enumerator interaction effect (1.56%), and the Language main effect (0.78%). The Enumerator main effect and the Item by Language interaction effect were set to zero due to negative estimates. The error variance component accounted for a large proportion of variability at 60.16%.

In the Swahili \times Tshiluba analysis, as with previous analyses, the Item main effect accounted for a smaller proportion of variance than previous analyses (17.17%). The Language effect accounted for a small portion of the total variance at 0.51%, and Error variability in this analysis accounted for 82.32%.

In the Lingala \times French analysis, the Item main effect accounted for 17.82% of the total variance, and the Language main effect accounted for 16.83% of the total variance, indicating some difference across the two languages. The error accounted for 65.35% of the total variance.

In the Lingala x Kikongo analysis, the Item main effect was the only facet other than error (16.13%) that accounted for any of the variance in the model with 83.87%.

Baseline – Subjective Measures

Table 26 shows the sample sizes used for the analyses completed on the General School Perception set of survey items administered at Baseline. Table 27 contains the results of the five analyses completed. In the Swahili x French analysis, 34.41% of the total variability in the model was attributable to the Item main effect, indicating a lack of redundancy in the item set. Aside from the error term, the Item by Enumerator interaction effect accounted for the second largest amount of variance at 23.18%. That is, enumerators appear to have been interacting with the items in such a way that regardless of the language of administration, there was a larger amount of variability in responses for some enumerators and not others. The next largest proportion was for the Enumerator main effect (1.53%), then the Language by Enumerator interaction effect (0.81%), and finally the Item by Language interaction effect (0.45%). The large unexplained error variance component accounting for 39.62% of the total variance shows that there may be factors of relevance that were not included in the model that are responsible for some amount of systematic error.

In the Swahili x Kilendu analysis, the proportion of variance accounted by the Item main effect was lower than the first analysis at 14.27%. The Item by Enumerator interaction effect was also the second largest in this analysis at 34.83%, followed by the Language by Enumerator interaction effect (0.80%), and the Language main effect (0.50%). The Enumerator main effect and the Item by Language interaction effect were set to zero due to negative estimates. The error variance component accounted for a large proportion of variability at 49.60%.

Table 26. Sample Size for Baseline General School Perception Survey Items

	Enumerator	Language	
<i>Analysis 1</i>		Swahili	French
	1	22	11
	2	31	15
	3	36	14
	4	15	13
	5	21	11
	<i>Total</i>	<i>125</i>	<i>64</i>
<i>Analysis 2</i>		Swahili	Kilendu
	1	14	21
	2	22	21
	3	14	21
	<i>Total</i>	<i>50</i>	<i>63</i>
<i>Analysis 3*</i>		Swahili	Tshiluba
	1	12	72
<i>Analysis 4*</i>		Lingala	French
	1	29	18
<i>Analysis 5*</i>		Lingala	Kikongo
	1	16	40

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

Table 27. GTheory Results for Baseline General School Perception Survey Items

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili x French</i>	Item	0.383	34.41%
	Language	0.000*	0.00%
	Enumerator	0.017	1.53%
	Item x Language	0.005	0.45%
	Item x Enumerator	0.258	23.18%
	Language x Enumerator	0.009	0.81%
	Item x Enumerator x Language, Error	0.441	39.62%
	Total	1.113	100.00%
<i>Swahili x Kilendu</i>	Item	0.143	14.27%
	Language	0.005	0.50%
	Enumerator	0.000*	0.00%
	Item x Language	0.000*	0.00%
	Item x Enumerator	0.349	34.83%

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
	Language x Enumerator	0.008	0.80%
	Item x Enumerator x Language, Error	0.497	49.60%
	Total	1.002	100.00%
<i>Swahili x Tshiluba</i> *	Item	0.235	19.81%
	Language	0.002	0.17%
	Item x Language, Error	0.949	80.02%
	Total	1.186	100.00%
<i>Lingala x French</i> *	Item	0.379	52.93%
	Language	0.000*	0.00%
	Item x Language, Error	0.337	47.07%
	Total	0.716	100.00%
<i>Lingala x Kikongo</i> *	Item	0.012	5.29%
	Language	0.000*	0.00%
	Item x Language, Error	0.215	94.71%
	Total	0.227	100.00%

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

In the Swahili x Tshiluba analysis, as with previous analyses, the Item main effect accounted for a slightly smaller than expected proportion of total variance at 19.81%. The Language effect also accounted for a small portion of the total variance at 0.17%, and Error variability in this analysis accounted for 80.02%.

In the Lingala x French analysis, the Item main effect attributed for over half of the total variance (52.93%), and the Language main effect was set to zero due to a negative estimate. The error accounted for 47.07% of the total variance. In the Lingala x Kikongo analysis, the Item main effect was the only main effect other than error (5.29%) that accounted for any of the variance in the model with 94.71%.

Table 28 **Error! Reference source not found.** shows the sample sizes used for the analyses completed on the Student Perception of Teacher set of survey items administered at

Baseline. Table 29 **Error! Reference source not found.** contains the results of the five analyses completed. In the Swahili \times French analysis, 24.31% of the total variability in the model was attributable to the Item main effect, indicating a lack of redundancy in the item set. Aside from the error term, the Enumerator main effect accounted for the second largest amount of variance at 12.10%, indicating that there are differences across enumerators even when language is not considered. The next largest proportion of variance is attributable to the Item by Enumerator interaction effect (10.66%). That is, enumerators appear to have been interacting with the items in such a way that regardless of the language of administration, there was a larger amount of variability in responses for some enumerators and not others. The next largest proportions were for the Language by Enumerator interaction effect (0.84%), and then the Item by Language interaction effect (0.48%). The error variance component accounted for a large proportion of variability at 51.62%.

Table 28. Sample Size for Baseline Teacher Perception Survey Items

	Enumerator	Language	
<i>Analysis 1</i>		Swahili	French
	1	22	11
	2	31	15
	3	38	18
	4	28	20
	5	18	31
	6	26	14
	<i>Total</i>	<i>163</i>	<i>109</i>
<i>Analysis 2</i>		Swahili	Kilendu
	1	14	21
	2	22	21
	3	14	21
	<i>Total</i>	<i>50</i>	<i>63</i>
<i>Analysis 3*</i>		Swahili	Tshiluba
	1	12	72
<i>Analysis 4*</i>		Lingala	French
	1	33	22
<i>Analysis 5*</i>		Lingala	Kikongo

	Enumerator	Language	
	1	20	52

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

In the Swahili \times Kilendu analysis, the proportion of variance accounted by the Item main effect was even larger at 44.41%. The Item by Enumerator interaction effect was also the second largest in this analysis at 7.19%, followed by the Item by Language interaction effect (1.83%), the Language main effect (1.26%), and the Language by Enumerator interaction effect (0.68%). The error variance component accounted for a large proportion of variability at 44.63%.

In the Swahili \times Tshiluba analysis, as with previous analyses, the Item main effect accounted for 29.56% of the total variance. The Language effect accounted for a small portion of the total variance at 0.11%, and Error variability in this analysis accounted for 52.38%.

In the Lingala \times French analysis, the Item main effect attributed for almost half of the total variance (47.53%), and the Language main effect was nearly zero, accounting for 0.09% of the variance. The error accounted for 52.38% of the total variance. In the Lingala \times Kikongo analysis, the Item main effect attributed for over half of the total variance (63.70%), and the Language main effect was nearly zero, accounting for 0.26% of the variance. The error accounted for 36.05% of the total variance.

Table 29. GTheory Results for Baseline Teacher Perception Survey Items

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili \times French</i>	Item	0.203	24.31%
	Language	0.000*	0.00%
	Enumerator	0.101	12.10%
	Item \times Language	0.004	0.48%
	Item \times Enumerator	0.089	10.66%
	Language \times Enumerator	0.007	0.84%

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
	Item x Enumerator x Language, Error	0.431	51.62%
	Total	0.835	100.00%
<i>Swahili x Kilendu</i>	Item	0.389	44.41%
	Language	0.011	1.26%
	Enumerator	0.000*	0.00%
	Item x Language	0.016	1.83%
	Item x Enumerator	0.063	7.19%
	Language x Enumerator	0.006	0.68%
	Item x Enumerator x Language, Error	0.391	44.63%
	Total	0.876	100.00%
<i>Swahili x Tshiluba*</i>	Item	0.277	29.56%
	Language	0.001	0.11%
	Item x Language, Error	0.659	70.33%
	Total	0.937	100.00%
<i>Lingala x French*</i>	Item	0.519	47.53%
	Language	0.001	0.09%
	Item x Language, Error	0.572	52.38%
	Total	1.092	100.00%
<i>Lingala x Kikongo*</i>	Item	0.493	63.70%
	Language	0.002	0.26%
	Item x Language, Error	0.279	36.05%
	Total	0.774	100.00%

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

Table 30 shows the sample sizes used for the analyses completed on the Student Perception of School Violence set of survey items administered at Baseline. Table 31 contains the results of the five analyses completed. In the Swahili x French analysis, 9.87% of the total variability in the model was attributable to the Item main effect, lower than expected. Aside from the error term, the Enumerator main effect accounted for the second largest amount of variance at 12.91%, indicating that there are differences across enumerators even when language is not

considered. The next largest proportion of variance is attributable to the Item by Enumerator interaction effect (2.91%). The Language main effect and Item by Language interaction effect both account for zero variance. The error variance component accounted for a large proportion of variability at 72.41%.

Table 30. Sample Size for Baseline School Violence Perception Survey Items

	Enumerator	Language	
<i>Analysis 1</i>		Swahili	French
	1	22	11
	2	31	15
	3	38	18
	4	28	20
	5	18	31
	6	26	14
	<i>Total</i>	<i>163</i>	<i>109</i>
<i>Analysis 2</i>		Swahili	Kilendu
	1	14	21
	2	22	21
	3	14	21
	<i>Total</i>	<i>50</i>	<i>63</i>
<i>Analysis 3*</i>		Swahili	Tshiluba
	1	12	72
<i>Analysis 4*</i>		Lingala	French
	1	33	22
<i>Analysis 5*</i>		Lingala	Kikongo
	1	20	52

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

In the Swahili x Kilendu analysis, the proportion of variance accounted by the Item main effect was also small at 9.87%. The Language by Enumerator interaction effect was also the second largest in this analysis at 4.42%, followed by the Enumerator main effect (1.01%), Item by Enumerator interaction effect (0.76%), and the Item by Language interaction effect (0.13%). The error variance component accounted for a large proportion of variability at 83.71%.

In the Swahili x Tshiluba analysis, the Item main effect accounted for a much larger proportion of total variance at 54.45%. The Language effect did not account for any variance, and Error variability in this analysis accounted for 45.55%.

In the Lingala x French analysis, the Item main effect attributed for the majority of the total variance (74.67%). The Language effect did not account for any variance, and Error variability in this analysis accounted for 45.55%. In the Lingala x Kikongo analysis, the Item main effect attributed to almost half of the total variance (47.35%), and the Language main effect was zero. The error accounted for 52.65% of the total variance.

Table 31. GTheory Results for Baseline School Violence Perception Survey Items

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili x French</i>	Item	0.078	9.87%
	Language	0.000*	0.00%
	Enumerator	0.102	12.91%
	Item x Language	0.000	0.00%
	Item x Enumerator	0.023	2.91%
	Language x Enumerator	0.015	1.90%
	Item x Enumerator x Language, Error	0.572	72.41%
	Total	0.790	100.00%
<i>Swahili x Kilendu</i>	Item	0.079	9.97%
	Language	0.000*	0.00%
	Enumerator	0.008	1.01%
	Item x Language	0.001	0.13%
	Item x Enumerator	0.006	0.76%
	Language x Enumerator	0.035	4.42%
	Item x Enumerator x Language, Error	0.663	83.71%
	Total	0.792	100.00%
<i>Swahili x Tshiluba*</i>	Item	0.599	54.45%
	Language	0.000*	0.00%
	Item x Language, Error	0.501	45.55%

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
	Total	1.100	100.00%
<i>Lingala x French*</i>	Item	1.026	74.67%
	Language	0.000*	0.00%
	Item x Language, Error	0.348	25.33%
	Total	1.374	100.00%
<i>Lingala x Kikongo*</i>	Item	0.501	47.35%
	Language	0.000	0.00%
	Item x Language, Error	0.557	52.65%
	Total	1.058	100.00%

*These analyses do not include Enumerator as a facet, as only one enumerator administered in the two indicated languages.

Annual – Subjective Measures

Table 32Error! Reference source not found. shows the sample sizes used for the analyses completed on the General School Perception set of survey items administered at the Annual evaluation. Table 33 contains the results of the analysis completed. In the Swahili x French analysis, the variability associated with the Item and Language main effects as well as the Language by Enumerator interaction effect was set to zero. The facet with the largest variance attributed to it was the Item by Enumerator interaction effect, with 55.69% of the total variance. This indicates that Enumerators were interacting with these items differentially. Next, the Enumerator main effect with 7.43%, indicating that there are differences across enumerators even when language is not considered. The Item by Language interaction effect attributed 1.02% of the total variance, and the Error attributed 35.86%.

Table 32. Sample Size for Annual General School Perception Survey Items

	Enumerator	Language	
<i>Analysis 1</i>		Swahili	French
	1	30	21
	2	21	25
	3	21	20

	<i>Total</i>	72	66
--	--------------	----	----

Table 33. GTheory Results for Annual General School Perception Survey Items

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili x French</i>	Item	0.000*	0.00%
	Language	0.000	0.00%
	Enumerator	0.051	7.43%
	Item x Language	0.007	1.02%
	Item x Enumerator	0.382	55.69%
	Language x Enumerator	0.000*	0.00%
	Item x Enumerator x Language, Error	0.246	35.86%
	Total	0.686	100.00%

Table 34**Error! Reference source not found.** shows the sample sizes used for the analyses completed on the Student Perception of Teacher set of survey items administered at the Annual evaluation. Table 35**Error! Reference source not found.** contains the results of the analysis completed. In the Swahili x French analysis, the variability associated with the Item main effect was largest, at 50.88%. Next largest, the Item by Enumerator interaction effect attributed 17.10%, indicating that Enumerators were interacting with these items differentially. The Language by Enumerator interaction effect (0.72%), Enumerator main effect (0.14%), and the Item by Language interaction effect (0.14%) were all small. Finally, the Language main effect was set to zero, and the Error accounted for 31.48% of the total variance.

Table 34. Sample Size for Annual Teacher Perception Survey Items

	Enumerator	Language	
<i>Analysis 1</i>		Swahili	French
	1	42	28
	2	48	10
	3	34	26
	4	14	11
	5	26	28
	<i>Total</i>	<i>164</i>	<i>103</i>

Table 35. GTheory Results for Annual Teacher Perception Survey Items

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili x French</i>	Item	0.375	50.88%
	Language	0.000	0.00%
	Enumerator	0.001	0.14%
	Item x Language	0.001	0.14%
	Item x Enumerator	0.126	17.10%
	Language x Enumerator	0.002	0.27%
	Item x Enumerator x Language, Error	0.232	31.48%
	Total	0.737	100.00%

Table 36 **Error! Reference source not found.** shows the sample sizes used for the analyses completed on the Student Perception of School Violence set of survey items administered at the Annual evaluation. Table 37 contains the results of the analysis completed. In the Swahili x French analysis, the variability associated with the Item main effect was largest, save for the Error, at 40.43%. Next largest, the Item by Enumerator interaction effect attributed 15.07%, indicating that Enumerators were interacting with these items differentially. The Item by Language interaction effect (1.58%) and Language by Enumerator interaction effect (0.24%) were small. Finally, Language and Enumerator main effects were set to zero, and the Error accounted for 31.48% of the total variance.

Table 36. Sample Size for Annual School Violence Perception Survey Items

	Enumerator	Language	
<i>Analysis 1</i>		Swahili	French
	1	42	28
	2	48	10
	3	34	26
	4	14	11
	5	26	28
	6	164	103
	<i>Total</i>	42	28

Table 37. GTheory Results for Annual School Violence Perception Survey Items

Analysis	Variance Component	Variance Estimate	Percent of Variance Accounted For
<i>Swahili x French</i>	Item	0.332	40.34%
	Language	0.000*	0.00%
	Enumerator	0.000*	0.00%
	Item x Language	0.013	1.58%
	Item x Enumerator	0.124	15.07%
	Language x Enumerator	0.002	0.24%
	Item x Enumerator x Language, Error	0.352	42.77%
	Total	0.823	100.00%

Many-Facet Model

Determining Datasets for Analysis

The process of determining datasets for the Many-Facet Model analyses was a simpler process, as it does not require a fully-crossed design in order to provide more detailed information. The only time cases were removed from datasets was due to missingness. If a case was missing data for any of the relevant chosen facets of interest, the case was removed. For the Baseline analyses, the facets of interest were Girls, Items, Province, Girl's Home Language, Enumerator, and Urbanicity. For the Annual analyses, the facets of interest were as above, but also included Enumerator's Home Language. For the Longitudinal analyses, the facets of interest were Girls, Items, Province, Girl's Home Language, Urbanicity, and Administration, as a proxy for translation. Note that informal translation of the subjective measure was done at Baseline, and formal adaptations of the measures were available at the Annual evaluation.

Analytic Procedure and Interpretation

The FACETS computer program (Linacre, 2007) was used to calibrate facets for all sets of analyses. FACETS provides many ways in which to analyze and review data, however, for these analyses the main output of interest is the variable map, and the summary statistics. As a review of the utility of the variable map (Figure 32, this example is taken from Chapter 3). The following example uses persons, items, and language. In this case, we use the EGRA as an example, and the three facets are persons, EGRA subtasks, and preferred home language of the girl being assessed. We can see in this example that there are 20 persons, five subtasks (LN – Letter Name, NW – Nonsense Word Reading, ORF – Oral Reading Fluency, RC – Reading

Comprehension, LC – Listening Comprehension), and six languages (FR- French, TS – Tshiluba, BE – Bemba, LI – Lingala, KIL – Kilendu, KIK – Kikongo SW – Swahili). Interpretation of the map is relatively intuitive with the understanding of the study design. In the person column, we see that girl 3 had the highest level of reading ability, and girl 10 had the lowest. In the item difficulty column we can see that reading comprehension was the most difficult task for the girls, and letter naming was the least difficult. Finally, those girls who indicated that their preferred language was French has the highest level of reading ability, and the girls who indicated Swahili as their preferred language had the lowest.

Logit	Person Score	Item Difficulty	Language
	High Reading Ability	More Difficult	More
4	3		FR
3	5 18	RC	TS BE
2	1 8	NW ORF	LI
1	7 9 11	LC	
0	2 6 12 13		
-1	7 15 19		KIL
-2	4 14 17	LN	
-3	16 20		KIK
-4	10		SW
	Low Reading Ability	Less Difficult	Less

Figure 32. Wright map for Many-Facet Model.

*Item Difficulty Definitions: LN = Letter Name; NW = Nonword Reading; ORF = Oral Reading Fluency; RC = Reading Comprehension; LC = Listening Comprehension

**Language Definitions: FR = French; TS = Tshiluba; BE = Bemba; LI = Lingala; KIL = Kilendu; KIK = Kikongo; SW = Swahili

In reviewing the summary statistics provided, it is useful to know that for the calibration, the items, girl's home province, girl's home language, enumerator's home language, enumerator,

urbanicity, and administration are anchored at zero by definition. This is done because only one facet (girls) is allowed to vary, allowing for an unambiguous result. Statistics provided in the results include the mean (M), standard deviation (SD), and sample size (N) providing basic context of location on the logit scale. Next, infit and outfit statistics provide information about the model fit. For each facet, the average infit and outfit are provided along with the standard deviation. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of .20. The outfit statistic is a more rigorous fit statistic, where the infit statistic limits the influence of outliers, of which there can be many in a dataset like the one used in this analysis.

Finally, the reliability of separation statistic provided, and it is conceptually equivalent to Cronbach's coefficient alpha. This statistic is also produced for each of the facets, and tests the hypothesis of whether or not there are significant differences between the elements within a facet. If the statistic is significant, it shows there is spread of the facet along whatever the latent variable is for the analysis at hand. This type of spread is what we look for in the Girls and Items facets, as we want to see spread along the latent variable for both the girls, in terms of their scores spanning the score scale, and the items, in terms of spanning the spectrum of difficulty (that is, we want to see items that are more difficult, less difficult, and moderately difficult). For the other facets of interest, however, significant spread is indicative of a possible problem. We do not want to see significant spread across the language of administration, for example, as this tells us that scores on the task are dependent on the language the assessment was administered in.

Many-Facet Model Results

Baseline – Objective Measures

Early Grades Reading Assessment (EGRA)

Figure 33 displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, enumerator, and urbanicity. The first column of Figure 33 represents the logit scale. The second column of the variable map displays the student measures on the Letter Name task of the EGRA. Girls with higher ability on the task appear at the top of the column, while girls with lower ability are at the bottom. Each asterisk represents 69 girls. The girl's achievement measures ranged from -12.81 logits to 12.44 logits ($M = -5.59$, $SD = 3.91$, $N = 3195$). The third column shows the locations of the Province facet on the latent variable where provinces appearing higher in this column showing higher achievement. The fourth, fifth, and sixth columns represent the Girl's Home Language, Enumerator ID, and Urbanicity Facets, respectively. As with the Province facet, values for these facets appearing higher on the map represent higher achievement on the task.

For the Letter Name subtask of the EGRA, we see that girls in Province Orientale showed slightly higher results than the rest of the provinces, while girls based in Katanga showed results slightly lower than the average. For the Enumerator facet, we can identify the groups of enumerators that yielded either higher or lower results on the subtask. We do not see any difference in scores on the Urbanicity facet, however, as girls in urban and rural areas performed similarly. The seventh and final column represents the location of EGRA Letter Name task items with item difficulty ranging from -8.70 logits to 9.30 logits ($M = 0.00$, $SD = 4.79$, $N = 100$).

Table 38 **Error! Reference source not found.** shows a set of summary statistics related to the FACETS analyses. As previously noted, the items, girl's home province, girl's home language, enumerator, and urbanicity are anchored at zero by definition. This is done because only one facet (girls) is allowed to vary. The overall model-data fit is mixed. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of .20, and only both statistics for the Girl facet are close to this expectation. For all other facets, while the infit is within reasonable limits, the outfit statistics are much larger than expected, which is conformation of the large number of (sometimes) extreme outliers.

However, as shown in Table 38, all six of the reliability of separation statistics are statistically significant ($p < .01$). The reliability of separation statistic is conceptually equivalent to Cronbach's coefficient alpha, used in this case to test whether or not there are significant differences between the elements within a facet. For the Letter Name task, the largest reliability of separation index is >0.99 for Items, Province, and Girl's Home Language. For this subtask, there is good differentiation for the Items (0.94) and Girls (0.94). However, as the reliability of separation for all other facets were also significant, this indicates there may be substantive differences between Enumerators (0.98), Provinces (>0.99), Urbanicity (0.92), and Girl's Home Language (>0.99).

Measr	+Girl	+Province-	+Language	+Enumerator	+Urbanicity	+Items
10 + .	+	+	+	+	+	10
9 + .	+	+	+	+	+	13
8 + .	+	+	+	+	+	1
7 + .	+	+	+	+	+	8
6 + .	+	+	+	+	+	11 3 4
5 + .	+	+	+	+	+	15 20 22 24 5 6 7
4 + .	+	+	+	+	+	14 2 23 25 26 27 29 9
3 + .	+	+	+	+	+	12 17
2 + .	+	+	+	+	+	28 32
1 + .	+	+	+	+	+	16 18 33
0 + .	+	+	+	+	+	30 31 34 35 38 40
-1 + .	+	+	+	+	+	21 36 37
-2 + .	+	+	+	+	+	19 43
-3 + .	+	+	+	+	+	39 41 42 44 45 46 47
-4 + .	+	+	+	+	+	49 50
-5 + .	+	+	+	+	+	51 52
-6 + .	+	+	+	+	+	53 54 55 56 57
-7 + .	+	+	+	+	+	58 60 61
-8 + .	+	+	+	+	+	62 63 65 66
-9 + .	+	+	+	+	+	68 69
-10 + .	+	+	+	+	+	64 67 70
-11 + .	+	+	+	+	+	72 73 75
-12 + .	+	+	+	+	+	74 77 78
-13 + .	+	+	+	+	+	76 79 80
-14 + .	+	+	+	+	+	81
-15 + .	+	+	+	+	+	82 83
-16 + .	+	+	+	+	+	84 85 86
-17 + .	+	+	+	+	+	87 88
-18 + .	+	+	+	+	+	89 90
-19 + .	+	+	+	+	+	91 92 93 94 95
-20 + .	+	+	+	+	+	100 96 97 98 99
-21 + .	+	+	+	+	+	
-22 + .	+	+	+	+	+	
-23 + .	+	+	+	+	+	
-24 + .	+	+	+	+	+	
-25 + .	+	+	+	+	+	
-26 + .	+	+	+	+	+	
-27 + .	+	+	+	+	+	
-28 + .	+	+	+	+	+	
-29 + .	+	+	+	+	+	
-30 + .	+	+	+	+	+	
-31 + .	+	+	+	+	+	
-32 + .	+	+	+	+	+	
-33 + .	+	+	+	+	+	
-34 + .	+	+	+	+	+	
-35 + .	+	+	+	+	+	
-36 + .	+	+	+	+	+	
-37 + .	+	+	+	+	+	
-38 + .	+	+	+	+	+	
-39 + .	+	+	+	+	+	
-40 + .	+	+	+	+	+	
-41 + .	+	+	+	+	+	
-42 + .	+	+	+	+	+	
-43 + .	+	+	+	+	+	
-44 + .	+	+	+	+	+	
-45 + .	+	+	+	+	+	
-46 + .	+	+	+	+	+	
-47 + .	+	+	+	+	+	
-48 + .	+	+	+	+	+	
-49 + .	+	+	+	+	+	
-50 + .	+	+	+	+	+	
-51 + .	+	+	+	+	+	
-52 + .	+	+	+	+	+	
-53 + .	+	+	+	+	+	
-54 + .	+	+	+	+	+	
-55 + .	+	+	+	+	+	
-56 + .	+	+	+	+	+	
-57 + .	+	+	+	+	+	
-58 + .	+	+	+	+	+	
-59 + .	+	+	+	+	+	
-60 + .	+	+	+	+	+	
-61 + .	+	+	+	+	+	
-62 + .	+	+	+	+	+	
-63 + .	+	+	+	+	+	
-64 + .	+	+	+	+	+	
-65 + .	+	+	+	+	+	
-66 + .	+	+	+	+	+	
-67 + .	+	+	+	+	+	
-68 + .	+	+	+	+	+	
-69 + .	+	+	+	+	+	
-70 + .	+	+	+	+	+	
-71 + .	+	+	+	+	+	
-72 + .	+	+	+	+	+	
-73 + .	+	+	+	+	+	
-74 + .	+	+	+	+	+	
-75 + .	+	+	+	+	+	
-76 + .	+	+	+	+	+	
-77 + .	+	+	+	+	+	
-78 + .	+	+	+	+	+	
-79 + .	+	+	+	+	+	
-80 + .	+	+	+	+	+	
-81 + .	+	+	+	+	+	
-82 + .	+	+	+	+	+	
-83 + .	+	+	+	+	+	
-84 + .	+	+	+	+	+	
-85 + .	+	+	+	+	+	
-86 + .	+	+	+	+	+	
-87 + .	+	+	+	+	+	
-88 + .	+	+	+	+	+	
-89 + .	+	+	+	+	+	
-90 + .	+	+	+	+	+	
-91 + .	+	+	+	+	+	
-92 + .	+	+	+	+	+	
-93 + .	+	+	+	+	+	
-94 + .	+	+	+	+	+	
-95 + .	+	+	+	+	+	
-96 + .	+	+	+	+	+	
-97 + .	+	+	+	+	+	
-98 + .	+	+	+	+	+	
-99 + .	+	+	+	+	+	
-100 + .	+	+	+	+	+	
-101 + .	+	+	+	+	+	
-102 + .	+	+	+	+	+	
-103 + .	+	+	+	+	+	
-104 + .	+	+	+	+	+	
-105 + .	+	+	+	+	+	
-106 + .	+	+	+	+	+	
-107 + .	+	+	+	+	+	
-108 + .	+	+	+	+	+	
-109 + .	+	+	+	+	+	
-110 + .	+	+	+	+	+	
-111 + .	+	+	+	+	+	
-112 + .	+	+	+	+	+	
-113 + .	+	+	+	+	+	
-114 + .	+	+	+	+	+	
-115 + .	+	+	+	+	+	
-116 + .	+	+	+	+	+	
-117 + .	+	+	+	+	+	
-118 + .	+	+	+	+	+	
-119 + .	+	+	+	+	+	
-120 + .	+	+	+	+	+	
-121 + .	+	+	+	+	+	
-122 + .	+	+	+	+	+	
-123 + .	+	+	+	+	+	
-124 + .	+	+	+	+	+	
-125 + .	+	+	+	+	+	
-126 + .	+	+	+	+	+	
-127 + .	+	+	+	+	+	
-128 + .	+	+	+	+	+	
-129 + .	+	+	+	+	+	
-130 + .	+	+	+	+	+	
-131 + .	+	+	+	+	+	
-132 + .	+	+	+	+	+	
-133 + .	+	+	+	+	+	
-134 + .	+	+	+	+	+	
-135 + .	+	+	+	+	+	
-136 + .	+	+	+	+	+	
-137 + .	+	+	+	+	+	
-138 + .	+	+	+	+	+	
-139 + .	+	+	+	+	+	
-140 + .	+	+	+	+	+	
-141 + .	+	+	+	+	+	
-142 + .	+	+	+	+	+	
-143 + .	+	+	+	+	+	
-144 + .	+	+	+	+	+	
-145 + .	+	+	+	+	+	
-146 + .	+	+	+	+	+	
-147 + .	+	+	+	+	+	
-148 + .	+	+	+	+	+	
-149 + .	+	+	+	+	+	
-150 + .	+	+	+	+	+	
-151 + .	+	+	+	+	+	
-152 + .	+	+	+	+	+	
-153 + .	+	+	+	+	+	
-154 + .	+	+	+	+	+	
-155 + .	+	+	+	+	+	
-156 + .	+	+	+	+	+	
-157 + .	+	+	+	+	+	
-158 + .	+	+	+	+	+	
-159 + .	+	+	+	+	+	
-160 + .	+	+	+	+	+	
-161 + .	+	+	+	+	+	
-162 + .	+	+	+	+	+	
-163 + .	+	+	+	+	+	
-164 + .	+	+	+	+	+	
-165 + .	+	+	+	+	+	
-166 + .	+	+	+	+	+	
-167 + .	+	+	+	+	+	
-168 + .	+	+	+	+	+	
-169 + .	+	+	+	+	+	
-170 + .	+	+	+	+	+	
-171 + .	+	+	+	+	+	
-172 + .	+	+	+	+	+	
-173 + .	+	+	+	+	+	
-174 + .	+	+	+	+	+	
-175 + .	+	+	+	+	+	
-176 + .	+	+	+	+	+	
-177 + .	+	+	+	+	+	
-178 + .	+	+	+	+	+	
-179 + .	+	+	+	+	+	
-180 + .	+	+	+	+	+	
-181 + .	+	+	+	+	+	
-182 + .	+	+	+	+	+	
-183 + .	+	+	+	+	+	
-184 + .	+	+	+	+	+	
-185 + .	+	+	+	+	+	
-186 + .	+	+	+	+	+	
-187 + .	+	+	+	+	+	
-188 + .	+	+	+	+	+	
-189 + .	+	+	+	+	+	
-190 + .	+	+	+	+	+	
-191 + .	+	+	+	+	+	
-192 + .	+	+	+	+	+	
-193 + .	+	+	+	+	+	
-194 + .	+	+	+	+	+	
-195 + .	+	+	+	+	+	
-196 + .	+	+	+	+	+	
-197 + .	+	+	+	+	+	
-198 + .	+	+	+	+	+	
-199 + .	+	+	+	+	+	
-200 + .	+	+	+	+	+	
-201 + .	+	+	+	+	+	
-202 + .	+	+	+	+	+	
-203 + .	+	+	+	+	+	
-204 + .	+	+	+	+	+	
-205 + .	+	+	+	+	+	
-206 + .	+	+	+	+	+	
-207 + .	+	+	+	+	+	
-208 + .	+	+	+	+	+	
-209 + .	+	+	+	+	+	
-210 + .	+	+	+	+	+	
-211 + .	+	+	+	+	+	
-212 + .	+	+	+	+	+	
-213 + .	+	+	+	+	+	
-214 + .	+	+	+	+	+	
-215 + .	+	+	+	+	+	
-216 + .	+	+	+	+	+	
-217 + .	+	+	+	+	+	
-218 + .	+	+	+	+	+	
-219 + .	+	+	+	+	+	
-220 + .	+	+	+	+	+	
-221 + .	+	+	+	+	+	
-222 + .	+	+	+	+	+	
-223 + .	+	+	+	+	+	
-224 + .	+	+	+	+	+	
-225 + .	+	+	+	+	+	
-226 + .	+	+	+	+	+	
-227 + .	+	+	+	+	+	
-228 + .	+	+	+	+	+	
-229 + .	+	+	+	+	+	
-230 + .	+	+	+	+	+	
-231 + .	+	+	+	+	+	
-232 + .	+	+	+	+	+	
-233 + .	+	+	+	+	+	
-234 + .	+	+	+	+	+	
-235 + .	+	+	+	+	+	
-236 + .	+	+	+	+	+	
-237 + .	+	+	+	+	+	
-238 + .	+	+	+	+	+	
-239 + .	+	+	+	+	+	
-240 + .	+	+	+	+	+	
-241 + .	+	+	+	+	+	
-242 + .	+	+	+	+	+	
-243 + .	+	+	+	+	+	
-244 + .	+	+	+	+	+	
-245 + .	+	+	+	+	+	
-246 + .	+	+	+	+	+	
-247 + .	+	+	+	+	+	
-248 + .	+	+	+	+	+	
-249 + .	+	+	+	+	+	
-250 + .	+	+	+	+	+	
-251 + .	+	+	+	+	+	
-252 + .	+	+	+	+	+	
-253 + .	+	+	+	+	+	
-254 + .	+	+	+	+	+	
-255 + .	+	+	+	+	+	
-256 + .	+	+	+	+	+	
-257 + .	+	+	+	+	+	
-258 + .	+	+	+	+	+	
-259 + .	+	+	+	+	+	

Table 38. Facets Results for Baseline EGRA – Letter Name Items

	Girls	Item	Province	Girl's Home Language	Enumerator	Urbanicity
Measures						
Mean	-5.59	0.00	0.00	0.00	0.00	0.00
SD	3.91	4.79	0.35	0.46	0.60	0.06
n	3195	100	5	6	54	2
INFIT						
Mean	0.95	0.82	0.94	0.92	0.94	0.95
SD	0.43	0.33	0.04	0.08	0.16	0.01
OUTFIT						
Mean	1.02	2.96	7.65	5.84	3.16	9.00
SD	2.01	3.86	3.02	3.73	3.23	0.00
Reliability of Separation						
Chi-Square Statistic	102604.7	107201.6	1013.3	1034.7	3971.8	25.1
Degrees of Freedom	3194	99	4	5	54	1

*p < 0.05

Figure 34Error! Reference source not found. displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, enumerator, and urbanicity for the Non-Word Reading items on the EGRA. For this map, each asterisk represents 117 girls. The girl's achievement measures ranged from -9.90 logits to 8.59 logits (M = -5.53, SD = 3.29, N = 3196).

For the Non-Word Reading subtask, we see that girls in Province Orientale showed slightly higher results than the rest of the provinces, while girls based the rest of the provinces performed similarly. Girls who identified French as their home language performed slightly better than all other languages. For the Enumerator facet, we can identify the groups of enumerators that yielded either higher or lower results on the subtask. We do not see any difference in scores on

the Urbanicity facet, however, as girls in urban and rural areas performed similarly. The seventh and final column represents the location of EGRA Non-Word Reading subtask items with item difficulty ranging from -4.85 logits to 5.98 logits ($M = 0.00$, $SD = 3.03$, $N = 50$).

Table 39 shows that the overall model-data fit is mixed here as well. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of $.20$, and while this subtask performed better than Letter Name, only the statistics for the Girl facet are close to this expectation. For all other facets, while the infit is within reasonable limits, the outfit statistics are much larger than expected.

However, as shown in Table 39 **Error! Reference source not found.**, all six of the reliability of separation statistics are statistically significant ($p < .01$). The reliability of separation statistic is conceptually equivalent to Cronbach's coefficient alpha, used in this case to test whether or not there are significant differences between the elements within a facet. For the Non-Word Reading subtask, the largest reliability of separation index is >0.99 for Items. For this subtask, there is good differentiation for the Items (0.94), and while the Girls index is smaller (0.81), it is also significant. This is not unexpected as we can see on the variable map that most of the girls have not performed well, and cluster together at the bottom of the logit scale. Province (0.97), Girl's Home Language (0.94), Enumerator (0.90), and Urbanicity (0.97) were also significant, indicating there may be substantive differences of note for these facets.

Table 39. Facets Results for Baseline EGRA – Non-Word Reading Items

	Girls	Item	Province	Girl's Home Language	Enumerator	Urbanicity
<i>Measures</i>						
<i>Mean</i>	-5.53	0.00	0.00	0.00	0.00	0.00
<i>SD</i>	3.29	3.03	0.18	0.18	0.45	0.10
<i>n</i>	3196	50	5	6	55	2
<i>INFIT</i>						
<i>Mean</i>	0.97	0.91	0.98	0.97	0.96	0.99
<i>SD</i>	0.40	0.25	0.05	0.09	0.16	0.00
<i>OUTFIT</i>						
<i>Mean</i>	0.89	1.65	3.05	2.61	1.42	3.20
<i>SD</i>	1.60	2.54	2.58	2.33	1.75	1.84
<i>Reliability of Separation</i>	0.81*	>0.99*	0.97*	0.94*	0.90*	0.97*
<i>Chi-Square Statistic</i>	42578.5	31195.0	124.2	137.3	1158.9	29.2
<i>Degrees of Freedom</i>	3195	49	4	5	54	1

*p < 0.05

Figure 35Error! Reference source not found. displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, enumerator, and urbanicity for the Oral Reading Fluency items on the EGRA. For this map, each asterisk represents 117 girls. The girl's achievement measures ranged from -10.16 logits to 7.61 logits (M = -5.13, SD = 3.98, N = 3195).

For the Oral Reading subtask, we see that girls in Katanga showed slightly higher results than the rest of the provinces, and girls based in Bandundu showed results slightly lower than the average, while girls based the rest of the provinces performed similarly. Girls who identified French as their home language outperformed others, followed by Swahili. Conversely, girls who reported Kilendu as their home language performed below all other languages. For the Enumerator facet, we can identify the groups of enumerators that yielded either higher or lower

results on the subtask. Here girls in rural areas performed better than those in urban areas, and Item facet showed item difficulty ranging from -3.62 logits to 5.54 logits ($M = 0.00$, $SD = 2.29$, $N = 50$).

Table 40**Error! Reference source not found.** shows the overall model-data fit is not good. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of $.20$, but none of the facets fall within this this expectation for the outfit statistic. For all facets, while the infit statistics are within reasonable limits, the outfit statistics are much larger than expected.

However, as shown in Table 40**Error! Reference source not found.**, all six of the reliability of separation statistics are statistically significant ($p < .01$). For the Oral Reading Fluency task, the largest reliability of separation index is >0.99 for Items. For this subtask, there is good differentiation for the Items (>0.99), and while the Girls index is smaller (0.85), it is also significant. This is not unexpected as we can see on the variable map that most of the girls have not performed well, and cluster together at the bottom of the logit scale. Province (0.98), Girl's Home Language (0.99), Enumerator (0.96), and Urbanicity (>0.99) were also significant, indicating there may be substantive differences of note for these facets.

Table 40. Facets Results for Baseline EGRA – Oral Reading Fluency Items

	Girls	Item	Province	Girl's Home Language	Enumerator	Urbanicity
Measures						
Mean	-5.13	0.00	0.00	0.00	0.00	0.00
SD	3.98	2.29	0.24	0.50	0.70	0.33
n	3195	50	5	6	54	2
INFIT						
Mean	0.94	0.92	0.93	0.93	0.92	0.94
SD	0.36	0.41	0.05	0.06	0.16	0.01
OUTFIT						
Mean	1.42	2.05	4.14	3.59	2.68	5.80
SD	2.16	2.84	2.77	2.82	2.79	3.84
Reliability of Separation	0.85*	>0.99*	0.98*	0.99*	0.96*	>0.99*
Chi-Square Statistic	52363.0	21975.0	273.9	515.7	2214.0	287.0
Degrees of Freedom	3194	49	4	5	53	1

*p < 0.05

Figure 36Error! Reference source not found. displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, enumerator, and urbanicity for the Reading Comprehension items on the EGRA. For this map, each asterisk represents 256 girls. The girl's achievement measures ranged from -4.87 logits to 4.77 logits (M = -2.92, SD = 1.60, N = 3195).

For the Reading Comprehension subtask, we see that girls in Equateur and Katanga showed slightly higher results than the rest of the provinces, and girls based in Province Orientale showed results slightly lower than the average, while girls based the rest of the provinces performed similarly. Girls who identified Kilendu as their home language outperformed others, followed by Tshiluba. Conversely, girls who reported French as their home language performed below all other languages. For the Enumerator facet, we can identify the groups of enumerators

that yielded either higher or lower results on the subtask. For this analysis, Urbanicity did not show differential results, and the Item facet showed item difficulty ranging from -2.50 logits to 1.69 logits ($M = 0.00$, $SD = 1.70$, $N = 5$).

Table 41 shows that the overall model-data fit is good. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of $.20$, and all of the infit and outfit statistics for all six facets fall in the acceptable range. Also shown in Table 41 **Error! Reference source not found.**, four of the six of the reliability of separation statistics are statistically significant ($p < .01$). For the Reading Comprehension subtask, the largest reliability of separation index is >0.99 for Items. For this subtask, there is good differentiation for the Items (>0.99), and while the Girls index is significant, it is 0.00 . This is not unexpected as we can see on the variable map that most of the girls have not performed well, and cluster together at the bottom of the logit scale. Girl's Home Language (0.55) and Enumerator (0.00) were also significant, indicating there may be substantive differences of note for these facets. Province (0.20) and Urbanicity (0.00) were not significant, indicating the variability of these facets along the latent variable was small.

Measr	+Girl	+Province	+Language	+Enumerator	+Urbanicity	+Items
4	+	+	+	+	+	+
3	+	+	+	+	+	+
2	+	+	+	+	+	+
1	+	+	+	27 37 39	+	1
0	+	+	+	56	+	4
-1	+	+	+	55	+	2
-2	+	+	+	47	+	2
-3	+	+	+	16 33 34 44 45 6 8	+	2
-4	+	+	+	1 10 2 21 24 30 31 35 50	+	2
-5	+	+	+	12 20 23 29 3 4 40 53	+	2
-6	+	+	+	13 18 48 49 54 7	+	2
-7	+	+	+	15 25 28 32 36 38 41 5 9	+	2
-8	+	+	+	11 14 17 19 26	+	2
-9	+	+	+	22 51	+	2
-10	+	+	+	46 52	+	2
-11	+	+	+	+	+	3
-12	+	+	+	+	+	5
-13	+	+	+	+	+	5
-14	+	+	+	+	+	5
-15	+	+	+	+	+	5
-16	+	+	+	+	+	5
-17	+	+	+	+	+	5
-18	+	+	+	+	+	5
-19	+	+	+	+	+	5
-20	+	+	+	+	+	5
-21	+	+	+	+	+	5
-22	+	+	+	+	+	5
-23	+	+	+	+	+	5
-24	+	+	+	+	+	5
-25	+	+	+	+	+	5
-26	+	+	+	+	+	5
-27	+	+	+	+	+	5
-28	+	+	+	+	+	5
-29	+	+	+	+	+	5
-30	+	+	+	+	+	5
-31	+	+	+	+	+	5
-32	+	+	+	+	+	5
-33	+	+	+	+	+	5
-34	+	+	+	+	+	5
-35	+	+	+	+	+	5
-36	+	+	+	+	+	5
-37	+	+	+	+	+	5
-38	+	+	+	+	+	5
-39	+	+	+	+	+	5
-40	+	+	+	+	+	5
-41	+	+	+	+	+	5
-42	+	+	+	+	+	5
-43	+	+	+	+	+	5
-44	+	+	+	+	+	5
-45	+	+	+	+	+	5
-46	+	+	+	+	+	5
-47	+	+	+	+	+	5
-48	+	+	+	+	+	5
-49	+	+	+	+	+	5
-50	+	+	+	+	+	5
-51	+	+	+	+	+	5
-52	+	+	+	+	+	5
-53	+	+	+	+	+	5
-54	+	+	+	+	+	5
-55	+	+	+	+	+	5
-56	+	+	+	+	+	5
-57	+	+	+	+	+	5
-58	+	+	+	+	+	5
-59	+	+	+	+	+	5
-60	+	+	+	+	+	5
-61	+	+	+	+	+	5
-62	+	+	+	+	+	5
-63	+	+	+	+	+	5
-64	+	+	+	+	+	5
-65	+	+	+	+	+	5
-66	+	+	+	+	+	5
-67	+	+	+	+	+	5
-68	+	+	+	+	+	5
-69	+	+	+	+	+	5
-70	+	+	+	+	+	5
-71	+	+	+	+	+	5
-72	+	+	+	+	+	5
-73	+	+	+	+	+	5
-74	+	+	+	+	+	5
-75	+	+	+	+	+	5
-76	+	+	+	+	+	5
-77	+	+	+	+	+	5
-78	+	+	+	+	+	5
-79	+	+	+	+	+	5
-80	+	+	+	+	+	5
-81	+	+	+	+	+	5
-82	+	+	+	+	+	5
-83	+	+	+	+	+	5
-84	+	+	+	+	+	5
-85	+	+	+	+	+	5
-86	+	+	+	+	+	5
-87	+	+	+	+	+	5
-88	+	+	+	+	+	5
-89	+	+	+	+	+	5
-90	+	+	+	+	+	5
-91	+	+	+	+	+	5
-92	+	+	+	+	+	5
-93	+	+	+	+	+	5
-94	+	+	+	+	+	5
-95	+	+	+	+	+	5
-96	+	+	+	+	+	5
-97	+	+	+	+	+	5
-98	+	+	+	+	+	5
-99	+	+	+	+	+	5
-100	+	+	+	+	+	5
-101	+	+	+	+	+	5
-102	+	+	+	+	+	5
-103	+	+	+	+	+	5
-104	+	+	+	+	+	5
-105	+	+	+	+	+	5
-106	+	+	+	+	+	5
-107	+	+	+	+	+	5
-108	+	+	+	+	+	5
-109	+	+	+	+	+	5
-110	+	+	+	+	+	5
-111	+	+	+	+	+	5
-112	+	+	+	+	+	5
-113	+	+	+	+	+	5
-114	+	+	+	+	+	5
-115	+	+	+	+	+	5
-116	+	+	+	+	+	5
-117	+	+	+	+	+	5
-118	+	+	+	+	+	5
-119	+	+	+	+	+	5
-120	+	+	+	+	+	5
-121	+	+	+	+	+	5
-122	+	+	+	+	+	5
-123	+	+	+	+	+	5
-124	+	+	+	+	+	5
-125	+	+	+	+	+	5
-126	+	+	+	+	+	5
-127	+	+	+	+	+	5
-128	+	+	+	+	+	5
-129	+	+	+	+	+	5
-130	+	+	+	+	+	5
-131	+	+	+	+	+	5
-132	+	+	+	+	+	5
-133	+	+	+	+	+	5
-134	+	+	+	+	+	5
-135	+	+	+	+	+	5
-136	+	+	+	+	+	5
-137	+	+	+	+	+	5
-138	+	+	+	+	+	5
-139	+	+	+	+	+	5
-140	+	+	+	+	+	5
-141	+	+	+	+	+	5
-142	+	+	+	+	+	5
-143	+	+	+	+	+	5
-144	+	+	+	+	+	5
-145	+	+	+	+	+	5
-146	+	+	+	+	+	5
-147	+	+	+	+	+	5
-148	+	+	+	+	+	5
-149	+	+	+	+	+	5
-150	+	+	+	+	+	5
-151	+	+	+	+	+	5
-152	+	+	+	+	+	5
-153	+	+	+	+	+	5
-154	+	+	+	+	+	5
-155	+	+	+	+	+	5
-156	+	+	+	+	+	5
-157	+	+	+	+	+	5
-158	+	+	+	+	+	5
-159	+	+	+	+	+	5
-160	+	+	+	+	+	5
-161	+	+	+	+	+	5
-162	+	+	+	+	+	5
-163	+	+	+	+	+	5
-164	+	+	+	+	+	5
-165	+	+	+	+	+	5
-166	+	+	+	+	+	5
-167	+	+	+	+	+	5
-168	+	+	+	+	+	5
-169	+	+	+	+	+	5
-170	+	+	+	+	+	5
-171	+	+	+	+	+	5
-172	+	+	+	+	+	5
-173	+	+	+	+	+	5
-174	+	+	+	+	+	5
-175	+	+	+	+	+	5
-176	+	+	+	+	+	5
-177	+	+	+	+	+	5
-178	+	+	+	+	+	5
-179	+	+	+	+	+	5
-180	+	+	+	+	+	5
-181	+	+	+	+	+	5
-182	+	+	+	+	+	5
-183	+	+	+	+	+	5
-184	+	+	+	+	+	5
-185	+	+	+	+	+	5
-186	+	+	+	+	+	5
-187	+	+	+	+	+	5
-188	+	+	+	+	+	5
-189	+	+	+	+	+	5
-190	+	+	+	+	+	5
-191	+	+	+	+	+	5
-192	+	+	+	+	+	5
-193	+	+	+	+	+	5
-194	+	+	+	+	+	5
-195	+	+	+	+	+	5
-196	+	+	+	+	+	5
-197	+	+	+	+	+	5
-198	+	+	+	+	+	5
-199	+	+	+	+	+	5
-200	+	+	+	+	+	5
-201	+	+	+	+	+	5
-202	+	+	+	+	+	5
-203	+	+	+	+	+	5
-204	+	+	+	+	+	5
-205	+	+	+	+	+	5
-206	+	+	+	+	+	5
-207	+	+	+	+	+	5
-208	+	+	+	+	+	5
-209	+	+	+	+	+	5
-210	+	+	+	+	+	5
-211	+	+	+	+	+	5
-212	+	+	+	+	+	5
-213	+	+	+	+	+	5
-214	+	+	+	+	+	5
-215	+	+	+	+	+	5
-216	+	+	+	+	+	5
-217	+	+	+	+	+	5
-218	+	+	+	+	+	5
-219	+	+	+	+	+	5
-220	+	+	+	+	+	5
-221	+	+	+	+	+	5
-222	+	+	+	+	+	5
-223	+	+	+	+	+	5
-224	+	+	+	+	+	5
-225	+	+	+	+	+	5
-226	+	+	+	+	+	5
-227	+	+	+	+	+	5
-228	+	+	+	+	+	5
-229	+	+	+	+	+	5
-230	+	+	+	+	+	5
-231	+	+	+	+	+	5
-232	+	+	+	+	+	5
-233	+	+	+	+	+	5
-234	+	+	+	+	+	5
-235	+	+	+	+	+	5
-236	+	+	+	+	+	5
-237	+	+	+	+	+	5
-238	+	+	+	+	+	5
-239	+	+	+	+	+	5
-240	+	+	+	+	+	5
-241	+	+	+	+	+	5
-242	+	+	+	+	+	5
-243	+	+	+	+	+	5
-244	+	+	+	+	+	5
-245	+	+	+	+	+	5
-246	+	+	+	+	+	5
-247	+	+	+	+	+	5
-248	+	+	+	+	+	5
-249	+	+	+	+	+	5
-250	+	+	+	+	+	5
-251	+	+	+	+	+	5
-252	+	+	+	+	+	5
-253	+	+	+	+	+	5
-254	+	+	+	+	+	5
-255	+	+	+	+	+	5
-256	+	+	+	+	+	5
-257	+	+	+	+	+	5
-258	+	+	+	+	+	5
-259	+	+	+	+	+	5
-260	+	+	+	+	+	5
-261	+	+	+	+	+	5
-262	+	+	+	+	+	5
-263	+	+	+	+	+	5
-264	+	+	+	+	+	5
-265	+	+	+	+	+	5
-266	+	+	+	+	+	5
-267	+	+	+	+	+	5
-268	+	+	+	+	+	5

Table 41. Facets Results for Baseline EGRA – Reading Comprehension Items

	Girls	Item	Province	Girl's Home Language	Enumerator	Urbanicity
<i>Measures</i>						
<i>Mean</i>	-2.92	0.00	0.00	0.00	0.00	0.00
<i>SD</i>	1.60	1.70	0.12	0.22	0.46	0.03
<i>n</i>	3195	5	5	6	54	2
<i>INFIT</i>						
<i>Mean</i>	1.00	1.01	1.02	1.01	1.02	1.01
<i>SD</i>	0.55	0.10	0.06	0.06	0.29	0.06
<i>OUTFIT</i>						
<i>Mean</i>	0.97	0.98	1.00	0.97	0.99	0.98
<i>SD</i>	1.22	0.19	0.11	0.07	0.51	0.05
<i>Reliability of Separation</i>	0.00*	>0.99*	0.20	0.55*	0.00*	0.00
<i>Chi-Square Statistic</i>	3824.5	715.5	5.1	12.2	73.4	0.2
<i>Degrees of Freedom</i>	3194	4	4	5	53	1

Figure 37Error! Reference source not found. displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, enumerator, and urbanicity for the Listening Comprehension items on the EGRA. For this map, each asterisk represents 185 girls. The girl's achievement measures ranged from -4.11 logits to 3.77 logits (M = -2.01, SD = 1.75, N = 3196).

For the Listening Comprehension subtask, we see that girls in Kasai Orientale showed slightly higher results than the rest of the provinces, and girls based in Province Orientale and Katanga showed results slightly lower than the average, while girls based the rest of the provinces performed similarly. Girls who identified Tshiluba as their home language outperformed others, and girls who reported Swahili as their home language performed below all other languages. For the Enumerator facet, we can identify the groups of enumerators that yielded either higher or

lower results on the subtask. For this analysis, Urbanicity did not show differential results, and the Item facet showed item difficulty ranging from -2.07 logits to 1.11 logits ($M = 0.00$, $SD = 1.24$, $N = 5$).

Table 42**Error! Reference source not found.** shows that the overall model-data fit is good. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of $.20$, and all of the infit and outfit statistics for all six facets fall in the acceptable range. Also shown in Table 42**Error! Reference source not found.**, four of the six of the reliability of separation statistics are statistically significant ($p < .01$). For the Listening Comprehension subtask, the largest reliability of separation index is >0.99 for Items. For this subtask, there is good differentiation for the Items (>0.99), and while the Girls index is small at 0.12 , it is significant. This is not unexpected as we can see on the variable map that most of the girls have not performed well, and cluster together at the bottom of the logit scale. Girl's Home Language (0.31) and Province (0.68) were also significant, indicating there may be substantive differences of note for these facets. Enumerator (0.09) and Urbanicity (0.68) were not significant, indicating the variability of these facets along the latent variable was small.

Measr	+Girl	+Province-	+Language	+Enumerator	+Urbanicity	+Items
3 + .	+	+	+	+	+	+
. . . .						
2 + .	+	+	+	+	+	+
.						
1 + .	+	+	+	40 + 37	+	1
.				29 33 39 45		3
.				19 2 28 30 34 48 55		4
* 0 *	* B E K PO	* F S	Kik Kil L	* 11 27 31 47 50 52 8 9 1 12 14 15 16 17 18 20 21 22 25 32 35 46 5 6 23 3 36 38 53 54 10 41 44 49 7 51 13 43	* Rural Urban *	2
.						
-1 + .	+	+	+	+	+	+
.						
-2 + .	+	+	+	+	+	5
.						
-3 + *****	+	+	+	+	+	+
Measr	* = 185	+Province-	+Language	+Enumerator	+Urbanicity	+Items

Figure 37. Variable Map for Baseline EGRA - Listening Comprehension Items

Table 42. Facets Results for Baseline EGRA – Listening Comprehension Items

	Girls	Item	Province	Girl's Home Language	Enumerator	Urbanicity
<i>Measures</i>						
<i>Mean</i>	-2.01	0.00	0.00	0.00	0.00	0.00
<i>SD</i>	1.75	1.24	0.13	0.12	0.39	0.08
<i>n</i>	3196	5	5	6	55	2
<i>INFIT</i>						
<i>Mean</i>	1.00	1.00	1.00	0.99	1.00	1.00
<i>SD</i>	0.42	0.12	0.14	0.13	0.21	0.01
<i>OUTFIT</i>						
<i>Mean</i>	0.99	0.99	0.98	0.96	0.97	1.00
<i>SD</i>	0.92	0.20	0.28	0.26	0.41	0.10
<i>Reliability of Separation</i>	0.12*	>0.99*	0.68*	0.31*	0.09	0.68
<i>Chi-Square Statistic</i>	4303.2	815.4	17.2	19.3	60.8	3.2
<i>Degrees of Freedom</i>	3195	4	4	5	54	1

Early Grades Mathematics Assessment (EGMA)

Figure 38**Error! Reference source not found.** displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, enumerator, and urbanicity for the Number Identification items on the EGMA. For this map, each asterisk represents 102 girls. The girl's achievement measures ranged from -8.49 logits to 8.25 logits ($M = 3.54$, $SD = 3.23$, $N = 3193$).

For the Number Identification subtask, we see that girls in Province Orientale showed slightly higher results than the rest of the provinces, and girls based in Equateur showed results slightly lower than the average, while girls based the rest of the provinces performed similarly. Girls who identified French as their home language outperformed others. Conversely, girls who reported Kikongo, Lingala, or Swahili as their home language performed below all other languages. For the Enumerator facet, we can identify the groups of enumerators that yielded either higher or lower results on the subtask. Here there was no difference in performance for Urbanicity, and the Item facet showed item difficulty ranging from -4.69 logits to 6.26 logits ($M = 0.00$, $SD = 3.15$, $N = 20$).

Table 43 shows that the overall model-data fit is mixed. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of .20, but only Girls falls within this expectation for the outfit statistic. For all facets, while the infit statistics are within reasonable limits, the outfit statistics are much larger than expected.

However, as shown in Table 43**Error! Reference source not found.**, all six of the reliability of separation statistics are statistically significant ($p < .01$). For the Number Identification subtask, the largest reliability of separation index is >0.99 for Items. For this subtask, there is good differentiation for the Items (>0.99), and while the Girls index is smaller

(0.84), it is also significant. This is not unexpected as we can see on the variable map that most of the girls have not performed well, and cluster together at the bottom of the logit scale. Province (0.88), Girl's Home Language (0.96), Enumerator (0.76), and Urbanicity (0.97) were also significant, indicating there may be substantive differences of note for these facets.

Measr	+Girl	+Province-	+Language	+Enumerator	+Urbanicity	+Items
7	*****.	+	+	+	+	+
	.	+	+	+	+	1 3
6	+	+	+	+	+	+
	.	+	+	+	+	+
5	+	+	+	+	+	+
	+	+	+	+	+	+
4	+	+	+	+	+	2
	+	+	+	+	+	+
3	+	+	+	+	+	7
	+	+	+	+	+	+
2	+	+	+	13	+	4 8
	.	+	+	+	+	+
1	+	+	F	14 41	+	6 5
	.	PO		1 10 12 15 16 22 28 4 44 48 49 54 7 8	+	13 9
*	0 *	* B K KO	* Kil T	* 11 18 2 23 24 26 27 3 30 31 33 35 40 47 5 55 56 6 9	* Rural Urban	* 15 *
	.	E	Kik L S	17 20 21 25 32 34 36 38 45 46 50 51 53		
-1	+	+	+	29 52	+	12
	.	+	+	37 39	+	11
-2	+	+	+	+	+	14
	.	+	+	+	+	10
-3	+	+	+	+	+	16
	.	+	+	+	+	17
-4	+	+	+	+	+	18 19
	.	+	+	+	+	20
-5	+	+	+	+	+	+
	.	+	+	+	+	+
-6	+	+	+	+	+	+
	.	+	+	+	+	+
-7	+	+	+	+	+	+
	.	+	+	+	+	+
Measr	* = 102	+Province-	+Language	+Enumerator	+Urbanicity	+Items

Figure 38. Variable Map for Baseline EGMA - Number Identification Items

Table 43. Facets Results for Baseline EGMA – Number Identification Items

	Girls	Item	Province	Girl's Home Language	Enumerator	Urbanicity
<i>Measures</i>						
<i>Mean</i>	0.37	0.00	0.00	0.00	0.00	0.00
<i>SD</i>	0.31	3.15	0.14	0.36	0.41	0.14
<i>n</i>	3193	20	5	6	53	2
<i>INFIT</i>						
<i>Mean</i>	0.95	0.97	0.94	0.92	0.94	0.94
<i>SD</i>	0.52	0.15	0.04	0.10	0.16	0.04
<i>OUTFIT</i>						
<i>Mean</i>	1.18	3.22	6.95	5.94	3.65	7.45
<i>SD</i>	2.14	3.17	2.81	3.56	3.17	2.19
<i>Reliability of Separation</i>	0.84*	>0.99*	0.88*	0.96*	0.76*	0.97*
<i>Chi-Square Statistic</i>	26384.8	19334.2	41.6	155.3	389.6	34.1
<i>Degrees of Freedom</i>	3193	19	4	5	52	1

Figure 39Error! Reference source not found. displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, enumerator, and urbanicity for the Number Discrimination items on the EGMA. For this map, each asterisk represents 99 girls. The girl's achievement measures ranged from -6.26 logits to 5.28 logits (M = 2.06, SD = 2.36, N = 3194).

For the Number Discrimination subtask, we see that girls in Province Orientale showed slightly higher results than the rest of the provinces, and girls based in Bandundu showed results slightly lower than the average, while girls based the rest of the provinces performed similarly. Girls who identified French, Kilendu, and Swahili as their home language outperformed others, and girls who reported Kikongo and Tshiluba as their home language performed below all other languages. For the Enumerator facet, we can identify the groups of enumerators that yielded

either higher or lower results on the subtask. For this analysis, Urbanicity did not show differential results, and the Item facet showed item difficulty ranging from -2.27 logits to 4.12 logits ($M = 0.00$, $SD = 1.97$, $N = 10$).

Table 44**Error! Reference source not found.** shows that the overall model-data fit is good. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of .20, and all of the infit and most of the outfit statistics for all six facets fall in the acceptable range. Also shown in Table 44**Error! Reference source not found.**, five of the six of the reliability of separation statistics are statistically significant ($p < .01$). For the Number Discrimination subtask, the largest reliability of separation index is >0.99 for Items. For this subtask, there is good differentiation for the Items (>0.99), and while the Girls index is smaller (0.67), it is also significant. This is not unexpected as we can see on the variable map that most of the girls have not performed well, and cluster together at the bottom of the logit scale. Province (0.69), Girl's Home Language (0.73), and Enumerator (0.68) were also significant, indicating there may be substantive differences of note for these facets. The Urbanicity (0.00) statistic was not significant, indicating less variability between urban and rural schools.

Table 44. Facets Results for Baseline EGMA – Number Discrimination Items

	Girls	Item	Province	Girl's Home Language	Enumerator	Urbanicity
<i>Measures</i>						
<i>Mean</i>	2.06	0.00	0.00	0.00	0.00	0.00
<i>SD</i>	2.36	1.97	0.11	0.19	0.40	0.02
<i>n</i>	3194	10	5	6	54	2
<i>INFIT</i>						
<i>Mean</i>	0.96	1.03	1.00	0.97	1.02	0.99
<i>SD</i>	0.47	0.19	0.12	0.09	0.25	0.06
<i>OUTFIT</i>						
<i>Mean</i>	0.98	1.32	1.38	1.18	1.53	1.32
<i>SD</i>	1.39	1.15	0.52	0.42	1.53	0.15
<i>Reliability of Separation</i>	0.67*	>0.99*	0.69*	0.73*	0.65*	0.00
<i>Chi-Square Statistic</i>	9141.6	4933.0	11.2	44.1	268.8	0.6
<i>Degrees of Freedom</i>	3193	9	4	5	53	1

Figure 40 displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, enumerator, and urbanicity for the Missing Number items on the EGMA. For this map, each asterisk represents 39 girls. The girl's achievement measures ranged from -9.90 logits to 6.42 logits ($M = -1.18$, $SD = 2.69$, $N = 3196$).

For the Missing Number subtask, we see that girls in Province Orientale showed slightly higher results than the rest of the provinces, and girls based in Bandundu, Katanga, and Kasai Orientale showed results slightly lower than the average. Girls who identified French or Kilendu, as their home language outperformed others, and girls who reported Kikongo or Swahili as their home language performed below all other languages. For the Enumerator facet, we can identify the groups of enumerators that yielded either higher or lower results on the subtask. For this

analysis, Urbanicity did not show differential results, and the Item facet showed item difficulty ranging from -3.23 logits to 6.99 logits ($M = 0.00$, $SD = .3.59$, $N = 10$).

Table 45**Error! Reference source not found.** shows that the overall model-data fit is mixed. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of $.20$, and all of the infit and one the outfit statistics (Girl) fall in the acceptable range. Also shown in Table 45**Error! Reference source not found.**, all of the reliability of separation statistics are statistically significant ($p < .01$). For the Missing Number subtask, the largest reliability of separation index is >0.99 for Items. For this subtask, there is good differentiation for Items (>0.99), and while the Girls index is smaller (0.73), it is also significant. This is not unexpected as we can see on the variable map that most of the girls have not performed well, and cluster together at the bottom of the logit scale. Province (0.99), Girl's Home Language (0.94), Enumerator (0.68), and Urbanicity (0.87) were also significant, indicating there may be substantive differences of note for these facets.

Measr	+Girl	+Province-	+Language	+Enumerator	+Urbanicity	+Items
7 + *	+	+	+		+	+ 1
6 +	+	+	+		+	+
5 +	+	+	+		+	+ 2
4 + .	+	+	+		+	+
3 + .	+	+	+		+	+
2 + *	+	+	+		+	+ 4
1 + *	+	+	+		+	+
0 + *	PO		F Kil	41 18 19 22 49 17 26 3 33 35 46 47 13 29 36 48 10 14 21 28 37 38 8 1 12 15 2 20 23 24 31 32 34 4 43 44 52 54 55 6 7 9 11 25 30 39 40 45 5 51 53 56	Rural Urban	+
-1 + *	B K KO Kik S	+	+	27	+	+ 6 5 8 + 3 9
-2 + *	+	+	+		+	+ 7
-3 + *	+	+	+		+	+ 10
-4 + .	+	+	+		+	+
-5 + *	+	+	+		+	+
-6 + .	+	+	+		+	+
-7 + .	+	+	+		+	+
-8 + *	+	+	+		+	+
Measr	* = 39	+Province-	+Language	+Enumerator	+Urbanicity	+Items

Figure 40. Variable Map for Baseline EGMA - Missing Number Items

Table 45. Facets Results for Baseline EGMA – Missing Number Items

	Girls	Item	Province	Girl's Home Language	Enumerator	Urbanicity
<i>Measures</i>						
<i>Mean</i>	-1.18	0.00	0.00	0.00	0.00	0.00
<i>SD</i>	2.69	3.59	0.43	0.28	0.65	0.08
<i>n</i>	3196	10	5	6	55	2
<i>INFIT</i>						
<i>Mean</i>	0.83	0.96	0.96	0.98	0.94	0.95
<i>SD</i>	1.04	0.12	0.07	0.07	0.56	0.04
<i>OUTFIT</i>						
<i>Mean</i>	0.86	3.21	4.38	4.27	2.13	5.25
<i>SD</i>	1.90	3.24	3.45	3.32	2.81	2.37
<i>Reliability of Separation</i>	0.73*	>0.99*	0.99*	0.94*	0.81*	0.87*
<i>Chi-Square Statistic</i>	11401.5	15969.7	207.3	61.4	981.1	7.7
<i>Degrees of Freedom</i>	3195	9	4	5	54	1

Figure 41 **Error! Reference source not found.** displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, enumerator, and urbanicity for the Addition items on the EGMA. For this map, each asterisk represents 14 girls. The girl's achievement measures ranged from -9.16 logits to 9.61 logits ($M = 0.03$, $SD = 3.42$, $N = 2971$).

For the Addition subtask, we see that girls in Province Orientale showed slightly higher results than the rest of the provinces, and girls based in Bandundu and Equateur showed results slightly lower than the average. Girls who identified Kikongo and Lingala as their home language performed below all other languages. For the Enumerator facet, we can identify the groups of enumerators that yielded either higher or lower results on the subtask. For this analysis,

Urbanicity did not show differential results, and the Item facet showed item difficulty ranging from -6.66 logits to 7.33 logits ($M = 0.00$, $SD = 4.02$, $N = 20$).

Table 46**Error! Reference source not found.** shows that the overall model-data fit is mixed. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of $.20$, and all of the infit and one the outfit statistics (Girls) fall in the acceptable range. Also shown in Table 46**Error! Reference source not found.**, all of the reliability of separation statistics are statistically significant ($p < .01$). For the Addition subtask, the largest reliability of separation index is >0.99 for Items. For this subtask, there is good differentiation for Items (>0.99) and Girls (0.92). Province (0.96), Girl's Home Language (0.89), Enumerator (0.84), and Urbanicity (0.97) were also significant, indicating there may be substantive differences of note for these facets.

Table 46. Facets Results for Baseline EGMA – Addition Items

	Girls	Item	Province	Girl's Home Language	Enumerator	Urbanicity
<i>Measures</i>						
<i>Mean</i>	0.03	0.00	0.00	0.00	0.00	0.00
<i>SD</i>	3.42	4.02	0.21	0.16	0.50	0.13
<i>n</i>	2971	20	5	5	55	2
<i>INFIT</i>						
<i>Mean</i>	0.93	0.93	0.92	0.90	0.94	0.93
<i>SD</i>	0.71	0.16	0.07	0.09	0.32	0.03
<i>OUTFIT</i>						
<i>Mean</i>	1.02	4.16	6.18	7.54	3.80	7.47
<i>SD</i>	2.09	3.38	3.36	2.43	3.66	2.17
<i>Reliability of Separation</i>	0.92*	>0.99*	0.96*	0.89*	0.84*	0.97*
<i>Chi-Square Statistic</i>	31314.0	36752.0	104.5	93.4	589.4	33.7
<i>Degrees of Freedom</i>	2970	19	4	4	54	1

Figure 42Error! Reference source not found. displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, enumerator, and urbanicity for the Subtraction items on the EGMA. For this map, each asterisk represents 72 girls. The girl's achievement measures ranged from -7.95 logits to 8.24 logits (M = -2.23, SD = 3.23, N = 3196).

For the Subtraction subtask, girls who identified French as their home language performed better than other languages, and those who reported Kilendu performed below all other languages. For the Enumerator facet, we can identify the groups of enumerators that yielded either higher or lower results on the subtask. For this analysis, Urbanicity did not show differential results, and the Item facet showed item difficulty ranging from -5.04 logits to 4.72 logits (M = 0.00, SD = 3.09, N = 20).

Table 47**Error! Reference source not found.** shows that the overall model-data fit is mixed. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of .20, and all of the infit and one the outfit statistics (Girls) fall in the acceptable range. Also shown in Table 47, five of the six reliability of separation statistics are statistically significant ($p < .01$). For the Subtraction subtask, the largest reliability of separation index is >0.99 for Items. For this subtask, there is good differentiation for the Items (>0.99) and Girls (0.97). Province (0.90), Girl's Home Language (0.95), and Enumerator (0.79) were also significant, indicating there may be substantive differences of note for these facets. The Urbanicity (0.16) statistic was not significant, indicating less variability between urban and rural schools.

Table 47. Facets Results for Baseline EGMA – Subtraction Items

	Girls	Item	Province	Girl's Home Language	Enumerator	Urbanicity
<i>Measures</i>						
<i>Mean</i>	-2.23	0.00	0.00	0.00	0.00	0.00
<i>SD</i>	3.23	3.09	0.12	0.23	0.39	0.02
<i>n</i>	3196	20	5	6	55	2
<i>INFIT</i>						
<i>Mean</i>	0.96	0.95	0.96	0.95	0.93	0.96
<i>SD</i>	0.61	0.15	0.09	0.13	0.30	0.01
<i>OUTFIT</i>						
<i>Mean</i>	1.03	1.90	1.73	1.74	1.80	1.90
<i>SD</i>	1.83	1.07	0.54	0.75	1.86	0.08
<i>Reliability of Separation</i>	0.87*	>0.99*	0.90*	0.95*	0.79*	0.16
<i>Chi-Square Statistic</i>	25970.3	24480.1	43.0	78.7	403.5	1.2
<i>Degrees of Freedom</i>	3195	19	4	5	54	1

Baseline – Subjective Measures

Figure 43 displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, enumerator, and urbanicity for the General School Perception items in the Girl's Survey done at the Baseline evaluation. For this map, each asterisk represents 21 girls. The girl's trait measures ranged from -0.77 logits to 3.95 logits ($M = 1.08$, $SD = 0.80$, $N = 2467$).

For the General School Perception items, we see that more positive perceptions of the general school environment were more common in Equateur. Girls in Province Orientale, however, were more likely to have an overall less positive perception of their school environment. Girls who identified Lingala and French as their home language were also more likely to have more positive views of their school, while girls who reported their home language as Kilendu has a less positive view. For the Enumerator facet, we can identify the groups of enumerators that yielded either higher or lower results on the subtask. For this analysis, girls who went to school in more rural settings had a more positive view of their schools than those in urban schools.

The Item facet showed item difficulty ranging from -1.04 logits to 0.68 logits ($M = 0.00$, $SD = 0.57$, $N = 9$). The most difficult item for girls to endorse is item 5, "Your classmates and you share books without fighting." with the location at 0.68 logits. The easiest item to agree with, however, is item 9, "Boys and girls have equal opportunity to succeed at this school." at -1.04 logits.

Table 48 shows that the overall model-data fit is good. The expected value of the mean square error statistics (infit and outfit) is >0.99 with a standard deviation of $.20$, and all of the infit

Measr	+Girl	+Province-	+Language	+Enumerator	+Urbanicity	+Items	GENER
3	*****						(4)
2	*****						
1	*****						
0	*****						
-1	*****						
-2	*****						(1)
Measr	* = 21	+Province-	+Language	+Enumerator	+Urbanicity	+Items	GENER

Figure 43. Variable Map for Baseline Survey - General School Perception Items

Table 48. Facets Results for Baseline General School Perception Survey Items

	Girls	Item	Province	Girl's Home Language	Enumerator	Urbanicity
Measures						
Mean	1.08	0.00	0.00	0.00	0.00	0.00
SD	0.80	0.57	0.27	0.18	0.28	0.14
n	2467	9	5	6	55	2
INFIT						
Mean	1.04	0.96	0.99	0.98	1.04	0.99
SD	0.54	0.19	0.15	0.12	0.36	0.15
OUTFIT						
Mean	1.00	1.00	0.99	0.99	0.99	1.00
SD	0.62	0.21	0.11	0.08	0.38	0.13
Reliability of Separation	0.33*	>0.99*	0.99*	0.98*	0.63*	0.99*
Chi-Square Statistic	3073.0	4036.8	578.0	289.9	1001.3	125.8
Degrees of Freedom	2466	8	4	5	54	1

and most of the outfit statistics for all six facets fall in the acceptable range. Also shown in Table 48, all of the reliability of separation statistics are statistically significant ($p < .01$). For the General School Perception survey items, the largest reliability of separation index is >0.99 for the Item facet. This indicates good differentiation for Items (>0.99), and while the Girls index is smaller (0.33), it is also significant. This is not unexpected as we can see on the variable map that most of the girls cluster together, toward the top of the logit scale. The statistics for Province (0.99), Girl's Home Language (0.98), Enumerator (0.63), and Urbanicity (0.99) are significant, indicating there may be substantive differences of note for these facets.

Figure 44**Error! Reference source not found.** displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, enumerator, and urbanicity for the Teacher Perception items in the Girl's Survey done at the Baseline evaluation. For this map, each asterisk represents 39 girls. The girl's trait measures ranged from -0.67 logits to 4.22 logits ($M = 0.92$, $SD = 0.44$, $N = 3049$).

For the Teacher Perception items, we see that more positive perceptions of the teachers were more common in Bandundu. Girls in Province Orientale, however, were more likely to have an overall less positive perception of their teacher. Girls who identified French, Kikongo, and Lingala as their home language were also more likely to have more positive views of their teacher, while girls who reported their home language as Kilendu has a less positive view. For the Enumerator facet, we can identify the groups of enumerators that yielded either higher or lower results on the subtask. For this analysis, there was no difference across urban and rural schools.

The Item facet showed item difficulty ranging from -1.17 logits to 1.90 logits ($M = 0.00$, $SD = 0.67$, $N = 22$). The most difficult item for girls to endorse is item 18, "Your teacher teaches less interesting lessons." with the location at 1.90 logits. The easiest item to agree with, however, is item 12, "Your teacher knows your name." at -1.17 logits.

Table 49**Error! Reference source not found.** shows that the overall model-data fit is good. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of $.20$, and all of the infit and outfit statistics for all six facets fall in the acceptable range. Also shown in Table 49**Error! Reference source not found.**, all of the six the reliability of separation statistics are statistically significant ($p < .01$). For the Teacher Perception survey items, the largest reliability of separation index is >0.99 for the Item facet. This indicates good differentiation for Items (>0.99), and while the Girls index is smaller (0.56), it is also

significant. This is not unexpected as we can see on the variable map that most of the girls cluster together, toward the top of the logit scale. The statistics for Province (0.99), Girl's Home Language (0.98), Enumerator (0.93), and Urbanicity (0.93) are significant, indicating there may be substantive differences of note for these facets.

Measr	+Girl	+Province	+Language	-Enumerator	-Urbanicity	-Items	TEACH
4	.	+	+	+	+		(4)
3	.	+	+	+	+		
2	.	+	+	+	+	Not interesting	
1	*****	+	+	+	+	Mastery	---
0	*****	B E K KO PO	F S T Kil	21 25 29 33 46 51 52 11 36 37 43 44 45 14 20 26 32 34 47 48 49 7 15 17 19 2 28 30 38 5 50 55 6 1 12 16 18 27 35 53 56 8 10 13 22 23 24 3 40 9 31 39 4 41 54	Rural Urban	Interested Notices you Fair Doesn't care Not encouraging Ignores you Criticizes you Knows your name	3 ---
-1	.	+	+	+	+	Helps Teaches math Expects success Respects	Helps cooperation Teaches reading Shares problems 2
-2	+	+	+	+	+		(1)
Measr	* = 39	+Province	+Language	-Enumerator	-Urbanicity	-Items	TEACH

Figure 44. Variable Map for Baseline Survey - Teacher Perception Items

Table 49. Facets Results for Baseline Teacher Perception Survey Items

	Girls	Item	Province	Girl's Home Language	Enumerator	Urbanicity
<i>Measures</i>						
<i>Mean</i>	0.92	0.00	0.00	0.00	0.00	0.00
<i>SD</i>	0.44	0.67	0.16	0.12	0.23	0.03
<i>n</i>	3049	22	5	6	55	2
<i>INFIT</i>						
<i>Mean</i>	1.09	1.04	1.01	0.96	1.04	1.04
<i>SD</i>	0.57	0.33	0.22	0.25	0.43	0.21
<i>OUTFIT</i>						
<i>Mean</i>	1.06	1.06	1.02	0.98	1.03	1.07
<i>SD</i>	0.76	0.39	0.22	0.24	0.48	0.24
<i>Reliability of Separation</i>	0.56*	>0.99*	0.99*	0.98*	0.93*	0.93*
<i>Chi-Square Statistic</i>	6272.3	15518.3	520.5	274.1	2199.5	13.8
<i>Degrees of Freedom</i>	3048	21	4	5	54	1

Figure 45Error! Reference source not found. displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, enumerator, and urbanicity for the Perception of School Violence items in the Girl's Survey done at the Baseline evaluation. For this map, each asterisk represents 26 girls. The girl's trait measures ranged from – 0.67 logits to 4.22 logits (M = 1.28, SD = 1.06, N = 3144).

For the Perception of School Violence items, it should be noted that items were reverse coded to be interpreted such that a lower rating on these items reflects a higher perception of violence occurring in the school and a higher rating reflects a lower perception of violence in the school. For this analysis, we see that perceptions of less school violence occurring were more common in Katango. Girls in Bandundu, however, were more likely to have a greater perception of violence occurring in their schools. Similarly, girls who identified French, Kilendu, and

Swahili as their home language had a perception of less school violence occurring in their schools, while girls who reported their home language as Kikongo, Lingala, and Tshiluba had a less positive view. For the Enumerator facet, we can identify the groups of enumerators that yielded either higher or lower results on the subtask. For this analysis, girls who attended rural schools tended to report less violence in their schools than girls attending urban schools.

The Item facet showed item difficulty ranging from -1.42 logits to 0.83 logits ($M = 0.00$, $SD = 0.71$, $N = 7$). The most difficult item for girls to endorse is item 7, “A student from school pushed, shoved, or hit you.” with the location at 0.83 logits. The easiest item to agree with, however, is item 2, “You are afraid of your teacher.” at -1.42 logits. It should be noted that both of these items were reverse coded for analysis, so that all items reflected a positive environment.

Table 50**Error! Reference source not found.** shows that the overall model-data fit is good. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of $.20$, and all of the infit and outfit statistics for all six facets fall in the acceptable range. Also shown in Table 50**Error! Reference source not found.**, all of the six the reliability of separation statistics are statistically significant ($p < .01$). For the Perception of School Violence survey items, the largest reliability of separation index is >0.99 for the Item facet. This indicates good differentiation for Items (>0.99), and while the Girls index is smaller (0.49), it is also significant. This is not unexpected as we can see on the variable map that most of the girls cluster together, toward the top of the logit scale. The statistics for Province (0.93), Girl’s Home Language (0.94), Enumerator (0.63), and Urbanicity (0.99) are significant, indicating there may be substantive differences of note for these facets.

Measr	+Girl	+Province-	+Language	+Enumerator	+Urbanicity	+Items	VIOLE
4	*****						(4)
3	.						
2	*****						
1	*****						---
0	*****	K E KO PO	F Kil S	15 21 23 28 32 5 7 10 11 12 13 14 16 2 22 27 3 31 38 39 4 40 47 48 52 6 1 18 24 25 30 33 35 37 41 45 50 54 8 9 17 36 44 49 53 55 19 20 43 56 26 29 34 46 51	Rural Urban	S bullies S leaves out S hits you S calls names S threatens T Hits you	3 ---
-1	.						2
-2	.					Afraid of T	---
-3	.						(1)
Measr	* = 26	+Province-	+Language	+Enumerator	+Urbanicity	+Items	VIOLE

Figure 45. Variable Map for Baseline Survey – School Violence Perception Items

Table 50. Facets Results for Baseline School Violence Perception Survey Items

	Girls	Item	Province	Girl's Home Language	Enumerator	Urbanicity
<i>Measures</i>						
<i>Mean</i>	1.28	0.00	0.00	0.00	0.00	0.00
<i>SD</i>	1.06	0.71	0.09	0.15	0.22	0.13
<i>n</i>	3144	7	5	6	55	2
<i>INFIT</i>						
<i>Mean</i>	1.01	1.00	1.03	1.04	1.04	1.03
<i>SD</i>	0.58	0.35	0.05	0.07	0.28	0.03
<i>OUTFIT</i>						
<i>Mean</i>	1.00	1.00	1.00	1.01	1.00	1.00
<i>SD</i>	0.77	0.42	0.10	0.10	0.34	0.03
<i>Reliability of Separation</i>	0.49*	>0.99*	0.93*	0.94*	0.63*	0.99*
<i>Chi-Square Statistic</i>	5602.2	5304.1	87.8	158.0	611.5	96.5
<i>Degrees of Freedom</i>	3143	6	4	5	54	1

Annual – Subjective Measures

Figure 46**Error! Reference source not found.** displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, enumerator's home language, enumerator, and urbanicity for the General School Perception items in the Girl's Survey done at the Annual evaluation. For this map, each asterisk represents 16 girls. The girl's trait measures ranged from -0.95 logits to 3.88 logits ($M = 1.10$, $SD = 0.65$, $N = 2214$).

For the General School Perception items, we see a more positive perception of the school environment from girls in Katanga. Girls in Bandundu and Equateur, however, were more likely to have a less positive perception of their schools. Similarly, girls who identified Swahili as their home language had a more positive perception their school, while girls who reported their home language as Kilendu had a less positive view. Interestingly, Enumerators who reported their home language as French were more likely to have interviewed girls who had more positive views of their schools, and those who reported Kilendu as their home language interviewed girls with a less positive view of their schools. For the Enumerator facet, we can identify the groups of enumerators that yielded either higher or lower results on the subtask. For this analysis, girls who attended rural schools tended to report more positive perceptions of their schools than girls attending urban schools.

The Item facet showed item difficulty ranging from -0.85 logits to 1.58 logits ($M = 0.00$, $SD = 0.73$, $N = 8$). The most difficult item for girls to endorse is item 7, "School is a welcoming place for all students." with the location at 1.58 logits. The easiest item to agree with, however, is item 8, "Boys and girls have the same chance of succeeding at this school." at -0.85 logits.

Table 51**Error! Reference source not found.** shows that the overall model-data fit is good. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of .20, and all of the infit and outfit statistics for all six facets fall in the acceptable range. Also shown in Table 51**Error! Reference source not found.**, all of the six the reliability of separation statistics are statistically significant ($p < .01$). For the General School Perception survey items, the largest reliability of separation index is >0.99 for the Item facet. This indicates good differentiation for Items (>0.99), and while the Girls index is smaller (0.22), it is also significant. This is not unexpected as we can see on the variable map that most of the girls cluster heavily together. The statistics for Province (0.95), Girl's Home Language (0.00), Enumerator's Home Language (0.98), Enumerator (0.84), and Urbanicity (0.99) are significant, indicating there may be substantive differences of note for these facets.

Measr	+Girl	+Province	+GLanguage	+ELanguage	+Enumerator	+Urbanicity	+Items	GENER
4	+	+	+	+	+	+	+	(4)
3	+	+	+	+	+	+	+	+
2	+	+	+	+	+	+	+	---
1	+	+	+	+	+	+	+	+
0	+	+	+	+	+	+	+	+
-1	+	+	+	+	+	+	+	+
Measr	* = 16	+Province	+GLanguage	+ELanguage	+Enumerator	+Urbanicity	+Items	GENER

Figure 46. Variable Map for Annual Survey - General School Perception Items

Table 51. Facets Results for Annual General School Perception Survey Items

	Girls	Item	Province	Girl's Home Language	Enumerator's Home Language	Enumerator	Urbanicity
<i>Measures</i>							
<i>Mean</i>	1.10	0.00	0.00	0.00	0.00	0.00	0.00
<i>SD</i>	0.65	0.73	0.13	0.14	0.45	34	0.11
<i>n</i>	2214	8	5	7	6	59	2
<i>INFIT</i>							
<i>Mean</i>	1.02	1.01	1.05	1.12	0.98	1.10	1.01
<i>SD</i>	0.62	0.59	0.19	0.22	0.12	0.45	0.04
<i>OUTFIT</i>							
<i>Mean</i>	0.99	0.99	1.01	1.10	0.95	1.06	0.99
<i>SD</i>	0.68	0.54	0.18	0.23	0.13	0.45	0.06
<i>Reliability of Separation</i>	0.22*	>0.99*	0.95*	0.00*	0.98*	0.84*	0.98*
<i>Chi-Square Statistic</i>	2697.2	2552.2	114.4	83.4	190.2	653.0	46.3
<i>Degrees of Freedom</i>	2213	7	4	6	5	58	1

Figure 47**Error! Reference source not found.** displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, enumerator's home language, enumerator, and urbanicity for the Teacher Perception items in the Girl's Survey done at the Annual evaluation. For this map, each asterisk represents 64 girls. The girl's trait measures ranged from -3.33 logits to 1.52 logits ($M = -0.07$, $SD = 0.36$, $N = 3120$).

For the Teacher Perception items, we see no differences across the Province, Girl's Home Language, Enumerator's Home Language, Enumerator, or Urbanicity. The Item facet showed item difficulty ranging from -2.69 logits to 2.39 logits ($M = 0.00$, $SD = 1.95$, $N = 22$). The most difficult item for girls to endorse is item 18, "Teachers at this school expect students like you to succeed in life." with the location at 2.39 logits. The easiest item to agree with, however, is item 12, "Your teacher knows your name." at -2.69 logits.

Table 52**Error! Reference source not found.** shows that the overall model-data fit is good. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of .20, and all of the infit and outfit statistics for all six facets fall in the acceptable range. However, as shown in Table 52, only three of the six the reliability of separation statistics are statistically significant ($p < .01$). For the Teacher Perception survey items, the largest reliability of separation index is >0.99 for the Item facet. This indicates good differentiation for Items (>0.99). The Girls index is 0.00, and not significant. This is not unexpected as we can see on the variable map that most of the girls cluster heavily together. The statistics for Province (0.59) and Enumerator (0.64) are significant, indicating there may be substantive differences of note for these facets. However, Girl's Home Language (0.00),

Enumerator's Home Language (0.00), and Urbanicity (0.00) are not significant, indicating that there is not substantive variability for these facets.

Table 52. Facets Results for Annual Teacher Perception Survey Items

	Girls	Item	Province	Girl's Home Language	Enumerator's Home Language	Enumerator	Urbanicity
<i>Measures</i>							
<i>Mean</i>	-0.07	0.00	0.00	0.00	0.00	0.00	0.00
<i>SD</i>	0.36	1.95	0.03	0.03	0.01	0.11	0.01
<i>n</i>	3120	22	5	7	6	59	2
<i>INFIT</i>							
<i>Mean</i>	1.00	0.99	0.98	1.08	0.94	1.03	0.99
<i>SD</i>	0.80	0.39	0.20	0.32	0.25	0.62	0.14
<i>OUTFIT</i>							
<i>Mean</i>	0.99	0.99	0.98	1.07	0.93	1.02	0.98
<i>SD</i>	0.79	0.39	0.20	0.33	0.25	0.62	0.14
<i>Reliability of Separation</i>	0.00	>0.99*	0.59*	0.00	0.00	0.64*	0.00
<i>Chi-Square Statistic</i>	2687.4	74434.3	11.6	7.5	1.2	242.1	0.6
<i>Degrees of Freedom</i>	3119	21	4	6	5	58	1

Figure 48**Error! Reference source not found.** displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, enumerator's home language, enumerator, and urbanicity for the Perception of School Violence items in the Girl's Survey done at the Annual evaluation. For this map, each asterisk represents 29 girls. The girl's trait measures ranged from -2.21 logits to 5.19 logits ($M = 1.59$, $SD = 0.84$, $N = 3136$).

For the Perception of School Violence items, it should be noted that items were reverse coded to be interpreted such that a lower rating on these items reflects a higher perception of violence occurring in the school and a higher rating reflects a lower perception of violence in the school. For this analysis, we see that perceptions of less school violence occurring were more common in Bandundu. Girls in Katanga, however, were more likely to have a greater perception of violence occurring in their schools; this is the opposite result from the Baseline results. Similarly, girls who identified French and Kikongo as their home language had a perception of less school violence occurring in their schools, while girls who reported their home language as Bemba and Kilendu had a less positive view. In addition, Enumerators who reported their home language to be Kikongo were more likely to have interviewed girls who reported slightly lower levels of school violence. For the Enumerator facet, we can identify the groups of enumerators that yielded either higher or lower results on the subtask. For this analysis, there were no differences between girls who attended urban vs. rural schools.

The Item facet showed item difficulty ranging from -1.49 logits to 1.15 logits ($M = 0.00$, $SD = 1.03$, $N = 7$). The most difficult item for girls to endorse is item 7, "A student from school pushed, shoved, or hit you." with the location at 1.15 logits. The easiest item to agree with,

however, is item 2, “Your teacher hits you.” at -1.49 logits. It should be noted that both of these items were reverse coded for analysis, so that all items reflected a positive environment.

Table 53**Error! Reference source not found.** shows that the overall model-data fit is good. The expected value of the mean square error statistics (infit and outfit) is >0.99 with a standard deviation of .20, and all of the infit and outfit statistics for all six facets fall in the acceptable range. Also shown in Table 53**Error! Reference source not found.**, five of the six the reliability of separation statistics are statistically significant ($p < .01$). For the Perception of School Violence survey items, the largest reliability of separation index is >0.99 for the Item facet. This indicates good differentiation for Items (>0.99) and Girls (0.42). This is not unexpected as we can see on the variable map that most of the girls cluster heavily together. The statistics for Province (0.89), Girl’s Home Language (0.60), Enumerator’s Home Language (0.53), and Enumerator (0.85) are significant, indicating there may be substantive differences of note for these facets. However, Urbanicity (0.00) is not significant, indicating that there is not substantive variability for these facets.

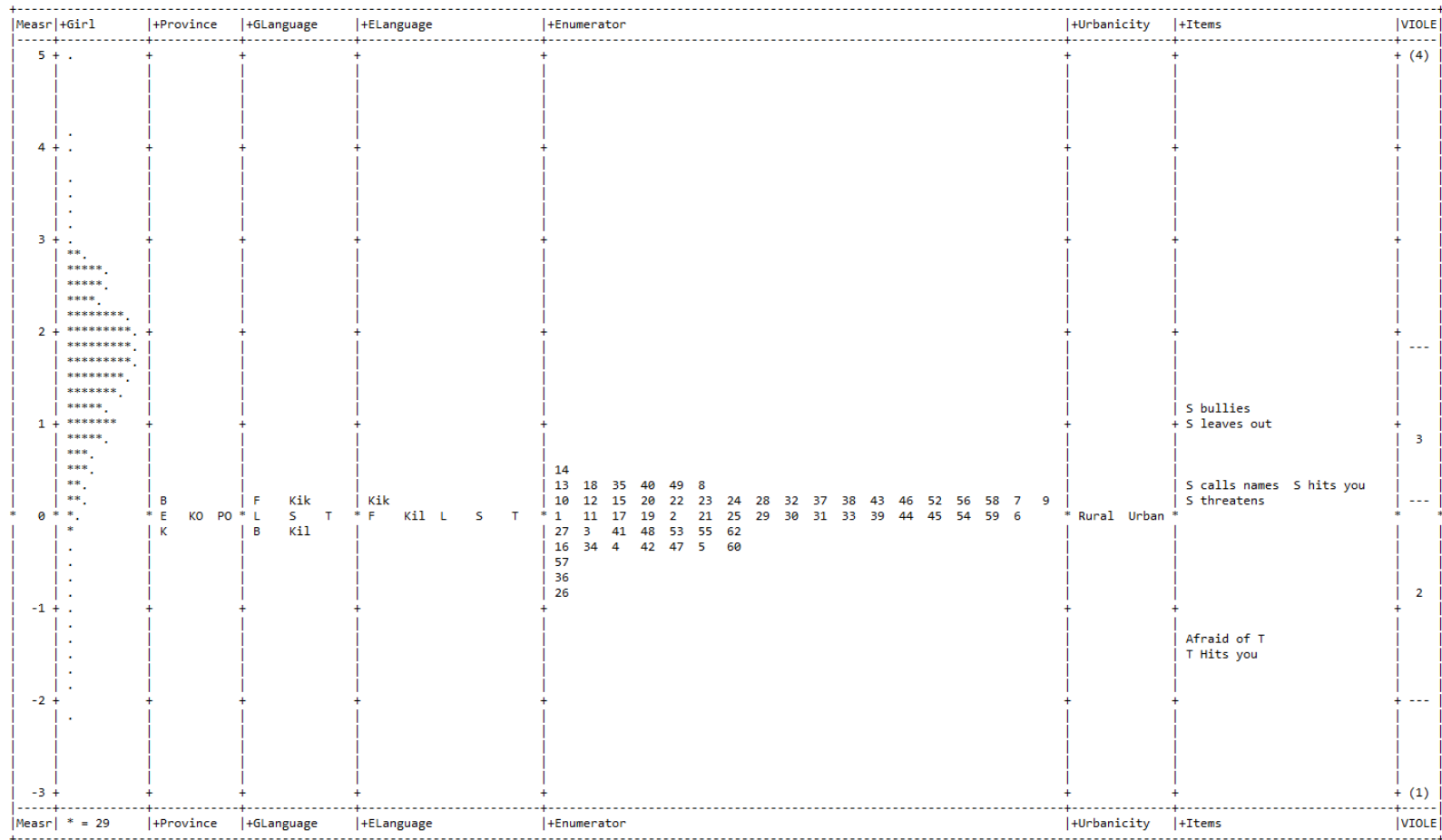


Figure 48. Variable Map for Annual Survey - School Violence Perception Items

Table 53. Facets Results for Annual School Violence Perception Survey Items

	Girls	Item	Province	Girl's Home Language	Enumerator's Home Language	Enumerator	Urbanicity
<i>Measures</i>							
<i>Mean</i>	1.59	0.00	0.00	0.00	0.00	0.00	0.00
<i>SD</i>	0.84	1.03	0.08	0.13	0.08	0.26	0.01
<i>n</i>	3136	7	5	7	6	59	2
<i>INFIT</i>							
<i>Mean</i>	0.96	1.06	1.05	1.08	0.99	1.02	1.07
<i>SD</i>	0.75	0.10	0.21	0.15	0.29	0.42	0.08
<i>OUTFIT</i>							
<i>Mean</i>	0.96	0.96	0.93	0.97	0.89	0.95	0.96
<i>SD</i>	0.89	0.21	0.19	0.15	0.27	0.46	0.08
<i>Reliability of Separation</i>	0.42*	>0.99*	0.89*	0.60*	0.53*	0.85*	0.00
<i>Chi-Square Statistic</i>	5389.5	8235.9	38.2	29.6	29.4	671.0	0.1
<i>Degrees of Freedom</i>	3135	6	4	6	5	58	1

Longitudinal – Subjective Measures

Figure 49 displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, urbanicity, and administration (Baseline vs Annual as a proxy for Informal Translation vs. Formal Translation) for the General School Perception items in the Girl's Survey. For this map, each asterisk represents 15 girls. The girl's trait measures ranged from -0.33 logits to 3.35 logits ($M = 0.86$, $SD = 0.55$, $N = 1398$).

For the General School Perception items, we see that there were slightly less positive perceptions of the general school environment Province Orientale. Girls who identified Kikongo and Lingala as their home language were also more likely to have more positive views of their school, while girls who reported their home language as Kilendu has a less positive view. For this analysis, girls who went to school in more rural settings had a more positive view of their schools than those in urban schools. In addition, more positive ratings were associated with the Baseline administration of the items where informal translations of the survey by the enumerators occurred.

The Item facet showed item difficulty ranging from -0.81 logits to 0.71 logits ($M = 0.00$, $SD = 0.51$, $N = 7$). The most difficult item for girls to endorse is item 7, "School is a welcoming place for all students." with the location at 0.71 logits. The easiest item to agree with, however, is item 8, "Boys and girls have equal opportunity to succeed at this school." at -0.81 logits.

Table 54 shows that the overall model-data fit is good. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of $.20$, and all of the infit

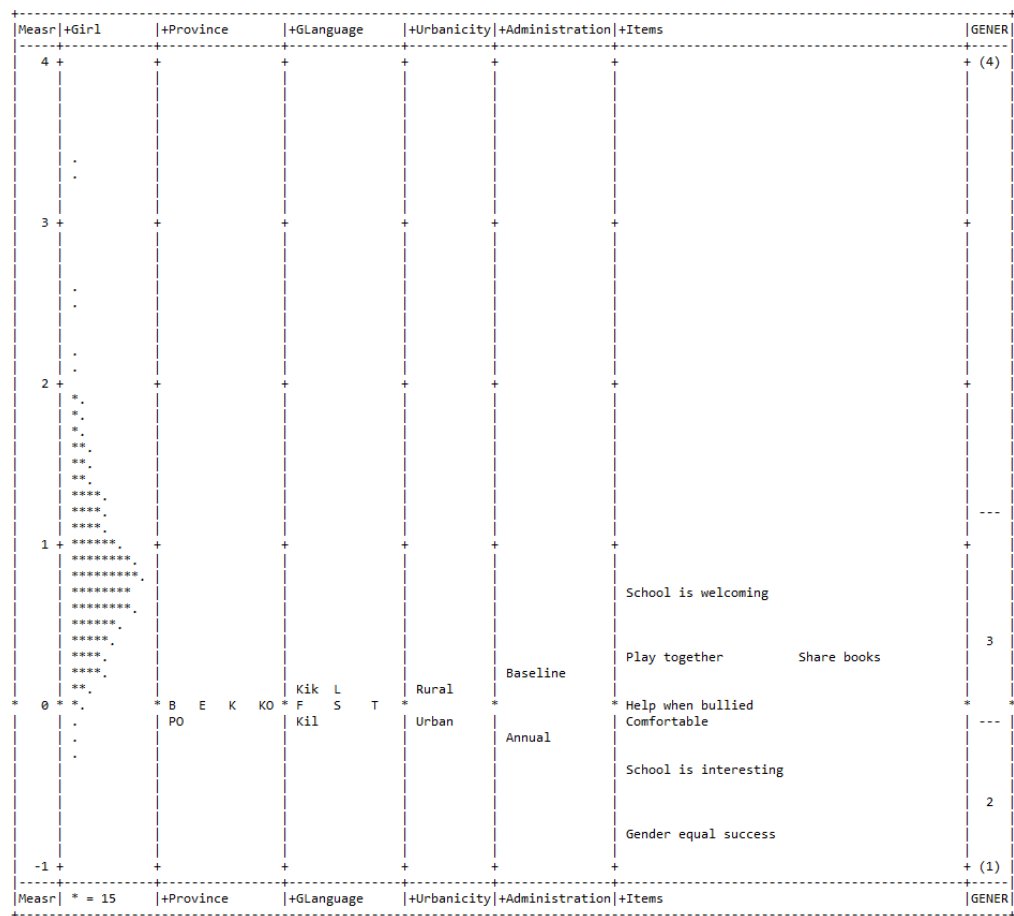


Figure 49. Variable Map for Longitudinal Survey Items - General School Perception Items

Table 54. Facets Results for Longitudinal General School Perception Survey Items

	Girls	Item	Province	Girl's Home Language	Urbanicity	Administration
<i>Measures</i>						
<i>Mean</i>	0.86	0.00	0.00	0.00	0.00	0.00
<i>SD</i>	0.55	0.51	0.06	0.08	0.07	0.21
<i>n</i>	1398	7	5	6	2	2
<i>INFIT</i>						
<i>Mean</i>	1.02	0.97	1.04	0.97	1.00	1.03
<i>SD</i>	0.37	0.35	0.12	0.16	0.07	0.33
<i>OUTFIT</i>						
<i>Mean</i>	0.98	0.98	1.00	0.94	0.98	0.98
<i>SD</i>	0.42	0.30	0.13	0.17	0.08	0.24
<i>Reliability of Separation</i>	0.54*	>0.99*	0.82*	0.28	0.97*	>0.99*
<i>Chi-Square Statistic</i>	2319.0	2619.8	9.3	4.5	30.0	277.4
<i>Degrees of Freedom</i>	1397	6	4	5	1	1

and most of the outfit statistics for all six facets fall in the acceptable range. Also shown in Table 54, five of the six reliability of separation statistics are statistically significant ($p < .01$). For the General School Perception survey items, the largest reliability of separation index is >0.99 for the Item facet. This indicates good differentiation for Items (>0.99), and while the Girls index is lower (0.54), it is significant. This is not unexpected as we can see on the variable map that most of the girls cluster heavily together. The statistics for Province (0.82), Urbanicity (0.97), and Administration (>0.99) are significant, indicating there may be substantive differences of note for these facets. However, Girl's Home Language (0.28) is not significant, indicating that there is not substantive variability for these facets.

Figure 50**Error! Reference source not found.** displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, urbanicity, and administration (Baseline vs Annual as a proxy for Informal Translation vs. Formal Translation) for the Teacher Perception items in the Girl's Survey. For this map, each asterisk represents 30 girls. The girl's trait measures ranged from -0.50 logits to 1.38 logits ($M = 0.54$, $SD = 0.27$, $N = 2230$).

For the Teacher Perception items, there were no differences for the Province, Girl's Home Language, or Urbanicity facets. we see that more positive perceptions of the teachers were more common in Bandundu. However, as with the General School Perception items, more positive ratings were associated with the Baseline administration of the items where informal translations of the survey by the enumerators occurred.

The Item facet showed item difficulty ranging from -0.71 logits to 0.82 logits ($M = 0.00$, $SD = 0.51$, $N = 22$). The most difficult item for girls to endorse is item 10, "Teacher at this school expect students like you to succeed." with the location at 0.82 logits. The easiest item to agree with, however, is item 18, "Your teacher teaches less interesting lessons." at -0.71 logits.

Table 55**Error! Reference source not found.** shows that the overall model-data fit is good. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of .20, and all of the infit and outfit statistics for all six facets fall in the acceptable range. Also shown in Table 55**Error! Reference source not found.**, all of the six the reliability of separation statistics are statistically significant ($p < .01$). For the Teacher Perception survey items, the largest reliability of separation index is >0.99 for the Item facet. This indicates good differentiation for Items (>0.99), and while the Girls index is lower (0.53), it is significant.

This is not unexpected as we can see on the variable map that most of the girls cluster heavily together. The statistics for Province (0.91), Girl's Home Language (0.00), Urbanicity

Measr	+Girl	+Province	+Language	+Urbanicity	+Administration	+Items		TEACH	
2 +	+		+	+		+		+ (4)	
.								---	
.									
1 + **	+		+	+		+		+	

*****.						Expects success	Respects		
*****.						Fair	Notifies you	3	
*****.					Baseline	Interested	Encourages participation	Helps	Shares problems
*****.						Cares			
*****.									
*****.									
**									
* 0 *	*	B E K KO PO F Kik Kil L S T	Rural Urban					*	
.						Criticizes you	Gets angry	Ignores you	Not encouraging
.						Doesn't care	Knows your name		
.									
.									
.					Annual	Helps cooperation	Teaches math	Teaches reading	
.						Mastery			2
.						Calls on you	Encouraging	Not interesting	
-1 +	+		+	+		+			+ (1)
Measr	* = 30	+Province	+Language	+Urbanicity	+Administration	+Items		TEACH	

Table 55. Facets Results for Longitudinal Teacher Perception Survey Items

	Girls	Item	Province	Girl's Home Language	Urbanicity	Administration
<i>Measures</i>						
<i>Mean</i>	0.54	0.00	0.00	0.00	0.00	0.00
<i>SD</i>	0.27	0.51	0.03	0.02	0.02	0.75
<i>n</i>	2230	22	5	6	2	2
<i>INFIT</i>						
<i>Mean</i>	0.99	0.98	0.97	0.95	1.00	1.04
<i>SD</i>	0.37	0.43	0.10	0.09	0.01	0.55
<i>OUTFIT</i>						
<i>Mean</i>	1.03	1.03	1.00	0.98	1.03	1.03
<i>SD</i>	0.40	0.44	0.10	0.10	0.03	0.53
<i>Reliability of Separation</i>	0.53*	>0.99*	0.91*	0.00*	0.94*	>0.99*
<i>Chi-Square Statistic</i>	4708.7	15422.8	39.3	30.4	16.1	18742.5
<i>Degrees of Freedom</i>	2229	21	4	5	1	1

(0.94), and Administration (>0.99) are significant, indicating there may be substantive differences of note for these facets.

Figure 51 displays a variable map representing the calibrations of the girls, items, girl's home province, girl's home language, urbanicity, and administration (Baseline vs Annual as a proxy for Informal Translation vs. Formal Translation) for the Perception of School Violence items in the Girl's Survey. For this map, each asterisk represents 24 girls. The girl's trait measures ranged from -0.57 logits to 4.00 logits ($M = 1.09$, $SD = 0.57$, $N = 2293$).

For the Perception of School Violence items, it should be noted that items were reverse coded to be interpreted such that a lower rating on these items reflects a higher perception of violence occurring in the school and a higher rating reflects a lower perception of violence in the school. For this analysis, we see that perceptions of slightly less school violence occurring were more common in Province Orientale. Similarly, girls who identified French as their home language had a perception of slightly less school violence occurring in their schools. For this analysis, girls who attended rural schools tended to report less violence in their schools than girls attending urban schools, and there was no difference between the administrations.

The Item facet showed item difficulty ranging from -1.03 logits to 0.78 logits ($M = 0.00$, $SD = 0.62$, $N = 7$). The most difficult item for girls to endorse is item 7, “A student from school pushed, shoved, or hit you.” with the location at 0.78 logits. The easiest item to agree with, however, is item 2, “You are afraid of your teacher.” at -1.03 logits. It should be noted that both of these items were reverse coded for analysis, so that all items reflected a positive environment. Table 56 shows that the overall model-data fit is good. The expected value of the mean square error statistics (infit and outfit) is 1.00 with a standard deviation of $.20$, and all of the infit and outfit statistics for all six facets fall in the acceptable range. Also shown in Table 56

Error!
Reference source not found., all of the six the reliability of separation statistics are statistically significant ($p < .01$). For the Perception of School Violence survey items, the largest reliability of separation index is >0.99 for the Item facets. This indicates good spread of the items on the latent variable. While the Girls index was smaller (0.53), it was significant. This is not surprising, as the girl’s scores cluster together. The statistics for, Province (0.91), Girl’s Home Language (0.00),

Urbanicity (0.94), and Administration (0.72) are significant, indicating there may be substantive differences of note for these facets.

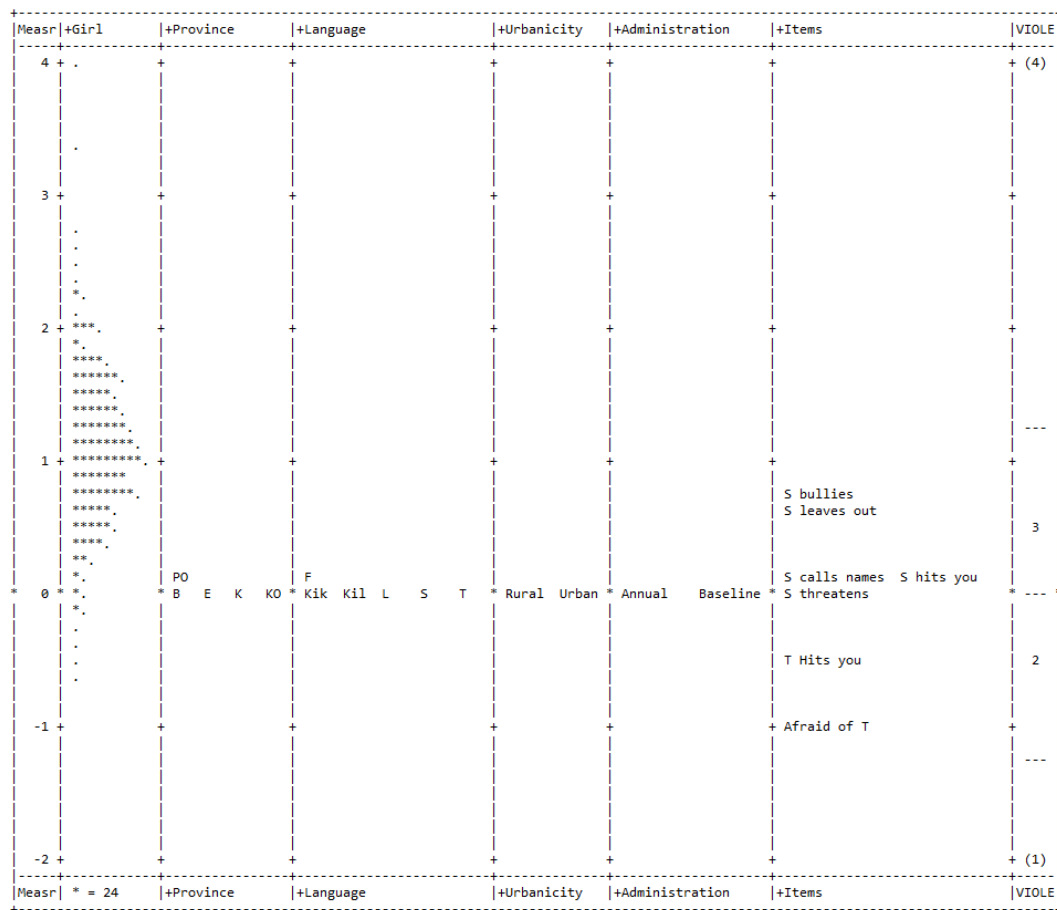


Figure 51. Variable Map for Longitudinal Survey Items - School Violence Perception Items

Table 56. Facets Results for Longitudinal School Violence Perception Survey Items

	Girls	Item	Province	Girl's Home Language	Urbanicity	Administration
<i>Measures</i>						
<i>Mean</i>	1.09	0.00	0.00	0.00	0.00	0.00
<i>SD</i>	0.57	0.62	0.05	0.05	0.04	0.02
<i>n</i>	2293	7	5	6	2	2
<i>INFIT</i>						
<i>Mean</i>	1.03	1.05	1.04	1.03	1.05	1.05
<i>SD</i>	0.40	0.13	0.10	0.08	0.01	0.15
<i>OUTFIT</i>						
<i>Mean</i>	0.98	0.98	0.96	0.95	0.99	0.99
<i>SD</i>	0.47	0.16	0.10	0.09	0.00	0.08
<i>Reliability of Separation</i>	0.57*	>0.99*	0.82*	0.00	0.93*	0.72*
<i>Chi-Square Statistic</i>	5076.1	6335.3	12.5	8.1	13.9	3.6
<i>Degrees of Freedom</i>	2292	6	4	5	1	1

CHAPTER V. DISCUSSION, LIMITATIONS AND IMPLICATIONS FOR RESEARCH, AND IMPLICATIONS FOR PRACTICE AND POLICY

Summary of the Study

The primary purpose of this study is to explore the use of Generalizability Theory (GTheory; Brennan, 1992, 2001; Shavelson & Webb, 1991) and Rasch Measurement Theory (Rasch, 1980; Wright & Masters, 1982) in the form of the Many-Facet (MF) model (Linacre, 1989) to assess possible sources of unreliability (error) in data from an international evaluation to be used as evidence of success in outcomes of an educational initiative. Recall that reliability is broadly defined as “the desired consistency (or reproducibility) of scores” (Crocker & Algina, 2008), and depends heavily “on characteristics of the test, the conditions of administration, and the group of examinees” (Traub & Rowley, 1991).

In the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014), eight standards, or aspects of reliability evidence are outlined (shown in Figure 5). This study focuses primarily on the fourth standard of reliability involving an examination of possible factors that may affect the reliability coefficient or the precision of measurement. These factors included the language of administration, mother tongue language of the examinees, the use of enumerators (or raters), province of administration, urbanicity, adaptation method, and the items of the assessments and surveys themselves.

To that end, in this dissertation I sought to answer three questions:

1. What are the largest sources measurement error in the current evaluation design, and how do they differ for subjective vs. objective measures?

2. What is the effect the of non-standard translation and adaptation procedures used on the assessments throughout the VAS-Y Fille! evaluation?
3. What facets are modifiable in a program such as VAS-Y Fille! that would allow for a decrease in the measurement error of the outcome measures?

In the following chapter, I begin with tables providing an overview of the results presented in Chapter IV, and then address each research question in order based on the results of the analyses. I conclude the chapter with the limitations of this study, and a discussion of the implications of the research for international development, evaluation, and measurement.

Discussion

The following three tables provide an overview of the results presented in CHAPTER IV. RESULTSTable 57 shows summary results of the GTheory analyses by outcome measure. Each row indicates the subtask or sub-scale of the objective or subjective measures used in the evaluation, and the columns indicate the facets for which variance components were estimated. Recall that there were five sets of analyses completed on each subtask/subscale, each of which included two home languages. The ratio in the corresponding column and row of this table signifies the number of analyses where the facet accounted for over 5% of the total variance, out of the total analyses completed including that facet. For example, for the Letter Name subtask of the EGRA, all five analyses had over 5% of the total variance accounted for by the Item facet, and one of the five for the Girl's Home Language Facet. For the two analyses that included the Enumerator facet, one of these analyses showed over 5% of the total variance accounted for by Enumerator. None of the three interaction facets had over 5% of the total variance accounted for on this subtask.

Table 57. GTheory Results by Subtask with the Number of Analyses Out of the Total Analyses Completed Including the Facet Accounting for Over 5% of the Total Variance

	Item	Girl's Home Lang.	Enum.	Item by Lang.	Item by Enum.	Lang. by Enum.
Objective Measures						
EGRA						
Letter Name	5/5	1/5	1/2	0/2	0/2	0/2
Nonword Reading	5/5	1/5	1/2	0/2	1/2	1/2
Oral Reading Fluency	3/5	2/5	1/2	0/2	0/2	1/2
Reading Comprehension	1/5	0/5	1/2	0/2	0/2	0/2
Listening Comprehension	3/5	2/5	1/2	0/2	1/2	0/2
EGMA						
Number Identification	5/5	2/5	1/2	0/2	0/2	1/2
Number Discrimination	4/5	1/5	1/2	0/2	0/2	0/2
Missing Number	5/5	0/5	2/2	0/2	2/2	1/2
Addition	5/5	1/5	0/2	0/2	0/2	0/2
Subtraction	5/5	1/5	0/2	0/2	0/2	1/2
Subjective Measures						
Baseline						
General School	5/5	0/5	0/2	0/2	2/2	0/2
Teacher Perceptions	5/5	0/5	1/2	0/2	2/2	0/2
School Violence	5/5	0/5	1/2	0/2	0/2	0/2
Annual*						
General School	0/1	0/1	1/1	0/1	1/1	0/1
Teacher Perceptions	1/1	0/1	0/1	0/1	1/1	0/1
School Violence	1/1	0/1	0/1	0/1	1/1	0/1

Table 58 provides a summary of the Facets results for the analyses including a single time point, and Table 59 provides the summary for the longitudinal analyses that include data from both the Baseline and Annual data collection points. Each row indicates the subtask or sub-scale of the objective or subjective measures used in the evaluation, and the columns indicate the

facets included in the analysis. For these tables, there are two indicators of practical differences. The first is the statistical significance of the reliability of separation statistics for each facet indicated in Table 38 through Table 56. Recall that the reliability of separation statistic tests the hypothesis of whether or not there are significant differences between the elements within a facet. If the statistic is significant, it shows there is spread of the facet along whatever the latent variable is for the analysis at hand. This type of spread is what we look for in the Girls and Items facets, as we want to see spread along the latent variable for both the girls, in terms of their scores spanning the score scale, and the items, in terms of spanning the spectrum of difficulty (that is, we want to see items that are more difficult, less difficult, and moderately difficult). For the other facets of interest, however, significant spread is indicative of a possible problem. We do not want to see significant spread across the language of administration, for example, as this tells us that scores on the task are dependent on the language the assessment was administered in. While this statistic can provide us with good information, it calculated using a chi-square statistic and is highly influenced by large sample sizes. That is, smaller and smaller differences are needed for statistical significance as sample size increases.

Therefore, the second indicator used in these tables is the difference between the elements in each facet with the smallest and the largest logit. If this difference is larger than 1.0 logit, we take this as practical significance as it indicates that, for example, girls who speak French at home versus Kilendu have higher scores, on average, on the Letter Name subtask of the EGRA (Figure 33). In the tables below, the ratios in each cell indicate how many of the two indicators were flagged for each facet and subtask/subtest.

Table 58. Facets Results for Single Time Points with the Number of Flags Indicating Variability of the Elements in the Facet on the Logit Scale

	Girls	Item	Prov.	Girl's Home Lang.	Enum.	Urban.
Objective Measures						
EGRA						
<i>Letter Name</i>	2/2	2/2	1/2	2/2	2/2	1/2
<i>Nonword Reading</i>	2/2	2/2	1/2	1/2	2/2	1/2
<i>Oral Reading Fluency</i>	2/2	2/2	1/2	2/2	2/2	1/2
<i>Reading Comprehension</i>	2/2	2/2	0/2	1/2	2/2	0/2
<i>Listening Comprehension</i>	2/2	2/2	1/2	1/2	2/2	0/2
EGMA						
<i>Number Identification</i>	2/2	2/2	1/2	2/2	2/2	1/2
<i>Number Discrimination</i>	2/2	2/2	1/2	1/2	2/2	0/2
<i>Missing Number</i>	2/2	2/2	2/2	1/2	2/2	1/2
<i>Addition</i>	2/2	2/2	1/2	1/2	2/2	1/2
<i>Subtraction</i>	2/2	2/2	1/2	1/2	2/2	1/2
Subjective Measures						
Baseline						
<i>General School</i>	2/2	2/2	1/2	1/2	2/2	1/2
<i>Teacher Perceptions</i>	2/2	2/2	1/2	1/2	1/2	1/2
<i>School Violence</i>	2/2	2/2	1/2	1/2	1/2	1/2
Annual						
<i>General School</i>	2/2	2/2	1/2	1/2	2/2	1/2
<i>Teacher Perceptions</i>	1/2	2/2	1/2	0/2	1/2	0/2
<i>School Violence</i>	2/2	2/2	1/2	1/2	2/2	0/2

Table 59. Summary Facets Results for Longitudinal Analyses

	Girls	Item	Prov.	Girl's Home Lang.	Urban.	Admin.
<i>Longitudinal</i>						
<i>General School</i>	2/2	2/2	1/2	0/2	1/2	1/2
<i>Teacher Perceptions</i>	2/2	2/2	1/2	1/2	1/2	2/2
<i>School Violence</i>	2/2	2/2	1/2	0/2	1/2	1/2

Research Question 1: What are the largest sources measurement error in the current evaluation design, and how do they differ for subjective vs. objective measures?

Overwhelmingly, for the objective measures, the Girl's Home Language/Language of Administration and Enumerator facets contribute to measurement error, both the Facets and GTheory results show this pattern. Girl's Home Language/Language of Administration was notable for both the EGRA and EGMA subtasks. In Table 58 we see both flags for the Facets results on two EGRA subtasks (Letter Name and Oral Reading Fluency), and one EGMA subtask (Number Identification), and in all cases, the girls who identified French as their home language performed better than all other languages. Table 57 also shows there were differences across results due to language in the GTheory analyses. These results give a similar picture as the Facets results. Those analyses flagged in Table 57 for Girls Home Language were those that included French. For all but one EGRA subtask (Reading Comprehension) and one EGMA subtask (Missing Number), the Lingala \times French analyses resulted in more than 5% of variance attributable to language differences. The Swahili \times French analyses resulted in flags for two EGRA subtasks (Oral Reading Fluency and Listening Comprehension) and one EGMA subtask (Number Identification). Other language combinations did not result in a substantive proportion

of variance. This pattern of results around language indicates that there are definite achievement gaps on the objective assessments based on the girl's home language.

The Enumerator facet was also of note for both the EGRA and EGMA, flagged for at least one analysis in all but two EGMA subtasks (Addition and Subtraction), and particularly for the Swahili \times French analysis results from the GTheory analyses. Facets results showed sometimes very large differences on the Enumerator facet across all subtasks for both the EGRA and EGMA.

From the GTheory analyses, two of the interaction terms also resulted in flags. The Item by Enumerator interaction appeared relevant for two EGRA subtasks (Nonword Reading and Listening Comprehension) and one EGMA subtasks (Missing Number). The girls tended to struggle with these subtasks, and all three require a great deal of instruction and active listening by the girls. In addition, the Language by Enumerator interaction was also large across two of the EGRA subtasks (Nonword Reading and Oral Reading Fluency) and three of the EGMA subtasks (Number Identification, Missing Number, and Subtraction), specifically for the Swahili \times Kilendu analysis results.

Results on the subjective measures differed somewhat. The Language of Administration of the surveys showed no effect in any of the GTheory analyses, and in no case were both Facets flags indicated. The Enumerator facet appeared relevant in very few cases, specifically in the Swahili \times French analysis results for the Teacher Perception and Perceptions of School Violence items at Baseline, and the General School items at Annual. Whereas Facets results showed differences across the enumerators at Baseline and Annual for the General School Perception survey items, and at Annual for the Perception of School Violence survey items. More so than for the objective measures, the Item by Enumerator interaction effect from the GTheory analyses was flagged for the subjective measures. In general, this indicates that responses depend on both the

enumerator and the language of administration, with some enumerators eliciting more positive responses when the Language of Administration is Swahili, and some when it is French.

Finally, for both the subjective and objective measure results, there are differences across the five GTheory analyses done for each outcome measure. That is, the facets that accounted for a substantive amount of variance in the model differed depending on which language combination was included in the analyses. As was noted, for example, the Lingala \times French and Swahili \times French analyses tended to result in significant error in the language facet. And, for the Language by Enumerator interaction, the Swahili \times Kilendu results tended to show a greater proportion of variance than other language combinations. These inconsistencies, particularly when language is included in the analyses, may be a further indication that language (be it of administration or home language) is a significant factor to consider in evaluations like this.

Research Question 2: What is the effect the of non-standard translation and adaptation procedures used on the assessments throughout the VAS-Y Fille! evaluation?

The results in Table 59 are the most relevant in answering this question, as these analyses in Facets included survey data from both Baseline and Annual for the same girls at each time point. As a reminder, the surveys were officially translated for the Annual administration, and done on-the-fly at Baseline by the enumerators. Facets results for these analyses showed that for the Teacher Perception survey items only, there was a difference in responses between the two time points, with Baseline responses being more positive, on average, than at Annual. This scale is significantly longer than the other two, with 22 items, including several negatively-worded items that may have proven difficult for enumerators to translate on-the-fly.

When reviewing GTheory results for the surveys across Baseline and Annual, none of the facets that include language had any notable error associated with them, and therefore, indicate no

real differences due to language between the two time points. These results suggest that there may be no translation effect of note. This could indicate that enumerators were well-versed in the required languages.

Research Question 3: What facets are modifiable in a program such as VAS-Y Fille! that would allow for a decrease in the measurement error of the outcome measures?

As has been addressed in the discussion of research questions 1 and 2, one of the biggest sources of error variance in the GTheory analyses is the Enumerator, and the facets that include Enumerator in the interaction effects. This is consistent for both the objective and subjective measures, and across both analyses. Facets results also show that not only are the reliability of separation statistics significant, but the practical differences across enumerators is a factor across all EGRA and EGMA subtasks, and most of the survey subtests. Given its prevalence in the results, focusing on reducing variability across the enumerators could result in a large change in overall measurement error.

The other facet of concern is Language of Administration/Girl's Home Language, particularly for the objective measures used in this evaluation. While we cannot control the alignment of Girl's Home Language and Language of Administration, it is imperative that results like those in this study are considered and discussed, when designing both literacy and numeracy interventions, and when choosing or designing outcome measures.

Limitations

It should be noted that unlike more typical GTheory analyses, the focus here was on enumerators, language, and items included in the assessments and the survey. As a result of the need for a crossed design, the person facet was not modeled. Including language as a facet

generally necessitates a nested model unless you require respondents to complete outcome measures in more than one language. Nested models, as discussed, do not allow for a thorough investigation of each facet individually.

As the Person facet was not modeled due to its nested nature in the design, this begs the question, where does the error associated with this facet go? What does this do to the remaining estimated components? This could mean that the error term is inflated. The decision to exclude the person facet also means we cannot calculate the typical GTheory reliability coefficients as they require variance estimates specific to the person. In almost all cases, there was a significant amount of variability attributable to the error term. This indicates the presence of facets not modeled in this set of analyses. Possibly, this is due to the absence of the Person facet, or it may be due to more of the nested facets not modeled such as Province, Urbanicity, etc. Still, the value of investigating the facets in the current study should not be understated. Where we are able to identify sources of measurement error, we know that there will be an effect on a final reliability estimate.

As discussed, a crossed design was created for the purpose of this analysis. Therefore, the dataset was very limited in size compared to the full dataset. The necessity of a crossed design for detailed information limits the utility of the analyses for this type of data. Future research could include the utility of a true crossed design at pilot stage, with appropriate sampling measures in place to inform the full evaluation design, and providing the points of concern to the team implementing the full evaluation.

The Item facet is also of interest from an overall reliability and validity standpoint. The variance associated with this facet in the GTheory analyses was highly differential across the different analyses, both subjective and objective, indicating possible large differences in

performance across Home Language/Language of Administration. These differences point to problems of instruction, and problems of administration. Further analyses with additional datasets, particularly for the EGRA and EGMA would shed some light on whether this is a problem specific to this evaluation, or is a problem generalizable to all evaluations using these assessments.

Finally, there were severe software limitations for the GTheory analyses, We were not able to run several analyses using SPSS (IBM Corp., 2019) even after providing over 24 hours for the software to run. In addition, the readily available software, including SPSS, requires sometimes tedious data formatting, with larger datasets requiring extended time for analysis, or not running at all. Newer software may be available and should be investigated for future use to respond to the above points. The data issues present with the GTheory analyses were not present for the Facets analyses. The full datasets were used across all analyses. However, specialized software is required at a cost to those using it, with some required time to learn the basics of formatting data and running the software.

In addition, for the objective measures, Facets results indicated that model fit was not good. This was particularly true for the fluency type measures, though it improved with the subjective measures significantly. As was evident in the descriptive analyses shown in Table 13 and Table 14, performance on the objective measures was generally poor with many zero scores. This is likely to have had an effect on both sets of analyses, but we are able to see it visually with the Wright maps in the Facets results. Unfortunately, the performance across these measures is relatively consistent with other projects like this one. This indicates that there is at the very least, a lack of alignment in expectation of the respondents wherein the difficulty of these assessments is perhaps not appropriate.

Given that there have been some significant changes as to how the EGRA is developed, including the requirements around inter-rater reliability for enumerators, and more strict protocols around data collection, a similar analysis with data collected after 2016 is recommended to confirm the results here. Future research could also include some assessment of the instruction methods in the classroom to identify if the students are receiving instruction in the national language – French – to see if this is responsible for the differential performance. It would also be prudent to look at the linguistic differences between the Girl's Home Language and the Language of Administration as there may be some home languages that are closer the French, allowing for a smoother transition to learning in French than others.

Implications

Implications for Evaluation in International Development

Due to the introduction of initiatives such as the Millennium Development Goals (<http://www.un.org/millenniumgoals/>), Sustainable Development Goals (<https://www.un.org/sustainabledevelopment/sustainable-development-goals/>), and pay-for-performance incentives (<http://go.worldbank.org/FVDDBVIZD0>), development aid organizations have shifted their focus to the desired long-term intended outcomes of a program (e.g., quality of life increases, employment rates, etc.; Gertler, Martinez, Premand, Rawlings, & Vermeersch, 2011).

Given the impetus of the use of impact evaluation in international development contexts, the rigor of outcome measures becomes even more important. The challenges surrounding the development and implementation of education programs in developing countries, particularly those experiencing conflict, are numerous and wide-reaching. Barriers such as a dearth of

supplies and infrastructure, underqualified teachers, and overall attitudes toward education present unique and serious challenges to development work (GCPEA, 2018; UNESCO, 2018). These barriers coupled with the noted difficulties with tight timelines and a lack of program staff trained in the process of adaptation of measures, make not only implementation but evaluation of programs extremely difficult. Therefore, a focus on the reliability and sensitivity of the measures used to evaluate program outcomes is required.

The results of this study provide implications for several phases of evaluations of educational initiatives in developing countries. First, the results indicate several factors to consider in the process of determining an evaluation design, as well as the importance of a pilot in assisting in refining the design and sampling plan. While in this chapter I have attempted to provide highlights and summary results, the detailed results presented in CHAPTER IV. RESULTS indicate the complexity of the context within which we are working. It may be said that in one case or another, every component of the evaluation matters, and has an effect on the measurement of program outcomes.

While results from this study pointed at several factors that consistently seem to effect measurement, such as the enumerators, and the mis-alignment of language of assessment administration and the respondent's home language, there may be many other factors of importance to include in analyses like these when in other contexts. For example, in this case province and urbanicity were not shown to introduce error, but each project and country comes with its own complexity. Previous studies may provide some guidance as to which factors to include in a study like this, and decisions around what is measured and how will constrain (or not) the available models for analyses. One major recommendation is to pilot, and to include as

many factors as possible in the pilot. This will allow evaluators to adjust the overall measurement plan as needed based on the pilot results.

Second, one of the consistent results throughout this study is the importance of the enumerators, and their training and development. Overwhelmingly, enumerators as a facet throughout both analyses, contribute to measurement error in the objective outcome measures. These measures of academic achievement are most often used to speak to the efficacy of educational interventions; at times, solely. The role of enumerators and enumerator training cannot be understated. These individuals often serve as translators and assessors, while trying to develop a rapport with the students at the same time. We can see that their performance influences scores above all other facets discussed in this evaluation, and they are pivotal to its success. Therefore, the recruitment, training, and development of enumerators in evaluations such as this should be at the top of mind when planning the evaluation. The methods of analyses used in this study may also provide some guidance on analyses that may be completed on enumerator training or pilot data, or may be used at intervals throughout official data collection to identify enumerators with additional training needs.

Finally, the results of this study add to the conversation regarding the importance of the appropriate selection of outcome measures. Given the increased use of quantitative measures, and the importance of said measures as required by impact evaluation and the use of pay-for-performance models, evaluators require a more complete picture of the possible effects on reliability and validity of measures used.

As we have seen that so many outside factors influence the reported scores on outcome measures for these types of evaluations, further discussion should be had around the use of payment-by-results (PbR) models used in development projects. Further research into determining appropriate reliability measures for EGRA and EMGA is required, allowing

substantive conversations about the appropriateness of the PbR targets set, and their use in such complex educational ecosystems.

Pay-for-incentives programs link monetary disbursements to tangible, transparent, and independently verifiable results (World Bank, 2011). These programs are most widely used to make determinations in Primary Health Care expenditures both in domestic and international development contexts (Appleby, Harrison, Hawkins, & Dixon, n.d.; Fritsche, Soeters, & Meessen, 2014; Perrin, 2013). Terminology in the area is inconsistent, DFID, for example, refers to the programs as Payment by Results (PbR) and they fall into one of two types: Results-Based Finance (RBF) programs provide incentives to service provider organizations and individuals, and Results-Based Aid (RBA) programs provide incentives to governments where aid is being provided (Perrin, 2013). The World Bank, however, uses the term program-for-results, and the Center for Global Development refers to it as Cash on Delivery (COD) aid. For the purposes of this paper, PbR will be used to refer to the general category of pay-for-incentives programs and models. Implementation of PbR has encountered consistent difficulties with questions regarding the choice of incentive, cost effectiveness, comparisons with other approaches, impact on equity, and sustainability (Perrin, 2013).

Key features of the World Bank's PbR program include the following stages: 1) the program finances and supports borrowers' programs, first assessing the quality of the program (including results and expenditure frameworks, support systems, and possibility of strengthening measures), 2) money is disbursed upon achievement of monitorable and verifiable program results rather than inputs, 3) there should be a focus on strengthening the institutional capacity needed for programs to achieve desired results by concentrating on transparency, accountability, and participation, and 4) the Bank will provide assurances that financing is used appropriately

and that the environmental and social impacts of the program are adequately addressed (World Bank, 2011).

DFID also provides 12 principles for proper use and implementation of PbR in international development: 1) the recipients of funding must invest first, and be compensated only if pre-agreed measured results are achieved, 2) the quality of the performance measures is the principle factor in determining if PbR should be used, and the strength of the incentive to be used, 3) as much attention paid to gaming of results, should be paid to distortion of incentives, 4) risk sharing is not a goal of PbR, but is relevant as a mechanism to sharpen performance incentives for the implementing agency, 5) when there is full alignment of objectives between the donor and recipient, performance incentives become irrelevant, 6) in cases where the monitoring of actions or inputs are difficult, PbR is advantageous, otherwise, other contracts may be better, 7) PbR is most appropriate where recipients have a large amount of control over the outcomes, 8) the most tangible cost of PbR is the verification of outcome measures used as they must be irrefutable, 9) attention should be paid to the use of fines or rewards such that they do not undermine the personal commitments of implementers, 10) non-payment must be possible, or the effect of the incentive is lost, 11) evidence of success measured only by the incentivized measure should be treated with caution, and 12) it should be noted that there are other forms of aid that can include financial incentives, and these should also be considered (Clist & Verschoor, 2014).

As evidenced above, the most important part of setting up a PbR program is the initial agreement between funders and program implementers regarding the outcome measures and the way in which they will be measured to show change. DFID uses a method call the “results chain” akin to a logic model used in evaluation, with an expectation of several indicators for the output, outcome, and impact levels. It is important that funders and recipients agree upon each piece of

the chain, and especially the indicators for each, ensuring that measures can be validated, and directly linked to indicators (Clist & Verschoor, 2014). Once the outcome measures are agreed upon, the validation of these measures is an integral step in ensuring reliable results.

However, in a 2018 review of ten years of data on PbR projects in international development, Clist was unable to find evidence that the funding models were being implemented appropriately, with due care to the outcome measures chosen. This evidence has not stalled efforts in the use of payment for performance models of development, with the advent of funds such as the Education Outcomes Fund, managed by UNICEF, existing solely to administer programs using these models.

More studies like this one focusing on the measurement aspect of evaluations like VAS-Y Fille! may also help to shift the focus onto the challenges of attaining precise and reliable outcome data in these complex systems, providing a platform for funders and implementers to allocate much needed resources and attention toward measuring what matters well. In addition, this study provides an introduction to two methods of analysis that can be used in these contexts and allow a more informed discussion around the commonly used EGRA and EGMA as outcome measures for these types of projects.

Implications for Measurement

As has been established, in international development projects it is not always feasible to carry out a quality adaptation of a measure before collecting data. Fortunately, there are many ways in which a researcher can determine the quality of the outcome of the efforts, or lack thereof. While GTheory and the Many-Facet Model are not necessarily the most common analyses to complete when attempting to formally determine measurement invariance, the

purpose of this study was to investigate the use of these two analyses, with highly complex data from an international development education evaluation in order to support a more thorough investigation of outcome measures for sources of error.

A thorough review of the results from these two analyses provides both general and specific actionable results, as well as a set of data to support a validity argument about the outcome measures being investigated. GTheory results provide general indications of sources of error for each of the main effect facets as well as the interaction effects. We can know, for example, that there is a substantive amount of error variance associated with enumerators as a result of the GTheory analyses. If we then move to the Facets results, the Wright map provides a visual of the enumerators along the logit scale, indicating if there are some enumerators who appear to have students performing consistently well or consistently poorly – something we would not expect given random selection. This information could then be used to target additional training or support for those enumerators, for example. This simple example suggests that both of the analyses provide very useful results when attempting to establish outcome measure reliability and validity.

Further, GTheory could be very useful at the pilot stage, when establishing a sampling frame and the data collection procedures and design. The visual nature of the Facets results are helpful for at-a-glance discussions, and for digging into the results in at a deeper level, in this case allowing one to see specifically which languages, provinces, enumerators, etc., are performing differentially and make appropriate changes.

In general, these evaluations are complex in nature, with more possible sources of error than those included in the current study. What these results indicate is that though we, as a field wish to standardize and assess in difficult settings, we cannot ignore the fact that context affects not only the results of assessments like the EGMA and EGRA, but their utility.

APPENDICES

Appendix A. Early Grades Reading Assessment (EGRA)

EGRA

3ème a 6eme Année

Évaluation des habiletés fondamentales en lecture

Nom de l'élève [] [] []
NOM POSTNOM Prénom

Classe [] Salle : [A] [B] [C] [D] [E] [F]

Nom de l'enseignant [] [] []
NOM POSTNOM PRÉNOM

LETTRE SD RX ECL CLAS ELEV
Code de l'élève [] [] [] [] [] [] [] []

LETTRE SD RX ECL
Code de l'école [] [] [] [] [] []

Code de l'enquêteur [] []

HH MM
Heure du début de l'entretien [] [] : [] []

🎧 Maintenant nous allons jouer quelques jeux en lecture !

Activité 1. Connaissance des graphèmes (lettres et groupes de lettres)																																																																																																																										
<p>🎧 Voici une page pleine de lettres. Lis-moi ces lettres en me donnant leur nom. Par exemple, cette lettre se lit / O / comme dans le mot "POT". [Indiquer le "O" dans la ligne des exemples]</p> <p>🎧 Essayons un autre maintenant. Lis-moi cette lettre: [Indiquer le "T" dans le rang des exemples]:</p> <ul style="list-style-type: none"> Si l'élève répond <u>correctement</u>, dites: "Très bien, cette lettre se lit / té / ou / t /" Si l'élève ne répond <u>pas correctement</u>, dites: " Cette lettre se lit / té / ou / t /" <p>🎧 D'accord ? On peut continuer ? Lorsque je dis "Commence", montre chaque lettre du doigt quand tu la lis. Prends soin de lire de gauche à droite, ligne par ligne. Mets ton doigt sur la première lettre. Es-tu prêt(e) ? Commence.</p>	<p>📄 S/1</p> <p>⌚ 60 secondes</p> <p>🎧 Si l'élève ne réussit pas à donner une seule bonne réponse parmi <u>les dix premiers graphèmes</u> (le premier rang)</p> <p>⌚ Si l'enfant s'arrête sur un nombre pendant <u>3 secondes</u></p>																																																																																																																									
<p>☒ (/) Incorrecte ou pas de réponse</p> <p>☐ () Après dernier graphème lu</p> <p>Dans le cas où l'élève a donné une réponse incorrecte mais s'est corrigé par la suite (auto-correction), entourez l'item que vous avez déjà barré. Comptez cette réponse comme étant correcte.</p>																																																																																																																										
<p>🎧 Exemple : O T ch</p>																																																																																																																										
<table border="1"> <thead> <tr> <th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th><th></th></tr> </thead> <tbody> <tr><td>E</td><td>R</td><td>d</td><td>C</td><td>L</td><td>p</td><td>a</td><td>T</td><td>b</td><td>i</td><td>*(10)</td></tr> <tr><td>m</td><td>F</td><td>o</td><td>S</td><td>n</td><td>G</td><td>u</td><td>H</td><td>Q</td><td>d</td><td>(20)</td></tr> <tr><td>l</td><td>m</td><td>P</td><td>t</td><td>d</td><td>B</td><td>e</td><td>b</td><td>A</td><td>r</td><td>(30)</td></tr> <tr><td>s</td><td>D</td><td>v</td><td>I</td><td>k</td><td>c</td><td>j</td><td>M</td><td>g</td><td>O</td><td>(40)</td></tr> <tr><td>h</td><td>c</td><td>N</td><td>G</td><td>f</td><td>p</td><td>V</td><td>q</td><td>n</td><td>p</td><td>(50)</td></tr> <tr><td>a</td><td>T</td><td>F</td><td>r</td><td>V</td><td>L</td><td>d</td><td>K</td><td>E</td><td>S</td><td>(60)</td></tr> <tr><td>m</td><td>J</td><td>c</td><td>Q</td><td>o</td><td>A</td><td>G</td><td>i</td><td>v</td><td>F</td><td>(70)</td></tr> <tr><td>k</td><td>b</td><td>P</td><td>q</td><td>e</td><td>q</td><td>s</td><td>n</td><td>O</td><td>l</td><td>(80)</td></tr> <tr><td>D</td><td>w</td><td>N</td><td>I</td><td>h</td><td>t</td><td>j</td><td>R</td><td>g</td><td>U</td><td>(90)</td></tr> <tr><td>J</td><td>f</td><td>Y</td><td>C</td><td>K</td><td>x</td><td>M</td><td>z</td><td>B</td><td>H</td><td>(100)</td></tr> </tbody> </table>	1	2	3	4	5	6	7	8	9	10		E	R	d	C	L	p	a	T	b	i	*(10)	m	F	o	S	n	G	u	H	Q	d	(20)	l	m	P	t	d	B	e	b	A	r	(30)	s	D	v	I	k	c	j	M	g	O	(40)	h	c	N	G	f	p	V	q	n	p	(50)	a	T	F	r	V	L	d	K	E	S	(60)	m	J	c	Q	o	A	G	i	v	F	(70)	k	b	P	q	e	q	s	n	O	l	(80)	D	w	N	I	h	t	j	R	g	U	(90)	J	f	Y	C	K	x	M	z	B	H	(100)	
1	2	3	4	5	6	7	8	9	10																																																																																																																	
E	R	d	C	L	p	a	T	b	i	*(10)																																																																																																																
m	F	o	S	n	G	u	H	Q	d	(20)																																																																																																																
l	m	P	t	d	B	e	b	A	r	(30)																																																																																																																
s	D	v	I	k	c	j	M	g	O	(40)																																																																																																																
h	c	N	G	f	p	V	q	n	p	(50)																																																																																																																
a	T	F	r	V	L	d	K	E	S	(60)																																																																																																																
m	J	c	Q	o	A	G	i	v	F	(70)																																																																																																																
k	b	P	q	e	q	s	n	O	l	(80)																																																																																																																
D	w	N	I	h	t	j	R	g	U	(90)																																																																																																																
J	f	Y	C	K	x	M	z	B	H	(100)																																																																																																																
Nombre exact de secondes restantes indiquées sur le chronomètre :	☒																																																																																																																									
Cochez ici si l'élève ne réussit pas à donner une seule bonne réponse parmi <u>les dix premiers graphèmes</u> (le premier rang) (auto-stop)	☒																																																																																																																									

Crochet	tok
# Incorrect	tok

Activité 2. Lecture des mots inventés

🗨️ Voici des mots que tu n'as peut-être jamais vus. Mais je voudrais que tu essayes de les lire. Par exemple, ce premier mot se lit « bi ». [Indiquer le mot « bi » avec le doigt] Peux-tu lire ce premier mot?

- Si l'élève répond correctement, dites-lui 🗨️ Très bien, ce mot se lit « bi »
- S'il répond pas correctement ou il ne répond pas, dites-lui, 🗨️ Ce mot se lit « bi ». Lis avec moi, « bi ».

🗨️ Et ce mot ? Peux-tu me lire ce mot ? [indiquer le mot « tok » avec le doigt].

- Si l'élève répond correctement, dites-lui 🗨️ Très bien, ce mot se lit « tok »
- S'il répond pas correctement ou il ne répond pas, dites-lui, 🗨️ Ce mot se lit « tok ». Lis avec moi, « tok ».

🗨️ Et ce mot ? Peux-tu me lire ce mot ? [indiquer le mot « sar » avec le doigt].

- Si l'élève répond correctement, dites-lui 🗨️ Très bien, ce mot se lit « sar »
- S'il répond pas correctement ou il ne répond pas, dites-lui, 🗨️ Ce mot se lit « sar ». Lis avec moi, « sar ».

🗨️ D'accord ? On peut continuer ? Lorsque je dis "Commence", montre chaque mot du doigt quand tu le lis. Prends soin de lire de gauche à droite, ligne par ligne. Mets ton doigt sur la premier mot. Es-tu prêt(e) ? Commence.

- 🗨️ (/) Incorrecte ou pas de réponse
- () Après dernier mot lu

🗨️ Exemple : bi tok sar

1	2	3	4	5	
co	ra	di	pa	so	*(5)
al	mé	fo	ce	gi	(10)
dan	tra	por	gra	cho	(15)
dir	nal	qua	pla	isa	(20)
au	in	oi	en	ai	(25)
min	poi	teau	gué	reu	(30)
bir	ja	cla	bige	pro	(35)
numi	moré	bamon	cabou	polan	(40)
chadin	gital	sineau	guira	louber	(45)
oti	lépa	ima	mouli	lorpe	(50)

📖 S/2

⌚ 60 secondes

🗨️ Si l'élève ne réussit pas à donner une seule bonne réponse parmi les cinq premiers mots (le premier rang)

🗨️ Si l'enfant s'arrête sur un nombre pendant 3 secondes

Nombre exact de secondes restantes indiquées sur le chronomètre :	00
Cochez ici si l'élève ne réussit pas à donner une seule bonne réponse parmi <u>les cinq premiers mots</u> (le premier rang) (auto-stop)	00
Crochet	00
# Incorrect	00

Activité 3a. Lecture d'une petite histoire	
<p>🔊 <i>Maintenant je voudrais que tu lises à haute voix l'histoire d'un enfant.</i></p> <p>🔊 <i>Essaye de lire rapidement et correctement ; après, je vais te poser quelques questions.</i></p> <p>🔊 <i>Commence ici lorsque je te le dis. [Montrez du doigt la première ligne (le titre)].</i></p> <p>🔊 <i>Es-tu prêt(e) ? Commence.</i></p>	<p>📖 5/3</p> <p>⌚ 60 secondes</p> <p>🔊 Si l'élève ne réussit pas à donner une seule bonne réponse parmi <u>les huit premiers mots</u> (deux premières lignes)</p> <p>🔊 Si l'enfant s'arrête sur un nombre pendant <u>3 secondes</u></p>
<p>00 (/) Incorrecte ou pas de réponse</p> <p>() Après dernier mot lu</p> <p>Le repas 2</p> <p>Il est midi. Claude rentre à la maison. *10</p> <p>Le repas n'est pas sur la table. 17</p> <p>il va à la cuisine. Maman prépare le riz. 26</p> <p>Les légumes et le poisson sont prêts. Il va se mettre à table. 39</p> <p>Claude est très content. Maman a préparé le plat qu'il aime. 50</p>	
Nombre exact de secondes restantes indiquées sur le chronomètre :	00
Cochez ici si l'élève ne réussit pas à donner une seule bonne réponse parmi <u>les dix premiers mots</u> (les deux premier rang) (auto-stop)	00
Crochet	00
# Incorrect	00

Activité 3b. Compréhension du texte lu			
<p>Lorsque l'élève a terminé de lire (Section 3a), posez la première question ci-après.</p> <p>Posez les questions qui correspondent aux lignes du texte <u>jusqu'à la ligne à laquelle se trouve le crochet (])</u>, c'est-à-dire, jusqu'à l'endroit où l'élève a cessé de lire .</p> <p>☛ Maintenant, tu vas répondre à quelques questions sur l'histoire.</p> <p>☒ Encerclez la case qui correspond à sa réponse par rapport à chaque question. Les réponses correctes peuvent être fournies en langue française ou en langue nationale.</p>	<div> <div>📖 X</div> <div>⌚ X</div> <div>🕒 X</div> </div> <p>➡ Si l'enfant s'arrête sur une question pendant <u>10 secondes</u>, répétez la question et donnez à l'enfant <u>encore 5 secondes</u> pour répondre.</p>		
	Correcte	Incorrecte	Pas de réponse
Quelle heure est-il? [Midi]	1	0	9
Que manque-t-il sur la table? [le repas]	1	0	9
Où va Claude? [à la cuisine]	1	0	9
Qu'est ce que maman prépare? [le riz, le legume, et le poisson]	1	0	9
Pourquoi est-il content? [il mange le plat qu'il aime]	1	0	9

Activité 4. Compréhension à l'Audition

🎧 Maintenant, je vais te lire une histoire deux fois. Après cela, je vais te poser quelques questions sur cette histoire. Tu vas bien écouter, et ensuite tu répondras aux questions le mieux que tu peux. D'accord ? Commençons ! Écoute bien.

- 🎧 Il y a un petit chat nommé Tino. Il est tombé dans la rivière.
- 🎧 Il essaie de sortir de l'eau.
- 🎧 Il s'enfonce dans la boue. Tino est fatigué.
- 🎧 Son père arrive
- 🎧 Il soulève Tino et le pose sur le sol. Tino est sauvé !

📖 X

🕒 X

🎧 X

🕒 Si l'enfant s'arrête sur une question pendant 3 secondes.

📌 Encerclez la case qui correspond à sa réponse par rapport à chaque question. Les réponses correctes peuvent être fournies en langue française ou en langue nationale.

	Correcte	Incorrecte	Pas de réponse
🎧 Comment s'appelle le petit chat? [Tino]	1	0	9
🎧 Qu'est-ce qui lui est arrivé? [il est tombé dans la rivière]	1	0	9
🎧 Ou est-ce que Tino s'enfonce ? [dans la boue]	1	0	9
🎧 Qui est ce qui le cherche ? [son père]	1	0	9
🎧 Ou est ce que son père lui a pose ? [sur le sol]	1	0	9

Activité 5. Entretien sur l'environnement de l'élève			
<p>🔊 On a presque terminé ! Il nous reste juste quelques questions sur les livres chez toi.</p> <p>🔊 Encercler la case qui correspond à sa réponse par rapport à chaque question.</p>	<p>📖 X</p> <p>🕒 X</p> <p>🗨️ X</p> <p>🔄 X</p>		
	Oui	Non	NSP/ Pas de réponse
As-tu un manuel de lecture?	1	0	NSP
Y a-t-il d'autres livres, journaux, ou autres choses à lire chez toi à la maison, autre que tes manuels scolaires?	1	0	NSP
[Si oui à la question précédente] Donne-moi quelques exemples.	Pas besoin d'enregistrer la réponse)		
[Si oui à la question 2:] Ces livres et autres sont en quelle(s) langue(s) ?			
7.1 Français	1	0	NSP
7.2 Lingala	1	0	NSP
7.3 Kikongo	1	0	NSP
7.4 Kiswahili	1	0	NSP
7.5 Tshiluba	1	0	NSP
7.6 Autre (A préciser)	1	0	NSP

Heure de fin du test	HH : MM [] : []
<p>🔊 [Si c'est le dernier outil à tester]: On a fini! Je suis très content. Maintenant, tu peux retourner en classe, vas-y directement.</p>	

Appendix B. Early Grades Mathematics Assessment (EGMA)

CODE fille hors école:	Lettre	SD	RX	MEN
	[]	[]	[]	[]

TEST EGMA /EGRA POUR LES FILLES HORS DE L'ECOLE DANS LE MENAGE

Évaluation des habiletés fondamentales en mathématiques et en lecture

Nom de la fille [] [] []
 NOM POSTNOM Prénom

Heure du début du test EGRA/EGMA [] : []

🔊 Maintenant nous allons jouer quelques jeux mathématiques

Activité 1A: Identification des Nombres

🔊 Voici quelques nombres. Je veux que tu pointes du doigt chaque nombre et que tu me dises de quel nombre s'agit-il. Lire de gauche à droite, ligne par ligne. Je vais me servir de ce chronomètre et je te dirai quand commencer et quand t'arrêter. [pointez du doigt le premier chiffre]
 🔊 Commence par ici. C'est quoi, ce nombre?

🔊 (/) Incorrecte ou pas de réponse
 () Après dernier nombre lu

					Tot. Cum.
2	9	0	22	16	(5)
35	12	28	60	72	(10)
67	89	39	97	55	(15)
234	107	692	528	912	(20)

📄 Feuille A

🕒 60 secondes

🔊 Si le temps sur le chronomètre est épuisé (60 secondes).

🔊 Si l'enfant s'arrête sur un nombre pendant 5 SECONDES

Temps restant: 🕒

Crochet: 🕒

Nombre(s) Incorrect(s): 🕒

Activité 2A: Comparaison des Quantités - EXEMPLE

👁️ Regarde ces nombres. Dis-moi lequel est le plus grand.

- Si l'élève répond correctement, dites-lui : 👁️ C'est correcte, 9 est plus grand..
- Si l'élève ne répond pas correctement, dites-lui 👁️ Le nombre 9 est plus grand. Ce nombre est 4. Ce nombre est 9. 9 est plus grand que 4. Continuons.

📖 Feuille B1

Activité 2A : Comparaison des Quantités - EXERCICE

👁️ Regarde ces nombres. Dis-moi lequel est le plus grand.

[Pointez du doigt les nombres de chaque ligne, une par une, en répétant la consigne]

✖ sur 1 si « Correct »

✖ sur 0 = « Incorrect » ou « pas de réponse »

8	5	<u>8</u>	<u>1</u>	<u>0</u>	94	77	<u>94</u>	<u>1</u>	<u>0</u>
15	26	<u>26</u>	<u>1</u>	<u>0</u>	357	319	<u>357</u>	<u>1</u>	<u>0</u>
27	34	<u>34</u>	<u>1</u>	<u>0</u>	323	443	<u>443</u>	<u>1</u>	<u>0</u>
40	58	<u>58</u>	<u>1</u>	<u>0</u>	708	780	<u>780</u>	<u>1</u>	<u>0</u>
78	75	<u>78</u>	<u>1</u>	<u>0</u>	955	958	<u>958</u>	<u>1</u>	<u>0</u>

📖 B2

🕒 ✖

👁️ Si l'enfant fait 4 erreurs successives

🕒 Si l'enfant ne répond pas après 5 SECONDES.

Total Correct: ✖

Activité 3A: Chiffre Manquant - EXEMPLE

👁️ Voici quelques nombres. Un, deux, trois, quatre. Quel est le nombre qui va ici ? [pointez du doigt à l'espace]

Si l'élève répond correctement, dites-lui : 👁️ C'est juste, cinq. Continuons.

Si l'élève ne répond pas correctement, dites-lui 👁️ Ici, le nombre est cinq.

Compte avec moi. [Pointez chaque nombre du doigt] Un, deux, trois, quatre. Ici, le nombre est cinq. Continuons.

📖 C1

👁️ Voici quelques nombres. Dix-sept, [pointez du doigt à l'espace sans rien dire], dix-neuf, vingt, vingt-et-un. Quel nombre va ici? [pointez du doigt à l'espace]

Si l'élève répond correctement, dites-lui : 👁️ C'est juste, dix-huit.

Si l'élève ne répond pas correctement, dites-lui 👁️ Ici, le nombre est dix-huit.

Compte avec moi. [Pointez chaque nombre du doigt] Dix-sept, dix-huit, dix-neuf, vingt, vingt-et-un. Continuons.

CODE fille hors école: Lettre SD RX MEN

[] [] [] [] [] []

Activité 3A: Chiffre Manquant - EXERCICE

👁️: Voici quelques nombres. Quel nombre va ici ? [Pointez l'espace du doigt] [Répétez pour chaque item si nécessaire]

- 👁️ ✖ sur 1 si « Correct »
✖ sur 0 = « Incorrect » ou « pas de réponse »

6	7	8	<u>9</u>	<input type="text" value="1"/>	<input type="text" value="0"/>	347	348	<u>349</u>	350	<input type="text" value="1"/>	<input type="text" value="0"/>
11	12	<u>13</u>	14	<input type="text" value="1"/>	<input type="text" value="0"/>	39	37	<u>35</u>	33	<input type="text" value="1"/>	<input type="text" value="0"/>
<u>49</u>	59	69	79	<input type="text" value="1"/>	<input type="text" value="0"/>	50	55	60	<u>65</u>	<input type="text" value="1"/>	<input type="text" value="0"/>
100	200	300	<u>400</u>	<input type="text" value="1"/>	<input type="text" value="0"/>	<u>440</u>	430	420	410	<input type="text" value="1"/>	<input type="text" value="0"/>
12	14	16	<u>18</u>	<input type="text" value="1"/>	<input type="text" value="0"/>	3	<u>8</u>	13	18	<input type="text" value="1"/>	<input type="text" value="0"/>

📖 C2

🕒 ✖

👁️ Si l'enfant fait 4 erreurs successives

👁️ Si l'enfant ne répond pas après 5 SECONDES.

Total correct:

👁️ ✖

Activité 4A: Addition

🕒 Voici quelques additions. Je vais maintenant utiliser ce chronomètre. Essaie de ton mieux. Si tu ne connais pas une réponse, passe au prochain problème. Es-tu prêt(e) ? ... Commence par ici. [Pointez le premier item]

D1

⌚ 60 secondes

☒ (/) Incorrecte ou pas de réponse
☐ () Après dernier item répondu

👉 Si le temps sur le chronomètre est épuisé (60 secondes).

➡ Si l'enfant s'arrête sur un nombre pendant 5 SECONDES

$1 + 1 = (2)$	$1 + 2 = (3)$	$3 + 2 = (5)$	(3)
$2 + 4 = (6)$	$3 + 3 = (6)$	$8 + 1 = (9)$	(6)
$3 + 6 = (9)$	$7 + 3 = (10)$	$5 + 5 = (10)$	(9)
$2 + 8 = (10)$	$7 + 4 = (11)$	$8 + 6 = (14)$	(12)
$5 + 7 = (12)$	$6 + 6 = (12)$	$7 + 8 = (15)$	(15)
$9 + 7 = (16)$	$8 + 8 = (16)$	$4 + 7 = (11)$	(18)
$10 + 3 = (13)$	$10 + 7 = (17)$		(20)

Total Correct:

Temps restants

Crochet

Activité 5A: Soustraction

Voici quelques soustractions. Je vais encore utiliser ce chronomètre. Essaie de ton mieux. Si tu ne connais pas une réponse, passe au prochain problème. Es-tu prêt(e) ? ... Commence par ici. [Pointez le premier item]

D2

⌚ 60 secondes

☐ (/) Incorrecte ou pas de réponse
☐ () Après dernier item répondu

👉 Si le temps sur le chronomètre est épuisé (60 secondes).

➡ Si l'enfant s'arrête sur un nombre pendant 5 SECONDES

$2 - 1 = (1)$	$3 - 2 = (1)$	$5 - 3 = (2)$	(3)
$6 - 2 = (4)$	$6 - 3 = (3)$	$9 - 1 = (8)$	(6)
$9 - 3 = (6)$	$10 - 3 = (7)$	$10 - 5 = (5)$	(9)
$10 - 2 = (8)$	$11 - 7 = (4)$	$14 - 8 = (6)$	(12)
$12 - 5 = (7)$	$12 - 6 = (6)$	$15 - 7 = (8)$	(15)
$16 - 7 = (9)$	$16 - 8 = (8)$	$11 - 7 = (4)$	(18)
$13 - 10 = (3)$	$17 - 7 = (10)$		(20)

Total Correct:



Temps restants

Crochet



Appendix C. In School Girls' Survey

Table 60. General School Perception Items from Girls' Survey

Item	Never	Sometimes	Almost Always	Always
<i>Do you feel comfortable when you are at school?</i>	1	2	3	4
<i>Do you feel uncomfortable when you are at school?*</i>	1	2	3	4
<i>You and your classmates help each other learn.</i>	1	2	3	4
<i>You and your classmates play together.</i>	1	2	3	4
<i>You and your classmates share books without fighting.</i>	1	2	3	4
<i>At school, if students see another student being picked on, they try to stop it.</i>	1	2	3	4
<i>The subjects you are studying at school are interesting.</i>	1	2	3	4
<i>The school is a welcoming place for all types of students.</i>	1	2	3	4
<i>Boys and girls have equal opportunity to succeed at this school.</i>	1	2	3	4
<i>At school, if students see another student being picked on, they try to stop it.</i>	1	2	3	4

*Item was reverse-coded; not included at Annual Evaluation.

Table 61. Teacher Perception Items from Girls' Survey

Item	Never	Sometimes	Almost Always	Always
<i>Your teachers treat you with respect.</i>	1	2	3	4
<i>Teachers at your school are interested in what students like you have to say.</i>	1	2	3	4
<i>Your teacher gives you help whenever you need it.</i>	1	2	3	4
<i>You can talk to your teacher if you have a problem.</i>	1	2	3	4
<i>Your teacher gets angry at you. *</i>	1	2	3	4
<i>Your teacher really cares about you.</i>	1	2	3	4
<i>Your teacher always tries to be fair.</i>	1	2	3	4
<i>Your teacher notices good things you do.</i>	1	2	3	4
<i>Every student is encouraged to participate in class discussions.</i>	1	2	3	4
<i>Teachers at this school expect students like you to succeed in life.</i>	1	2	3	4
<i>When students master the lesson, they are given more difficult work.</i>	1	2	3	4
<i>Your teacher knows your name.</i>	1	2	3	4
<i>Your teacher ignores you. *</i>	1	2	3	4
<i>Your teacher criticizes you without reason. *</i>	1	2	3	4
<i>Your teacher calls on you when you raise your hand.</i>	1	2	3	4
<i>Your teacher helps you learn how to read.</i>	1	2	3	4
<i>Your teacher helps you learn math.</i>	1	2	3	4

Item	Never	Sometimes	Almost Always	Always
<i>Your teacher teaches less interesting lessons.*</i>	1	2	3	4
<i>Your teacher encourages you when you have a problem.</i>	1	2	3	4
<i>Your teacher makes you feel dumb and not want to continue.*</i>	1	2	3	4
<i>Your teacher does not care if you learn.*</i>	1	2	3	4
<i>Your teacher helps students get along.</i>	1	2	3	4

*Items were reverse-coded.

Table 62. Perception of School Violence from Girls' Survey

Item	Never	Sometimes	Almost Always	Always
<i>Your teacher hits you.*</i>	1	2	3	4
<i>You are afraid of your teacher.*</i>	1	2	3	4
<i>Other kids from school push or hit you.*</i>	1	2	3	4
<i>Other kids from school call you mean names.*</i>	1	2	3	4
<i>Other kids from school tell you they want to hit you.*</i>	1	2	3	4
<i>Other kids from school leave you out on purpose.*</i>	1	2	3	4
<i>Another kid from school did something to make the other kids not like you.*</i>	1	2	3	4

*Items were reverse-coded.

REFERENCES

- Ahmed, A. U., Rabbani, M., Sulaiman, M., & Das, N. C. (2009). *The Impact of Asset Transfer on Livelihoods of the Ultra Poor in Bangladesh* (No. 39). Retrieved from www.brac.net/research
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.
- Andrich, D. (1978). Relationships Between the Thurstone and Rasch Approaches to Item Scaling. *Applied Psychological Measurement*, 2(3), 451–462.
<https://doi.org/10.1177/014662167800200319>
- Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2002). Vouchers for Private Schooling in Columbia: Evidence from a Randomized Natural Experiment. *American Economic Review*, 92(5), 1535–1558. <https://doi.org/10.1257/000282802762024629>
- Appleby, J., Harrison, T., Hawkins, L., & Dixon, A. (n.d.). *Payment by Results: How can payment systems help to deliver better care?* Retrieved from www.kingsfund.org.uk/publications
- Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. L. (2007). Remediating Education: Evidence from Two Randomized Experiments in India. *Quarterly Journal of Economics*, 122(3), 1235–1264. <https://doi.org/10.1162/qjec.122.3.1235>
- Barrera-Orsorio, F., Bertrand, M., Linden, L. L., & Perez-Calle, F. (2011). Improving the design of conditional transfer programs: Evidence from a randomized education experiment in

- Colombia. *American Economic Journal: Applied Economics*, 3(2), 167–195.
<https://doi.org/10.1257/app.3.2.167>
- Barrera-Osorio, F., & Linden, L. L. (2009). *The Use and Misuse of Computers in Education: Evidence from a Randomized Experiment in Colombia* (No. 4836). Retrieved from <http://econ.worldbank.org>.
- Bategeka, L., & Okurut, N. (2005). *Universal Primary Education: Uganda* (No. 10).
<https://doi.org/10.4135/9781483345727.n829>
- Bayerl, P. S., & Paul, K. I. (2007). Squibs and Discussions Identifying Sources of Disagreement: Generalizability Theory in Manual Annotation Studies. *Computational Linguistics*, 33(1), 3–8. Retrieved from <http://journals.sagepub.com/doi/10.1177/0265532216686999>
- Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of Cross Cultural adaptation of Self Report measures. *Spine*, 25(24), 3186–3191.
- Behrman, J R, Parker, S. W., & Todd, P. E. (2009). Schooling Impacts of Conditional Cash Transfers on Young Children: Evidence from Mexico. *Econ.Dev.Cult.Change.*, 57(3), 439–477. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/20209076>
<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2832207/pdf/nihms-148140.pdf>
- Behrman, Jere R., Sengupta, P., & Todd, P. (2005). Progressing through PROGRESA: An Impact Assessment of a School Subsidy Experiment in Rural Mexico. *Economic Development and Cultural Change*, 54(1), 237–275. <https://doi.org/10.1086/431263>
- Berlinski, S., Galiani, S., & Manacorda, M. (2008). Giving children a better start: Preschool

attendance and school-age profiles. *Journal of Public Economics*, 92(5–6), 1416–1440.

<https://doi.org/10.1016/j.jpubeco.2007.10.007>

Berthet, J. (2013). *How to improve enrollment, attendance and learning for primary girls in the DRC*. New York, NY.

Bettinger, E., Kremer, M., & Saavedra, J. E. (2009). Education Vouchers in Columbia. In F. Barrera-Orsorio, H. A. Patrinos, & Q. Wodon (Eds.), *Emerging Evidence on Vouchers and Faith-Based Providers in Education: Case Studies from Africa, Latin America, and Asia* (pp. 71–78). Washington, D.C.: The World Bank.

Brennan, R. L. (1992). *Elements of generalizability theory*. American College Testing Program.

Brennan, R. L. (2001). *Generalizability theory*.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24(1), 1–21. <https://doi.org/10.1080/08957347.2011.532417>

Bretherton, D., Weston, J., & Zbar, V. (2003). Peace education in a post-conflict environment: The case of sierra leone. *Prospects*, 33(2), 219–230.
<https://doi.org/10.1023/A:1023651031633>

Burde, D., & Linden, L. L. (2013). Bringing education to afghan girls: A randomized controlled trial of village-based schools. *American Economic Journal: Applied Economics*.
<https://doi.org/10.1257/app.5.3.27>

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, 10, 107–132. <https://doi.org/10.1080/15305051003637306>

- Carlson, E. D. (2000). A case study in translation methodology using the health-promotion lifestyle profile II. *Public Health Nursing, 17*(1), 61–70. <https://doi.org/10.1046/j.1525-1446.2000.00061.x>
- Chaaban, J., & Cunningham, W. (2011). *Measuring the Economic Gain of Investing in Girls; the girl effect dividend* (No. 5753).
- Clinton Foundation, Bill & Melinda Gates Foundation, & WORLD Policy Analysis Center. (2015). *The Full Participation Report*. New York, NY.
- Clist, P. (2019). Payment by results in international development: Evidence from the first decade. *Development Policy Review, 37*(6), 719–734. <https://doi.org/10.1111/dpr.12405>
- Clist, P., & Verschoor, A. (2014). *The Conceptual Basis of Payment by Results* *.
- Conn, K. M. (2014). *Identifying Effective Education Interventions in Sub-Saharan Africa: A meta-analysis of rigorous impact evaluations* (Columbia University). <https://doi.org/10.1017/CBO9781107415324.004>
- Conn, K. M. (2017). Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Impact Evaluations. *Review of Educational Research*. <https://doi.org/10.3102/0034654317712025>
- Contreras, D., Sepúlveda, P., & Bustos, S. (2010). When schools are the ones that choose: The effects of screening in Chile. *Social Science Quarterly, 91*(5), 1349–1368. <https://doi.org/10.1111/j.1540-6237.2010.00735.x>
- Cor, M. K., & Peeters, M. J. (2015). Using generalizability theory for reliable learning assessments in pharmacy education. *Currents in Pharmacy Teaching and Learning, 7*(3),

332–341. <https://doi.org/10.1016/j.cptl.2014.12.003>

Crocker, L. M., & Algina, J. (2008). *Introduction to classical and modern test theory*. Retrieved from <http://www.worldcat.org/title/introduction-to-classical-and-modern-test-theory/oclc/268675245>

Cronbach, L. J. (1963). On estimates of test reliability. *British Journal of Statistical Psychology*, 16(2), 137–163. <https://doi.org/10.1111/j.2044-8317.1963.tb00206.x>

Cronbach, Lee J. (1951). *Coefficient alpha and the internal structure of tests** lf~ j. *cronbach*. 16(3), 297–298.

Cronbach, Lee J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of Generalizability: A Liberalization of Reliability Theory. *British Journal of Statistical Psychology*, 16(2), 137–163. <https://doi.org/10.1111/j.2044-8317.1963.tb00206.x>

De Herdt, T., & Titeca, K. (2016). Governance with Empty Pockets: The Education Sector in the Democratic Republic of Congo. In *Development and Change*. <https://doi.org/10.1111/dech.12235>

Deininger, K. (2003). Does cost of schooling affect enrollment by the poor? Universal primary education in Uganda. *Economics of Education Review*, 22(3), 291–305. [https://doi.org/10.1016/S0272-7757\(02\)00053-5](https://doi.org/10.1016/S0272-7757(02)00053-5)

Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. *American Economic Review*. <https://doi.org/10.1257/aer.102.4.1241>

Durvasula, S., Netemeyer, R. G., Andrews, J. C., & Lysonski, S. (2006). Examining the cross-national applicability of multi-item, multi-dimensional measures using generalizability

- theory. *Journal of International Business Studies*, 37(4), 469–483.
<https://doi.org/10.1057/palgrave.jibs.8400210>
- Dzhambov, A. M., & Dimitrova, D. D. (2014). Validating a short Bulgarian version of a psychometric instrument for multidimensional noise sensitivity assessment. *Folia Medica*, 56(2), 116–125. <https://doi.org/10.2478/folmed-2014-0017>
- Engelhard, G. J., & Wind, S. A. (2018). *Invariant Measurement with Raters and Rating Scales: Rasch Models for Rater-Mediated Assessments*. New York, NY: Routledge.
- Evans, D. K., & Popova, A. (2016). What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. *World Bank Research Observer*. <https://doi.org/10.1093/wbro/lkw004>
- Evans, D., Kremer, M., & Ngatia, M. (2008). *The Impact of Distributing School Uniforms on Children's Education in Kenya*. Retrieved from <http://documents.worldbank.org/curated/en/311551491549236505/pdf/114092-WP-evans-kenya-uniforms-PUBLIC.pdf>
- Fernando, D., De Silva, D., Carter, R., Mendis, K. N., & Wickremasinghe, R. (2006). A Randomized, Double-Blind, Placebo-Controlled, Clinical Trial of the Impact of Malaria PRevention on the Educational Attainment of School Children. *The American Journal of Tropical Medicine and Hygiene*, 74(3), 386–393. <https://doi.org/10.4269/ajtmh.2006.74.386>
- Force, L. M. T. (2013). *Toward universal learning: Recommendations from the learning metrics task force*. Retrieved from http://www.ghbook.ir/index.php?name=فرهنگ و رسانه های نوین&option=com_dbook&task=readonline&book_id=13650&page=73&chkhask=ED9C9

491B4&Itemid=218&lang=fa&tmpl=component

Fritsche, G. B., Soeters, R., & Meessen, B. (2014). *Performance-Based Financing Toolkit*.

<https://doi.org/10.1596/978-1-4648-0128-0>

Gadbury-Amyot, C. C., McCracken, M. S., Woldt, J. L., & Brennan, R. L. (2014). Validity and reliability of portfolio assessment of student competence in two dental school populations: A four-Year study. *Journal of Dental Education*, 78(5), 657–667.

Gallego, F. A. (2006). *Voucher-School competition , incentives, and outcomes: Evidence from Chile*.

GCPEA. (2014). *Education Under Attack 2014*. Retrieved from

http://protectingeducation.org/sites/default/files/documents/eua_2014_full_0.pdf

GCPEA. (2018). *Education Under Attack 2018*. Retrieved from

http://www.protectingeducation.org/sites/default/files/documents/eua_2018_full.pdf
<http://www.protectingeducation.org/country-profile/syria>

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2011). *Impact Evaluation in Practice*. Retrieved from <http://www.worldbank.org/pdt>

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, 30(4), 395–418.

<https://doi.org/10.1007/BF02289531>

Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics*, 2(3), 205–227. <https://doi.org/10.1257/app2.3.205>

- Goldin, I., Pitt, A., Nabarro, B., & Boyle, K. (2018). *Migration and The Economy: Economic Realities, Social Impacts & Political Choices*.
- Grogan, L. (2009). Universal primary education and school entry in Uganda. *Journal of African Economies*, 18(2), 183–211. <https://doi.org/10.1093/jae/ejn015>
- Groleau, G. (2017). *Improved Management and Accountability: Conditions for Better Access and Quality of Primary Education in the Democratic Republic of Congo?* New York, NY.
- Hamazaki, K., Tunru, I. S., Azwir, M. F., Bs, P., Bsc, A., Sawazaki, S., & Hamazaki, T. (2008). The effects of docosahexaenoic acid-rich fish oil on behavior, school attendance rate and malaria infection in school children-a double-blind, randomized, placebo-controlled trial in Lampung, Indonesia. *Asia Pacific Journal of Clinical Nutrition*, 17(2), 258–263.
- Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2004). *Adapting educational and psychological tests for cross-cultural assessment* (R. K. Hambleton, P. F. Merenda, & C. D. Spielberger, Eds.). London: Psychology Press.
- Harkness, J. (1999). In pursuit of quality: Issues for cross-national survey research. *International Journal of Social Research Methodology*, 2(2), 125–140. <https://doi.org/10.1080/136455799295096>
- Harkness, J. A., Villar, A., & Edwards, B. (2010). Translation, adaptation, and design. In Janet A. Harkness, M. Braun, B. Edwards, T. P. Johnson, La. Lyberg, P. P. Mohler, ... T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 115–140). Hoboken, New Jersey: John Wiley & Sons, Inc.
- Herz, B. K., & Sperling, G. B. (2004). *What Works in Girls' Education: Evidence and Policies*

from the Developing World. Washington, D.C.

I.L.O. (2017). *Global Estimates of Child Labour*. Retrieved from
[https://www.ilo.org/wcmsp5/groups/public/---dgreports/---
dcomm/documents/publication/wcms_575499.pdf](https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/publication/wcms_575499.pdf)

I.O.M. (2018). *Global Report Indicators 2018*. Retrieved from www.iom.int

IBM Corp. (2019). *IBM SPSS Statistics for Windows*. Armonk, NY: IBM Corp.

International Test Commission. (2017). *International Test Commission*. (Second Edi).
<https://doi.org/10.1111/j.1464-0597.1975.tb00322.x>

Iramaneerat, C., Yudkowsky, R., Myford, C. M., & Downing, S. M. (2008). Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. *Advances in Health Sciences Education*, 13(4), 479–493. <https://doi.org/10.1007/s10459-007-9060-8>

Jones, N., Jones, H., Steer, L., & Datta, A. (2009). *Improving impact evaluation production and use* (No. 300). London.

Kagitcibasi, C., Sunar, D., & Bekman, S. (2001). Long-term effects of early intervention: Turkish low-income mothers and children. *Journal of Applied Developmental Psychology*, 22(4), 333–361. [https://doi.org/10.1016/S0193-3973\(01\)00071-5](https://doi.org/10.1016/S0193-3973(01)00071-5)

Kang, M., Bjornson, K., Barreira, T. V, Ragan, B. G., & Song, K. (2014). The Minimum Number of Days Required to Establish Physical Activity Estimates in Children Aged 2-15 Years. *Physiological Measurement*, 35(1), 2229–2237. <https://doi.org/10.1177/0272431614561261>.

Kazianga, H., De Walque, D., & Alderman, H. (2009). *Educational and Health Impacts of Two*

- School Feeding Schemes Evidence from a Randomized Trial in Rural Burkina Faso* (No. 4976). Retrieved from <http://econ.worldbank.org>.
- Kim, J., Alderman, H., & Orazem, P. (1998). *Can Cultural Barriers Be Overcome in Girls' Schooling?: The Community Support Program in Rural Balochistan* (No. 10).
- Kline, R. (2015). *Principles and Practice of Structural Equation Modeling* (Fourth Edi). New York, NY: Guilford.
- Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to Learn. *The Review of Economics and Statistics*, 91(3), 437–456.
- Krishnaratne, S., White, H., & Carpenter, E. (2013). *Quality Education for All Children? What Works in Education in Developing Countries* (No. 20).
- Lai, F., Zhang, L., Qu, Q., Hu, X., Shi, Y., Boswell, M., & Rozelle, S. (2012). *Does Computer-Assisted Learning Improve Learning Outcomes? Evidence from a Randomized Experiment in Public Schools in Rural Minority Areas in Qinghai, China* (No. 237).
- Levine, R., Lloyd, C., Greene, M., & Grown, C. (2008). *Girls Count: A Global Investemtn & Action Agenda*. Washington, D.C.: Center for Global Development.
- Levy, D., Sloan, M., Linden, L., & Kazianga, H. (2009). *Impact Evaluation of Burkina Faso's BRIGHT Program*. Retrieved from <https://files.eric.ed.gov/fulltext/ED507466.pdf>
- Linacre, J. M. (1989). *Many-Facet Rasch Measurement*. Chicago, Illinois: MESA Press.
- Lloyd, C. B., & Young, J. (2009). *New Lessons: The power of educating adolescent girls*. Retrieved from <http://www.gsdr.org/go/display&type=Document&id=4298>

- Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet rasch measurement in the development of performance assessments of The ESL speaking skills of immigrants. *Language Testing*, 15(2), 158–180. <https://doi.org/10.1177/026553229801500202>
- Macours, K., Schady, N., & Vakis, R. (2012). Cash transfers, behavioral changes, and cognitive development in early childhood: Evidence from a randomized experiment. *American Economic Journal: Applied Economics*, 4(2), 247–273. <https://doi.org/10.1257/app.4.2.247>
- Malhotra, M. K., & Sharma, S. (2008). Generalizability Theory : An Examination. *Decision Sciences Journal of Innovative Education*, 39(4), 643–669.
- Martinez, S., Naudeau, S., & Pereira, V. (2013). *The Promise of Preschool in Africa : A Randomized Impact Evaluation of Early Childhood Development in Rural Mozambique*. Washington, D.C.
- Maydeu-Olivares, A., Morera, O., & D’Zurilla, T. J. (1998). Using graphical methods in assessing measurement invariance in inventory data. *Multivariate Behavioral Research*, 34(3), 397–420. https://doi.org/10.1207/S15327906MBR3403_5
- Mcgorry, S. Y. (2000). Measurement in a cross-cultural environment: Survey translation issues. *Qualitative Market Research: An International Journal*, 3(2), 74–81. <https://doi.org/10.1108/13522750010322070>
- Meng, X., & Ryan, J. (2010). Does a food for education program affect school outcomes? The Bangladesh case. *Journal of Population Economics*, 23(2), 415–447. <https://doi.org/10.1007/s00148-009-0240-0>
- Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: applications

- in cross-cultural research. *International Journal of Psychological Research*, 3(1), 111–121.
<https://doi.org/10.21500/20112084.857>
- Moul, T. R. (2017). *Promotion and Implementation of Global Citizenship Education in Crisis Situations*. Geneva, Switzerland.
- Nguyen, T. (2008). *Information, Role Models and Perceived Returns to Education: Experimental Evidence from Madagascar*. Boston.
- Oh, Y., Osgood, D. W., & Smith, E. P. (2015). Measuring Afterschool Program Quality Using Setting-Level Observational Approaches. *Journal of Early Adolescence*, 35(5–6), 681–713.
<https://doi.org/10.1177/0272431614561261>
- Ölmez, İ. B., & Ölmez, S. B. (2019). Validation of the Math Anxiety Scale with the Rasch Measurement Model. *Mathematics Education Research Journal*, 31(1), 89–106.
<https://doi.org/10.1007/s13394-018-0244-8>
- ParahitaAnandi, R., & Zailaini, M. A. (2019). Using Rasch model to assess self-assessment speaking skill rubric for non-native Arabic language speakers. *Pertanika Journal of Social Sciences and Humanities*, 27(3), 1469–1480.
- Pereznieto, P., Magee, A., & Fyles, N. (2017). *Evidence Review: Mitigating threats to girls' education in conflict-affected contexts: Current practice*. Retrieved from
http://www.ungei.org/resources/files/Girls_in_Conflict_Review-Final-Web.pdf
- Perrin, B. (2013). Evaluation of Payment by Results (PBR): Current Approaches, Future Needs. In *Department for International Development Working Paper* (Vol. 39).
- Pigozzi, M. J. (2006). What is the “quality of education”? (A UNESCO perspective). In K. N.

- Ross & I. J. Genevois (Eds.), *Cross-national studies of the quality of education: planning their design and managing their impact* (pp. 39–50). UNESCO, International Institute for Educational Planning.
- Randall, J., & Engelhard, G. (2010). Using confirmatory factor analysis and the Rasch model to assess measurement invariance in a high stakes reading assessment. *Applied Measurement in Education*, 23(3), 286–306. <https://doi.org/10.1080/08957347.2010.486289>
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Rihani, M. A., Kays, L., & Psaki, S. (2006). *Keeping the Promise: Five Benefits of girls' secondary education*. Washington, D.C.: Academy for Educational Development.
- RTI International. (2014). *Early Grade Mathematics Assessment (EGMA) Toolkit*. Retrieved from [http://ierc-publicfiles.s3.amazonaws.com/public/resources/EGMA Toolkit_March2014.pdf](http://ierc-publicfiles.s3.amazonaws.com/public/resources/EGMA_Toolkit_March2014.pdf)
- RTI International. (2016). *Early Grade Reading Assessment (EGRA): Toolkit (Second Edition)*.
- Sapelli, C., & Vial, B. (2005). Private vs Public Voucher Schools in Chile: New Evidence on Efficiency and Peer Effects. In *Documento de Trabajo* (No. 289). Santiago.
- Schady, N., & Araujo, M. C. (2006). *Cash transfers, conditions, school enrollment, and child work: Evidence from a randomized experiment in Ecuador* (No. 3930). Washington, D.C.
- Schaffer, B. S., & Riordan, C. M. (2003). A Review of cross-cultural methodologies for organizational research: A best-practices approach. *Organizational Research Methods*, 6(2), 169–215. <https://doi.org/10.1177/1094428103251542>

- Sharma, S., & Weathers, D. (2003). Assessing generalizability of scales used in cross-national research. *International Journal of Research in Marketing*, 20, 287–295.
[https://doi.org/10.1016/S0167-8116\(03\)00038-7](https://doi.org/10.1016/S0167-8116(03)00038-7)
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.
- Skoufias, E., & Parker, S. W. (2001). Conditional Cash Transfers and Their Impact on Child Work and Schooling: Evidence from the PROGRESA Program in Mexico. *Economía*.
<https://doi.org/10.1353/eco.2001.0016>
- Skoufias, E., & Shapiro, J. (2006). Evaluating The Impact Of Mexico's Quality Schools Program : The Pitfalls Of Using Nonexperimental Data: Policy Research Working Papers. *Policy Research Working Papers*, 48. <https://doi.org/10.1596/1813-9450-4036>
- Smith, E. V., & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement*, 64(4), 617–639.
<https://doi.org/10.1177/0013164404263876>
- Solano-Flores, G., & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice*, 25(1), 13–22.
<https://doi.org/10.1111/j.1745-3992.2006.00048.x>
- Sperling, G. B., Winthrop, R., & Kwauk, C. (2016). *What works in girls' education: Evidence for the world's best investment*. Washington, D.C.: Brookings Institution Press.
- Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing*

Writing, 9(3), 239–261. <https://doi.org/10.1016/j.asw.2004.11.001>

Summers, L. H. (1992). *Investing in all the people* (No. 905).

Tabatabaee-Yazdi, M., Motallebzadeh, K., Ashraf, H., & Baghaei, P. (2018). Development and validation of a teacher success questionnaire using the rasch model. *International Journal of Instruction*, 11(2), 129–144. <https://doi.org/10.12973/iji.2018.11210a>

Tieffenberg, J. A., Wood, E. I., Alonso, A., Tossutti, M. S., & Vicente, M. F. (2000). A randomized field trial of acindes: a child-centered training model for children with chronic illnesses (asthma and epilepsy). *Journal of Urban Health*, 77(2), 280–297. <https://doi.org/10.1007/BF02390539>

Torgerson, C. J., Torgerson, D. J., & Taylor, C. A. (2015). Randomized Controlled Trials and Nonrandomized Designs. In K. E. Newcomer, H. P. Hatry, & J. S. Wholey (Eds.), *Handbook of Practical Program Evaluation* (4th Editio, pp. 158–176). Hoboken, New Jersey: John Wiley & Sons, Inc.

Traub, R. E., & Rowley, G. L. (1991). An NCME Instructional Module on. Understanding Reliability. *Educational Measurement: Issues and Practice*, 10(1), 37–45. <https://doi.org/10.1111/j.1745-3992.1991.tb00183.x>

U.N. (2015a). The Millennium Development Goals Report 2015. In *United Nations*. <https://doi.org/978-92-1-101320-7>

U.N. (2015b). *Transforming our World: The 2030 Agenda for Sustainable Development*. <https://doi.org/10.1201/b20466-7>

U.N. (2017). The Sustainable Development Goals Report 2017. In *United Nations*.

<https://doi.org/10.18356/3405d09f-en>

U.N. (2018). *The Sustainable Development Goals Report 2018*. New York, NY.

U.N. (2019a). *Financing for sustainable development report 2019*. New York, NY.

U.N. (2019b). *The Sustainable Development Goals Report 2019*. Retrieved from

<https://unstats.un.org/sdgs/report/2019/>

U.S. Department of State. (2019). *Trafficking in Persons Report June 2019*. Washington, D.C.

UK Aid. (2013). *VAS-Y Fille Monitoring & Evaluation Framework*.

UK Aid. (2015). *Girls ' Education Challenge Project profiles*.

UK Aid. (2018). *Thematic Review: Teaching , Learning and Assessment*.

UNESCO. (2011). *The Hidden Crisis: Armed Conflict and Education: EFA Global Monitoring Report 2011*. Paris, France.

UNESCO. (2014). *Education for All Global Monitoring Report 2013/4: Teaching and Learning*. Paris, France.

UNESCO. (2015). *Education for All 2000-2015: Achievements and Challenges*. Paris, France.

UNESCO. (2018). *Gloabl Education Monitoring Report 2017/18*.

<https://doi.org/10.4324/9781315021348>

UNHCR. (2019). *Stepping Up: Refugee Education in Crisis*. Geneva, Switzerland.

UNICEF. (2016). *UNICEF Annual Report 2016: Democratic Republic of Congo*.

<https://doi.org/10.2499/9780896297852>

- USAID, & ECCN. (2016). *USAID ECCN Alternative Education in DRC Final Research Report DRAFT 3*. Washington, D.C.
- van de Glind, H., & Kou, A. (2013). *Migrant Children in Child Labour: A Vulnerable Group in Need of Attention*. Geneva, Switzerland: International Organization for Migration.
- van de Vijver, F. J. R., & Matsumoto, D. (2011). *Introduction to the Methodological Issues Associated with Cross-Cultural Research* (D. Matsumoto & F. J. R. van de Vijver, Eds.). New York, NY: Cambridge University Press.
- Van Zile-Tamsen, C. (2017). Using Rasch Analysis to Inform Rating Scale Development. *Research in Higher Education*, 58(8), 922–933. <https://doi.org/10.1007/s11162-017-9448-0>
- Vermeersch, C., & Kremer, M. (2005). *School Meals, Educational Achievement, and School Competition: Evidence from a Randomized Evaluation*. <https://doi.org/10.2139/ssrn.667881>
- White, H. (2005). Challenges in Evaluating Development Effectiveness. In *IDS Working Paper* (No. 242). Brighton, Sussex.
- White, Howard, & Raitzer, D. A. (2017). *Impact evaluations of development interventions: A practical guide*. <https://doi.org/10.22617/TCS179188-2>
- Winthrop, R., Anderson, K., & Cruzalegui, I. (2015). A review of policy debates around learning in the post-2015 education and development agenda. *International Journal of Educational Development*, 40, 297–307. <https://doi.org/10.1016/j.ijedudev.2014.11.016>
- Wodon, Q., Male, C., Onagoruwa, A., & Yedan, A. (2017). *Girls' education and child marriage in West and Central Africa: Key findings ahead of the October 2017 high level meeting on ending child marriage in West and Central Africa*. Washington, D.C.

- World Bank. (2011). *A new instrument to advance development effectiveness: program-for-results financing*. 111. Retrieved from <http://documents.worldbank.org/curated/en/2011/12/15590386/new-instrument-advance-development-effectiveness-program-for-results-financing>
- World Health Organization. (2019). *Ebola Virus Disease Democratic Republic of the Congo: External Situation Report 72*. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/324996/SITREP_EVD_DRC_20190528-eng.pdf?ua=1%0Ahttps://www.afro.who.int/health-topics/ebola-virus-disease
- Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. San Diego: MESA Press.
- Yamauchi, F., & Liu, Y. (2013). Impacts of an Early Stage Education Intervention on Students' Learning Achievement: Evidence from the Philippines. *Journal of Development Studies*, 49(2), 208–222. <https://doi.org/10.1080/00220388.2012.700395>