Analytic Concerns in Designing and Assessing Child Outcomes: Case Study: EGRA, Impact Evaluations, and Factor Scoring

Peter Halpin
Assistant Professor of Applied Statistics
Carly Tubbs
Advanced Doctoral Candidate
New York University
The Issue

Validly and reliably measuring student competencies or social processes for use as outcomes in impact evaluations

Pre-IE

Develop new measures or re-calibrate existing ones

Post-IE

Appropriate statistical methods to address “data challenges”

Today: A case study of how factor analysis can be applied to address challenges with EGRA data from an IE in the DRC

---

The Intervention

- The first large-scale rigorous evaluation of a program in a conflict-affected country designed to
  - Improve classroom practices and processes
  - Promote children’s academic and socio-emotional skills

This talk:
- Literacy outcomes (EGRA) for fourth grade students at baseline (AY2010 – 2011) and midline (AY2011 – 2012)

Opportunities for Equitable Access to Basic Education

2 The program was implemented by the International Rescue Committee, in partnership with RTI International, the Flemish Association for Development Cooperation and Technical Assistance, and the Institute of Human Development and Social Change at New York University, and was funded by USAID under an initiative entitled Opportunities for Equitable Access to Quality Basic Education (OPEQ).
Evaluation Design and Sample

- 3-CRT (students nested within schools nested within school administrative clusters)
  - Simplify mathematical notation by talking about students within schools

Table 1. Sample summary: Fourth grade students

<table>
<thead>
<tr>
<th>Summary</th>
<th>Baseline</th>
<th>Midline</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>Treatment</td>
</tr>
<tr>
<td>Administrative Clusters</td>
<td>53</td>
<td>20</td>
</tr>
<tr>
<td>Schools</td>
<td>84</td>
<td>33</td>
</tr>
<tr>
<td>Students</td>
<td>928</td>
<td>286</td>
</tr>
</tbody>
</table>

Note: Entries denote sample sizes at each level of clustering.
The Outcome

- Early Grade Reading Assessment (EGRA) developed by RTI International
  - Consists of 11 subtasks
- Has been implemented in 50 countries, 70 languages (Gove & Wetterberg, 2011)
- Increasingly used as an outcome in IEs (e.g., AIR, 2014; Davidson & Hobbs, 2013; Ralaingita & Wetterberg, 2011).
EGRA: Instrument Design

Table 2. Summary of EGRA sub-tests administered in OPEQ

<table>
<thead>
<tr>
<th>Variable Code</th>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGRA1</td>
<td>Vocabulary</td>
<td>Basic vocabulary and understanding of commands</td>
</tr>
<tr>
<td>EGRA2</td>
<td>Initial sound identification</td>
<td>Relationships between sounds and letters (phonological awareness)</td>
</tr>
<tr>
<td>EGRA3</td>
<td>Knowledge of graphemes</td>
<td>Read the letters of the alphabet, or say their sounds</td>
</tr>
<tr>
<td>EGRA4</td>
<td>Familiar word reading</td>
<td>Read high-frequency words quickly (fluency)</td>
</tr>
<tr>
<td>EGRA5</td>
<td>Invented word decoding</td>
<td>Decode made-up words fluently</td>
</tr>
<tr>
<td>EGRA6a</td>
<td>Oral passage reading</td>
<td>Read a short passage (50 words long)</td>
</tr>
<tr>
<td>EGRA6b</td>
<td>Reading comprehension</td>
<td>Questions about the comprehended part of oral passage</td>
</tr>
<tr>
<td>EGRA7</td>
<td>Listening comprehension</td>
<td>Follow and understand a simple oral story</td>
</tr>
<tr>
<td>EGRA8</td>
<td>Writing a complete sentence</td>
<td>Correctly write a sentence (3 words)</td>
</tr>
</tbody>
</table>
EGRA: DRC Administration and Data

- The full test took about 15 minutes and was administered orally by a trained examiner
  - Instructions were in native language, subtasks were in French
- Not all grades received each subtask
- Subtask can be scored variously
  - e.g., number correct out of total; number correct out of attempted; number correct in a certain amount of time
- Item-level data (within subtests) not available for some subtests
EGRA: DRC Subtest Data

Why this is important:

- Methods for scoring typically make assumptions about the marginal and bivariate distributions
- This includes sum scores (e.g., total score, percent correct mean score)

- Extreme floor effects have also been reported with EGRA subtasks in Ethiopia, Liberia, Mali, South Africa, and Uganda (Piper, 2009, 2010, 2011; Spratt et al., 2013)

Figure 1. Example data for fourth grade students (baseline)
How Can We Best Assess the Impact of OPEQ on Student Literacy Using EGRA?

Option 1: Subtask-by-Subtask

Appears to be the most common treatment of EGRA

1. EGRA subtask distributions do not follow single functional form
2. Recoding (e.g. dichotomization) affects psychometric properties
3. Multiple comparisons??
4. Reporting on only one or a few subtests raises concerns about reporting bias and use of resources
5. If subtasks are not perfectly reliable, decreased power and effect size for treatment effects in CRT

- Difficult to compare across subtasks/studies (e.g. increased measurement error at higher proficiency)
- Family-wise error control
- Undermines rationale for EGRA
- Loss of power
Introducing (additive, homoskedastic) measurement error into the outcome variable is equivalent to introducing an additional variance component at the within level of a two-level model (Raudenbush & Sadoff, 2008).

They showed that this decreases power for treatment effects at level two.

Halpin et al. pointed out the (obvious) implicated for Hedges’ (2007, 2011) effect sizes.

Effects standardized by total or student-level variances are too small.

Arguments are “instructive”.

General conclusion: multiple endpoint studies with unreliable outcomes yield a large number of under-powered inferences and attenuated effect sizes.
How Can We Best Assess the Impact of OPEQ on Student Literacy Using EGRA?

Option 2: Aggregate Over Subtests (e.g., summary score)

- **Advantage 1:** It is easy
  - No fuss

- **Advantage 2:** When tests are properly calibrated, it is also usually “good enough” for group level analysis
  - Works with calibrated tests

- **Disadvantage 1:** It is not easy to quantify measurement error when standard assumptions are not met
  - Don’t know how well it works in tougher situations

- **Disadvantage 2:** Interpretation
  - What does a sum score mean?
How Can We Best Assess the Impact of OPEQ on Student Literacy Using EGRA?

**Option 3: Factor Analysis (etc.)**

1. **Step 1:** Exploratory factor analysis baseline year
2. **Step 2:** Confirmatory factor analysis / measurement invariance at follow up
3. **Step 3:** Factor scoring
4. **Step 4:** Assess measurement error

**Questions:**
- What do EGRA subtasks measure?
- Are we measuring it across waves and Tx groups?
- “Estimate” of literacy factor
- How reliable are the scores?
Step 1: Overview of Factor Analysis

- **Specification:** Common factor model (Maxwell & Lawley, 1971; Bartholomew et al. 2011)

- **Exploratory uses:** How many common constructs? Which subtasks measure which constructs, and how reliably?

- **“Recent” advances:**
  - Can deal with dichotomous, ordered categorical, censored, and Gaussian outcomes via two-step WLS (Muthen & Satorra, 1995)
  - Robust and clustered SEs and goodness of fit (Satorra & Bentler, 2001)
  - Multilevel extensions (Muthen & Asparouhov, 2011)
  - Readily available (e.g., Mplus; GLLAMM in Stata; LISREL)
Step 1: Exploratory Factor Analysis for EGRA

\[ Y_{ij}^* = \nu + \Lambda X_{ij} + e_{ij} \]

where

- \( Y_{ij}^* \) is the \( K \)-vector of latent response variables for student \( i \) in school \( j \);
- \( \nu \) is the \( K \)-vector of intercepts;
- \( \Lambda \) is the \( K \times F \) matrix of factor loadings;
- \( X_{ij} \sim N(\mu_j, \Sigma) \) is the \( F \)-vector of factors or latent constructs;
- \( e_{ij} \sim N(0, \Theta) \) is the \( K \)-vector of unique factors with diagonal covariance matrix \( \Theta \);

and it is further assumed that \( X_{ij} \) and \( e_{ij} \) are uncorrelated both within and across schools.
Step 1: Exploratory Factor Analysis: Results

- **Number of factors:** A single factor model was acceptable for describing the correlations between all subtasks except EGRA7 (cf. Jimenez et al. 2014)

### Table 3. Summary of model fit

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2 (df)$</th>
<th>RSMEA (90% CI)</th>
<th>CFI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-factor</td>
<td>114.20 (27)</td>
<td>.058 (.048, .070)</td>
<td>.928</td>
</tr>
<tr>
<td>1-factor + correlated errors</td>
<td>56.40 (24)</td>
<td>.038 (.025, .051)</td>
<td>.973</td>
</tr>
<tr>
<td>1-factor without EGRA7</td>
<td>52.21 (20)</td>
<td>.042 (.028, .055)</td>
<td>.966</td>
</tr>
<tr>
<td>2-factor</td>
<td>56.06 (20)</td>
<td>.044 (.030, .058)</td>
<td>.970</td>
</tr>
</tbody>
</table>

*Note:* “1-factor + correlated errors” refers to a model with correlated errors of EGRA7 with EGRA1, EGRA2, and EGRA6B; $\chi^2 (df)$ denotes the chi-square test of model fit and its degrees of freedom; RSMEA (90% CI) denotes the root mean square error of approximation and its 90% confidence interval; CFI denotes the comparative fit index; Recommended values: RMSEA ≤ .05; CFI ≥ .95
Step 1: Exploratory Factor Analysis: More Results

- **Interpretation via factor loadings:** e.g., the factor is correlated .85 with the latent response variable for Reading Comprehension.

### Table 4. Factor loadings for EGRA subtasks

<table>
<thead>
<tr>
<th>Subtask</th>
<th>Name</th>
<th>Baseline</th>
<th>Treatment</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGRA1</td>
<td>Vocabulary</td>
<td>.422</td>
<td>.449</td>
<td>.448</td>
</tr>
<tr>
<td>EGRA2</td>
<td>Phonetic Awareness</td>
<td>.579</td>
<td>.513</td>
<td>.485</td>
</tr>
<tr>
<td>EGRA3</td>
<td>Knowledge of Graphemes</td>
<td>.853</td>
<td>.785</td>
<td>.785</td>
</tr>
<tr>
<td>EGRA4</td>
<td>Familiar Word Reading</td>
<td>.932</td>
<td>.900</td>
<td>.909</td>
</tr>
<tr>
<td>EGRA5</td>
<td>Unfamiliar Word Reading</td>
<td>.947</td>
<td>.930</td>
<td>.931</td>
</tr>
<tr>
<td>EGRA6A</td>
<td>Passage Reading</td>
<td>.932</td>
<td>.937</td>
<td>.935</td>
</tr>
<tr>
<td>EGRA6B</td>
<td>Reading Comprehension</td>
<td>.852</td>
<td>.827</td>
<td>.827</td>
</tr>
<tr>
<td>EGRA7</td>
<td>Listening Comprehension</td>
<td>.585</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>EGRA8</td>
<td>Writing a Complete Sentence</td>
<td>.719</td>
<td>.710</td>
<td>.710</td>
</tr>
</tbody>
</table>

*Note:* Baseline refers to results from the exploratory analysis presented in Table 3; Treatment and Control refer to the results from the metric invariance model described in Table 5. All factor loadings were significant at $\alpha = .001$, after adjusting standard errors to account for clustering at the school administrative level.
Step 1: Exploratory Factor Analysis: More Results

- **Interpretation via conditional expectation of subtasks:** e.g., students who are 2SD above the mean on the literacy factor are expected to score modestly on familiar and unfamiliar word reading (EGRA 4 and 5).

*Figure 2. Expected subtasks score as a function of the factor*
Step 2: Confirmatory Factor Analysis

- **Specification:** Common factor model (Maxwell & Lawley, 1971; Bartholomew et al. 2011)

- **Confirmatory Uses:** Test whether a hypothesized internal structure holds in a new population

- **For IEs:**
  - **Replication:** e.g., to confirm using midline data the one-factor structure identified using EFA with baseline data
  - **Test measurement invariance:** To ensure that the underlying literacy factor has the same interpretation in treatment and control groups (see Millsap, 2011)

- **Summary of results for EGRA:** The same single-factor model fits the treatment and control conditions, after removing EGRA7 (Listening Comprehension)
Step 3: Factor Scoring

- **Approach:** Empirical Bayes treatment of factors (cf. predicting values for random effects)

- **Use:** To obtain realizations of the factors for each student that can be used in further analysis

- **Mathematical model:** \( f(X_{ij} \mid Y_{ij}) \propto g(Y_{ij} \mid X_{ij}) \times h(X_{ij}) \)

- **Complications arise in IE context:** The prior should address nesting of observations; not dealt with in current software

- Kludgey two-step approach proposed in Halpin et al.: first get \( g \), then shrink according to \( h \)
Step 4: Estimating Measurement Error

Figure 3. Estimated standard error of measurement for the MLE of the factor at midline. SE obtained via the Hessian of the log-likelihood, evaluated at the MLE.
Step 4: Estimating Measurement Error

Figure 4. Distributions of factor scores at midline

Marginal Reliability

Before shrinkage: \[
\frac{\text{var}(\hat{Y}) - \text{mean}(S^2 E(\hat{Y})^2)}{\text{var}(\hat{Y})} = .81
\]

After shrinkage (via equation 12 in Halpin et al.): .89
Step 4: Comparison in terms of treatment effects

Table 5. Estimated effect sizes, standard errors, and intra-class correlations

<table>
<thead>
<tr>
<th>Outcome</th>
<th>(d_{WT})</th>
<th>(SE)</th>
<th>(ICC_2)</th>
<th>(ICC_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGRA1</td>
<td>0.263</td>
<td>0.127</td>
<td>0.054</td>
<td>0.049</td>
</tr>
<tr>
<td>EGRA6A</td>
<td>0.098</td>
<td>0.141</td>
<td>0.012</td>
<td>0.111</td>
</tr>
<tr>
<td>EGRA median</td>
<td>0.134</td>
<td>0.141</td>
<td>0.024</td>
<td>0.118</td>
</tr>
<tr>
<td>APC</td>
<td>0.197</td>
<td>0.158</td>
<td>0.032</td>
<td>0.140</td>
</tr>
<tr>
<td>log(APC)</td>
<td>0.209</td>
<td>0.166</td>
<td>0.067</td>
<td>0.146</td>
</tr>
<tr>
<td>Adjusted Factor Scores</td>
<td>0.285</td>
<td>0.164</td>
<td>0.077</td>
<td>0.135</td>
</tr>
<tr>
<td>Unadjusted Factor Scores</td>
<td>0.270</td>
<td>0.156</td>
<td>0.071</td>
<td>0.117</td>
</tr>
</tbody>
</table>

Note: \(d_{WT}\) denotes Hedges’ (2011) within-treatments effect size (Hedges’ equation 31), \(SE\) denotes its standard error (Hedges’ equation 32), and \(ICC_L\) denotes the intra-class correlation at level \(L\). EGRA subtasks codes are given in Table 2 and the median is over all subtasks. APC denotes the average percent correct. Adjusted factor scores were obtained using the procedures outlined in the previous section. Unadjusted refers to scores produced without taking into account clustering.
Summary

- Measuring educational and psychological constructs in new populations presents many challenges.

- Factor analysis can be used to obtain relatively small number of reliable scores, even with relatively intractable data.
  - Can improve on this approach in various ways.

- Modest but appreciable improvements in reported effect sizes.