Evaluating Assessors: Assessor Accuracy Measure and Field IRR (Inter-Rater Reliability)

Jeff Davis, Ed.D., Ph.D.

(with Gäelle Simon, Thomaz Alvares, and Abdullah Ferdous)

Management Systems International (MSI)

April 27, 2016

About the Presentation

- This presentation was prepared for the USAID workshop/webinar "Release of the *EGRA Toolkit, Second Edition:* Updated Guidance and Tools for Conducting Early Grade Reading Assessments," Bethesda, Maryland, April 27, 2016.

Outline

- 2016 *EGRA Toolkit, Second Edition*
- Practical Experiences
- Recommendations

2016 EGRA Toolkit (1)

Assessor Accuracy and IRR

- Methods have been developed to measure assessor accuracy (validity) in training (i.e., agreement with a standard) (p. 87; Annex J)
- In addition to training, it is required that assessors use IRR for evaluating consistency (reliability) with each other when they are collecting data in the field (p. 89-90; Annex J)

2016 EGRA Toolkit (2)

IRR Guidelines (Save the Children)

- Sample a minimum of 150 students for double-assessment (p. 207)
- Organize assessor pairings and collect data in the field (pp. 208-209)
- Calculate IRR statistics (pp. 205-206)
- Use benchmarks for evaluating strength of agreement (p. 206)

Practical Experiences (1)

QITABI Project (Lebanon): Quality Instruction Towards Access and Basic Education Improvement

- USAID-funded (Sept. 2014-Sept. 2018)
- Improve equitable access and learning outcomes
- Led by World Learning with subcontractors
- 2 cohorts each with 2 years of intervention
  - Cohort 1 from October 2015 to June 2017
  - Cohort 2 from October 2016 to June 2018

MSI MANAGEMENT SYSTEMS INTERNATIONAL
A TETRA TECH COMPANY

TETRA TECH

Practical Experiences (2)

Student Assessment

- EGRA in grades 2 and 3 (10 students/grade)
- 120 schools in each of cohorts 1 and 2
- Cohort 1 baseline in 2015 (completed)
- Cohort 2 baseline in 2016 (in process now)
- 12 assessor teams (4 assessors + 1 supervisor)
- 6 quality control officers (QCOs)

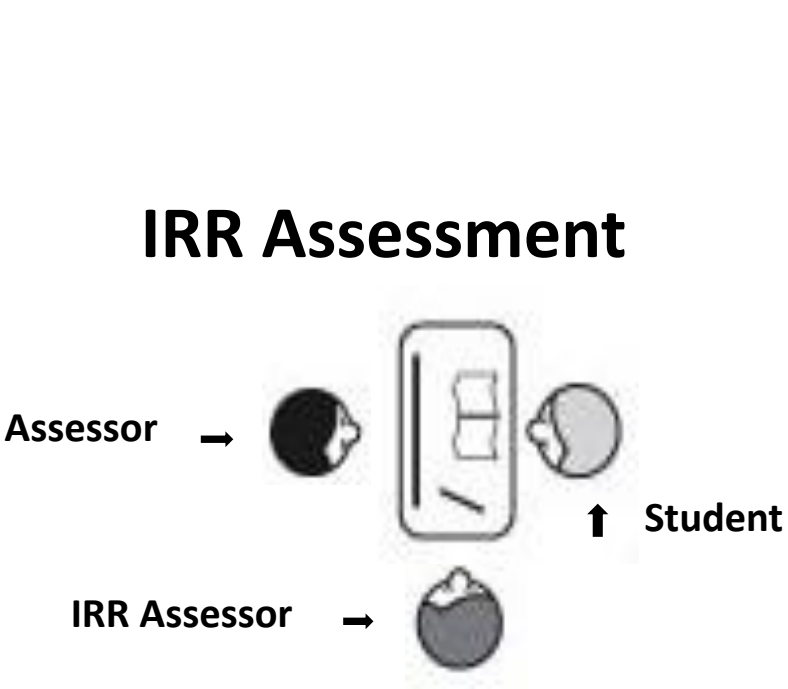
Practical Experiences (3)

Accuracy Training and IRR in the Field (Cohort 1)

- Conduct accuracy training for all assessors
  - Ensure 90% agreement with a gold standard
- Conduct a sample-based IRR study in the field
  - 240 students out of 2,400 total (10%)
  - 6 IRR assessors (1 IRR assessor / 2 teams)
  - Each IRR assessor double-assesses 4 students/day (4 students x 6 IRR assessors x 10 days = 240 students)

Practical Experiences (4)

IRR Assessment



MSI MANAGEMENT SYSTEMS INTERNATIONAL
A TETRA TECH COMPANY

TETRA TECH

Practical Experiences (5)

IRR Benchmarks

- Kappa and intraclass correlation (ICC) have interpretation categories by strength of agreement (p. 206)*
  - Less than 0.40 = Poor
  - 0.40 to 0.75 = Intermediate to Good
  - Greater than 0.75 = Excellent

* Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Practical Experiences (6)

IRR Results (Cohort 1)

- Overall estimates for the average assessor pairs
  - Kappa = 0.87 (“excellent”)
  - ICC = 0.86 (“excellent”)
- Subtask estimates for the average assessor pairs (10 subtasks across the 2 grade levels)
  - Kappa = 0.76 to 0.97 (all “excellent”)
  - ICC = 0.77 to 0.95 (all “excellent”)

Practical Experiences (7)

IRR Changes (Cohort 2)

- All assessors trained in IRR, including use of the IRR function on the data collection application
- Assessors conduct double-assessments for each other (4 per day x 12 teams x 10 days = 480 students or 20% of 2,400 total)

ANNEX 1: DAILY TEAM TESTING PLANNER

	Grade 2		Grade 3	
Time	**Enumerator A**	**Enumerator B**	**Enumerator C**	**Enumerator D**
T1	Operational 1	Operational 2	Operational 3	Operational 4
T2	IRR C	Operational 5	Operational 6	Operational 7
T3	IRR D	Operational 8	Operational 9	Operational 10
T4	Operational 11	Operational 12	IRR A	Operational 13
T5	Operational 14	Operational 15	IRR B	Operational 16
T6	Operational 17	Operational 18	Operational 19	Operational 20

Team Assignments

Date	**Enumerator A**	**Enumerator B**	**Enumerator C**	**Enumerator D**
	4 Tests + 2 IRR – Enumerators C, D	**6 Tests**	**4 Tests + 2 IRR – Enumerators A, B**	**6 Tests**
Day 1 – Example	Luna	Bilal	Rita	Nassim
Day 2 – Example	Bilal	Rita	Nassim	Luna
Day 3 – Example	Rita	Nassim	Luna	Bilal
Day 4 – Example	Nassim	Luna	Bilal	Rita

Training Dashboard

View Monitoring Dashboard

Enumerator Overview

Click on an enumerator to view their individual assessments.
Then, click on an assessment to view it compared to its Gold Standard

Gold 1	Gold 2	Gold 3	Gold 4
Gold 5	**Gold 6**		

Overall Score	Phoneme	Letter Sound	Syllable
Letter Name	Invented	Reading Vocabulary	Familiar Word

0 - 50	0047									
51 - 70										
71 - 80	0029	0046								
81 - 85	0061									
86 - 90	0042									
91 - 95										
96 - 100	0013	0014	0015	0016	0017	0018	0019	0020	0021	0022
	0023	0024	0027	0028	0030	0031	0032	0033	0034	0035
	0036	0037	0038	0039	0041	0043	0044	0045	0048	0049

MANAGEMENT SYSTEMS INTERNATIONAL
A TETRA TECH COMPANY

TETRA TECH

School Count	Grade 2	Grade 3	Total	IRR
76	761 379 M | 382 F	758 382 M | 376 F	1519 761 M | 758 F	300 151 M | 149 F

Wednesday, April 20, 2016

School ID	Enumerators	Grade 2		Grade 3		Total	IRR
		M	F	M	F		
0023	0027 0028 0029 0031	6	4	5	5	**20**	4
0049	0030 0033 0038 0049	5	5	5	5	**20**	4
0402	0046 0051 0052 0053	0	10	0	10	**20**	4
0405	0054 0055	10	0	10	0	**20**	4
0547	0044 004	5	5	5	5	**20**	4
0548	0040 004	5	5	5	5	**20**	4
0921	0058 0059 0060 0061	6	4	5	5	**20**	4
0954	0062 0063 0064 0065	5	5	5	5	**20**	4
1095	0034 0035 0036 0037	5	5	5	5	**20**	4
1156	0013 0015 0016 0021	5	5	5	5	**20**	4
1169	0022 0023 0024 0025	4	4	5	6	**19**	4

Eliane Saade (0055)

2/M/08 2/M/10 2/M/07 2/M/08

IRR: 3/M/09 IRR: 3/M/09

MANAGEMENT SYSTEMS INTERNATIONAL
A TETRA TECH COMPANY

TETRA TECH

Recommendations

- Both assessor accuracy training and IRR in the field are essential for high accuracy and consistency
- Having assessors conduct double-assessments for each other is preferred over using IRR assessors
- Programming an IRR function on the data collection application is necessary for the field work
- Calculating IRR statistics should be based on standardized, well-researched methods*

* Hallagren, A. K. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Quantitative Methods in Psychology, 8*(1), 23-34.