USAID
FROM THE AMERICAN PEOPLE

RTI International, Research Triangle Park, North Carolina, USA

Day 2 – Session 1

Measuring and Reporting Results

Africa Regional Education Sector Workshop
Dakar, Senegal
June 5-7, 2013

Acknowledgments and Purpose

- This presentation was prepared for the Africa Bureau's Education Officer Training Course under EdData II Task Order 19: Data for Education Research and Programming (DERP).

- The USAID EdData II project is led by RTI International. Task Order 19 is EdData II Task Order Number 19, EHC-E-0X-04-00004-00.

- The presentation was developed by Amber Gove with contributions from Ash Hartwell and Alison Pflepsen. Kristi Fair (USAID) provided slides related to counting against the numerical targets. Karen Tietjen provided slides related to establishing performance standards. Alison Pflepsen provided slides related to the Nigeria case study on setting performance standards. Stephen Kowal provided slides and content related to PPRs. Kristi Fair prepared the reference slides on measuring change.

- The contents of this presentation are the professional opinions of the authors and do not represent the official position of either RTI International or USAID.

Learning Objectives for Day 2

- Participants will understand how USAID's Education Goals 1 should be measured in terms of indicators, performance standards, target setting, using sampling and evaluation design to track change.

- Learn how PPR reporting differs from reporting toward the strategy's numerical targets: preparing exemplary PPRs.

- Understand SART contract and issues in submitting datasets to SART.

KEY TERMS AND DEFINITIONS

Definitions: Goal

- **Objective that a program, system or agency plans to achieve.**
- Sample goals:
  - Eradicate extreme poverty in the next two decades
  - Improve reading for 100 million children in primary grades by 2015
  - Increase equitable access to education in crisis and conflict environments for 15 million learners

Definitions: Indicator

- **A metric used to monitor or evaluate the achievement of the goal/objective over time.**
- An indicator can include specification of quantifiable targets and measures of quality.
- Examples:
  - Rate of infant deaths per 1,000 live births (www.healthindicators.gov)
  - Proportion of students who can read and understand the meaning of a grade-level text by the end of two years of primary schooling

Definitions: Performance standard

- **An established norm or requirement that provides clear and consistent understanding of what children are expected to learn, so teachers and parents know what they need to do to help them.**
- Example from U.S. “Common Core” standards for education:
- Grade 2: Read with sufficient accuracy and fluency to support comprehension.
  - Read on-level text with purpose and understanding.
  - Read on-level text orally with accuracy, appropriate rate, and expression on successive readings.
  - Use context to confirm or self-correct word recognition and understanding, rereading as necessary

Definitions: Benchmark

- Minimum level of performance that pupils need to reach in order to meet a performance standard
- Used to track pupil progress
- Know when pupils are lacking or need more instruction with a particular skill or concept
- Based on research and predict future reading success

DIBELS Oral Reading Fluency benchmarks

Second Grade: Three Assessment Periods Per Year

DIBELS Measure	Beginning of Year Months 1 - 3		Middle of Year Months 4 - 6		End of Year Months 7 - 10	
	Scores	**Status**	**Scores**	**Status**	**Scores**	**Status**
ORF	0 - 25 26 - 43 44 and above	At Risk Some Risk Low Risk	0 - 51 52 - 67 68 and above	At Risk Some Risk Low Risk	0 - 69 70 - 89 90 and above	At Risk Some Risk Low Risk

Third Grade: Three Assessment Periods Per Year

DIBELS Measure	Beginning of Year Months 1 - 3		Middle of Year Months 4 - 6		End of Year Months 7 - 10	
	Scores	**Status**	**Scores**	**Status**	**Scores**	**Status**
ORF	0 - 52 53 - 76 77 and above	At Risk Some Risk Low Risk	0 - 66 67 - 91 92 and above	At Risk Some Risk Low Risk	0 - 79 80 - 109 110 and above	At Risk Some Risk Low Risk

Putting all the terms together

- A goal in the U.S. is to have all children reading by the end of grade 3.
- The proportion of pupils meeting *basic level* proficiency on the NAEP (a reading assessment test used in the U.S.) is an indicator of progress toward achieving that goal.
- The *basic level* performance standard for 3rd grade requires that students "locate relevant information, make simple inferences, and use their understanding of the text to identify details that support a given interpretation or conclusion."
- A benchmark for the above performance standard might be a specific score on an assessment that pupils need to receive to be considered proficient

Goal 1 & 3 Standard indicators

- Proportion of students who, by the end of two grades of primary schooling, demonstrate they can read and understand grade-level text
- Proportion of students who, by the end of the primary cycle, are able to read and demonstrate understanding as defined by a country curriculum, standards or national experts
- Number of learners enrolled in primary schools and/or equivalent non-school based settings
- Number of learners enrolled in secondary schools or equivalent non-school based settings

Source: 2011 USAID Education Strategy: Technical Notes (p. 20)

http://transition.usaid.gov/our_work/education_and_universities/pdfs/2012/ED_Technical_Notes_2011.pdf

Measuring against the indicators requires:

- Clear performance standards
- Data that measures against those standards
- Results from a sample that is representative of the target population
- To demonstrate change: results from at least two, preferably three points in time (baseline, midterm and end line) from both treatment (project) and control (non-project) populations

IMPACT EVALUATION 101

Latest guidelines from USAID - Definitions

- **Impact evaluations** measure the change in a development outcome that is attributable to a defined intervention; impact evaluations are based on models of cause and effect and require a credible and rigorously defined counterfactual to control for factors other than the intervention that might account for the observed change. Impact evaluations in which comparisons are made between beneficiaries that are randomly assigned to either a treatment or a control group provide the strongest evidence of a relationship between the intervention under study and the outcome measured.
- **Performance evaluations** focus on descriptive and normative questions: what a particular project or program has achieved (either at an intermediate point in execution or at the conclusion of an implementation period); how it is being implemented; how it is perceived and valued; whether expected results are occurring; and other questions that are pertinent to program design, management and operational decision making. Performance evaluations often incorporate before-after comparisons, but generally lack a rigorously defined counterfactual.

Latest guidelines from USAID - Requirements

- All large projects should have at least a "performance" evaluation with baseline plus change
- Any project involving untested methods (i.e., in my interpretation: any method that is not more or less exactly the same as one that has previously been tested for impact) has to have an "impact" evaluation (pre- and post-, and treatment and control, or similar methods); randomization preferred, others acceptable if randomization infeasible
- Ideally externally-done

Sampling 101

- Why sampling?
- Do you have to drink whole 2-gallon pot to know how salty the soup is?
- Lowers the cost of knowing characteristics of a population, such as fluency (saltiness) levels
  - As compared to measuring EVERYONE (drinking whole pot)
  - Measuring everyone is expensive (and sometimes destructive! – soup is gone if you drink it all)
- Purpose is to select representative individuals of a specific population so as to allow generalization back to the that total population
- If you cannot generalize (not representative)—sample is no good

Sampling 101

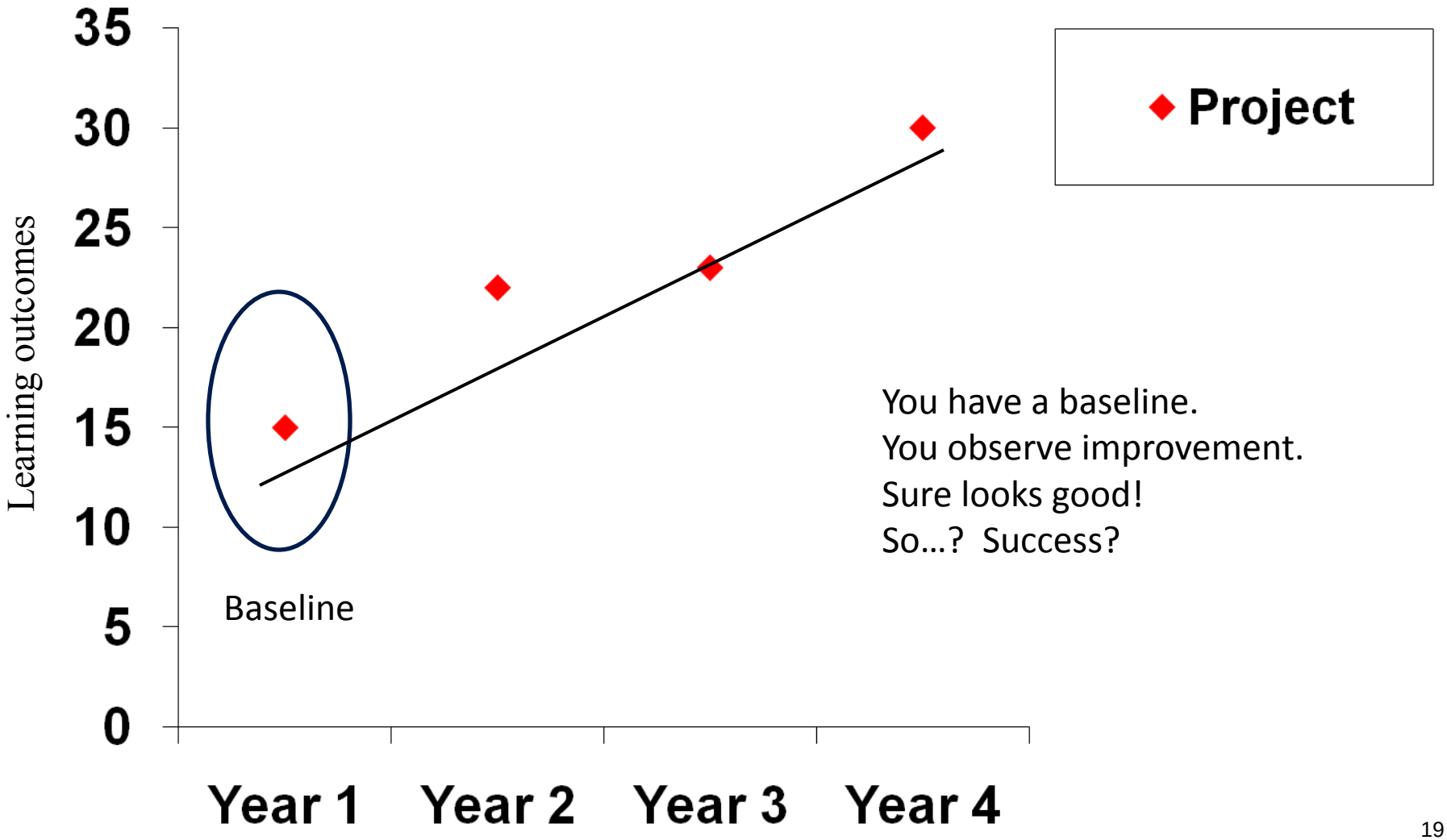
- Why random?
  - Surest way to ensure representativeness
- Why not get representativeness by saying “choose some males, some females, some urban, some rural?”
  - Cannot know all the important characteristics ahead of time, cannot enumerate them all (what else would you add?)
  - If sample is large enough, usually don’t need to anyway, you’ll get enough women by luck of the draw, IF the sample is properly random
  - We don’t know our unconscious biases, also we don’t know how characteristics distributed… maybe urban are more male, etc.

Tracking Change: The key question in impact evaluation is:

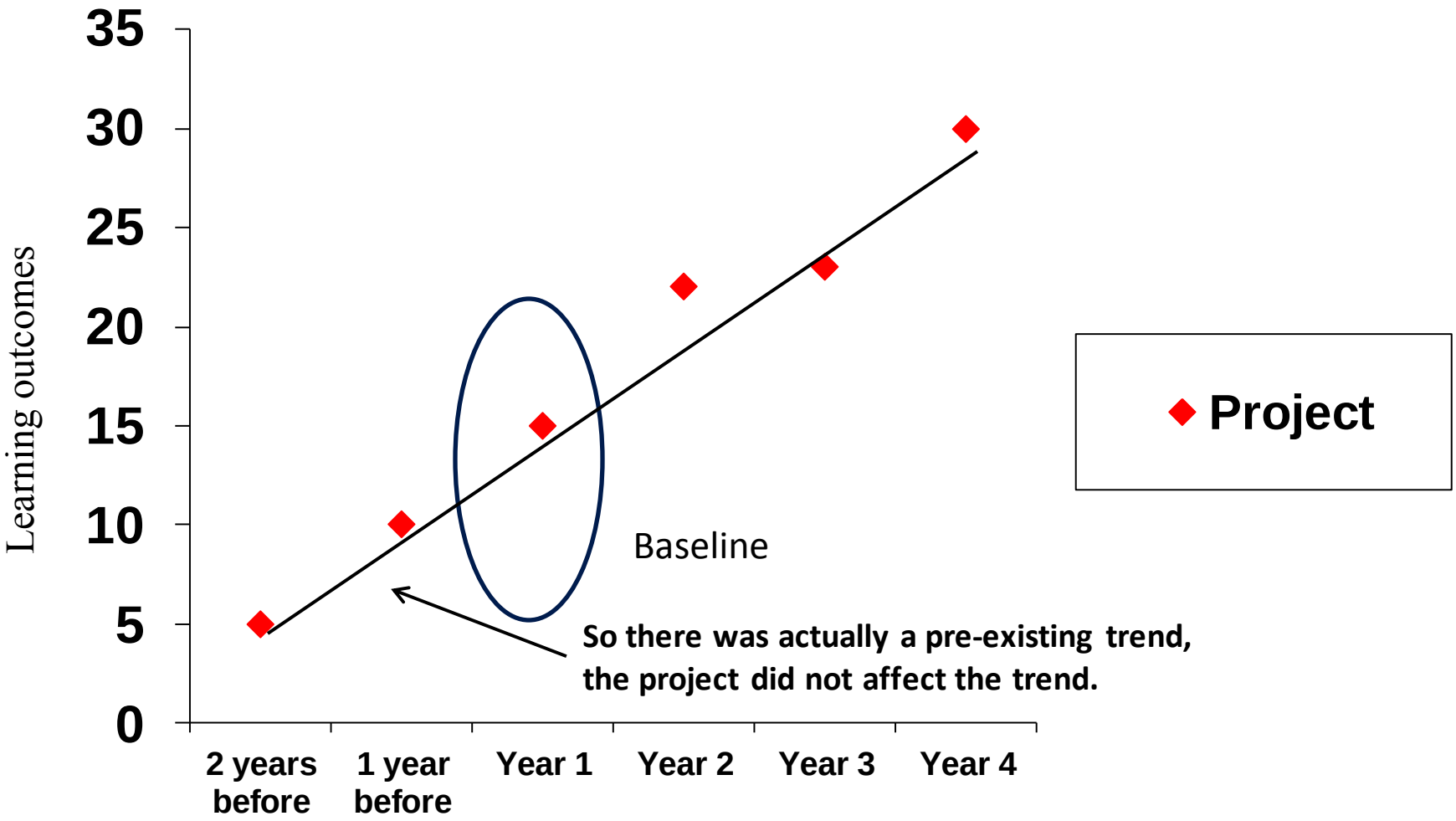
1. "What would have happened without our project, in schools that are otherwise exactly the same?"
   - If you can answer that honestly and rigorously you have a good design

2. Are we comparing our schools to schools that are—except for our project—exactly the same? In other words, do we have a reliable control group???

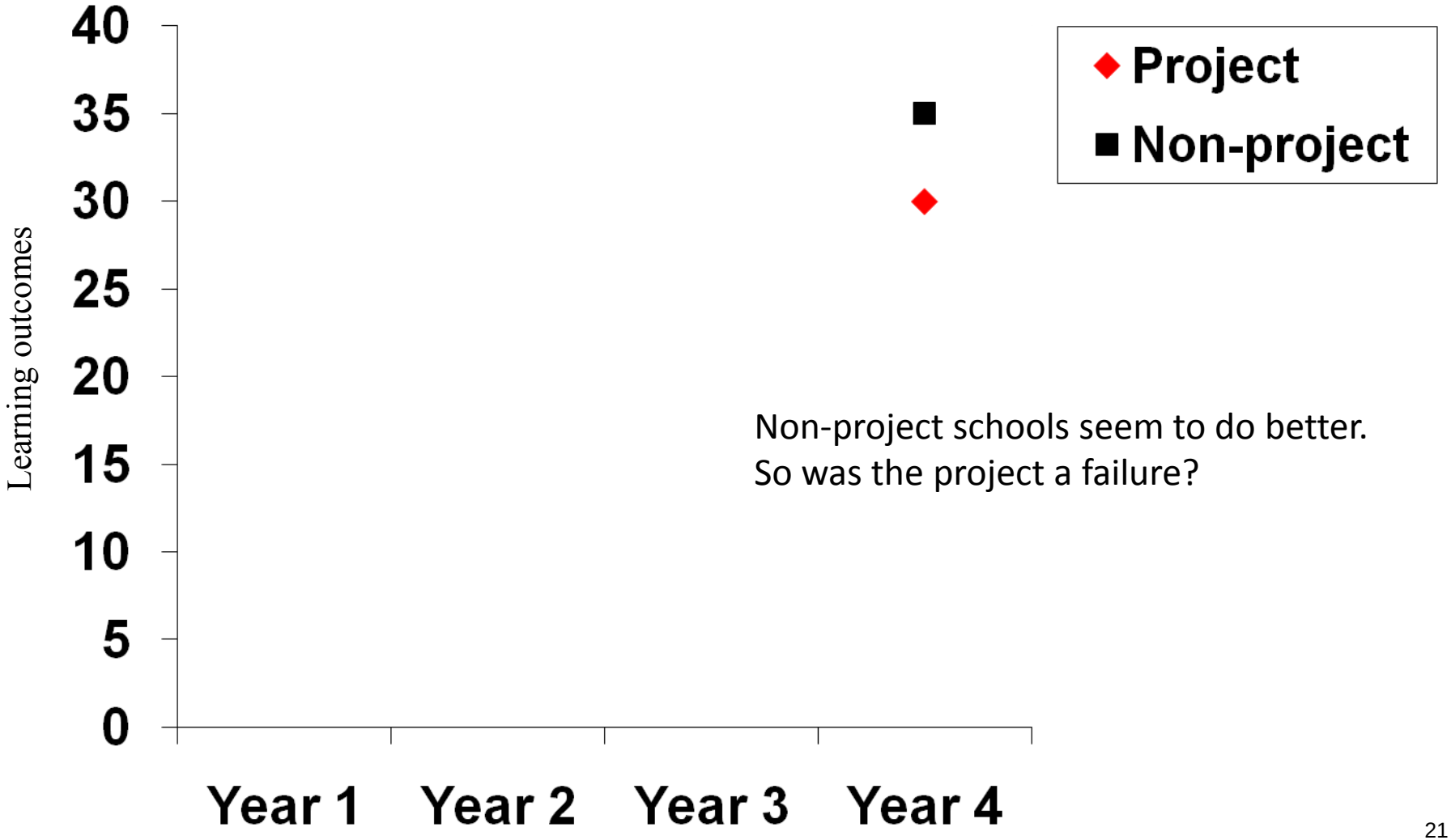
Case 1: Project measured outcomes at beginning and end of the project. Was there impact?



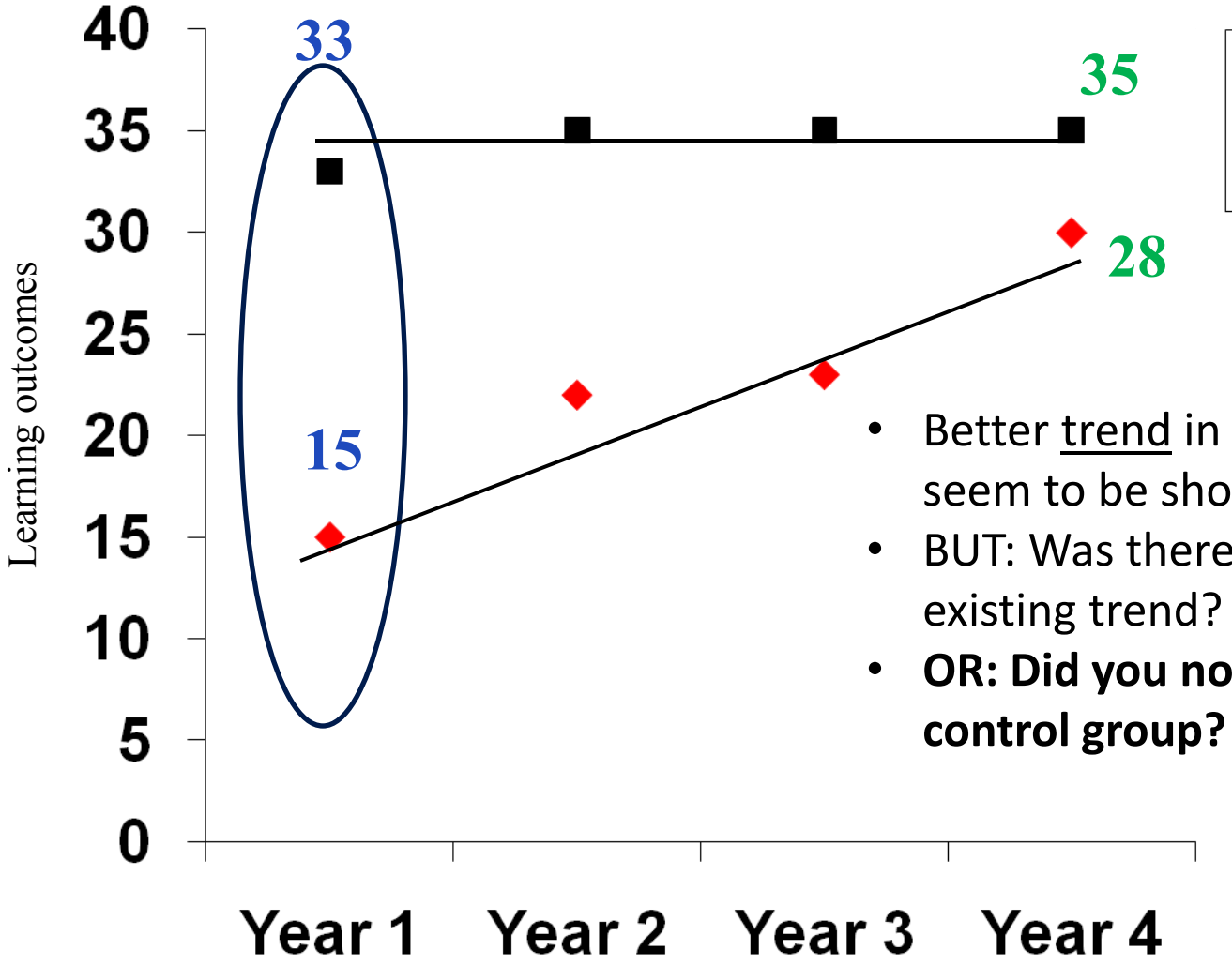
But what if this was really the case?



Case 2: Project and control (non-project) groups sampled



How about now? Sure seems good!



- Better trend in the project, so you seem to be showing impact
- BUT: Was there already a pre-existing trend?
- **OR: Did you not select a good control group? Hard to know.**

Why control groups matter so much

- Your control group (non-project schools/pupils/parents) should be as similar as possible to your treatment group.
- This helps to identify whether your control group reliable comparison group, or population whose only difference—all things being equal—was the LACK of treatment (i.e., project)
- If you do see significant differences between control and treatment groups, you may not have identified an appropriate control group—or you may have sampled incorrectly.
- How can we avoid this problem and accurately measure impact?

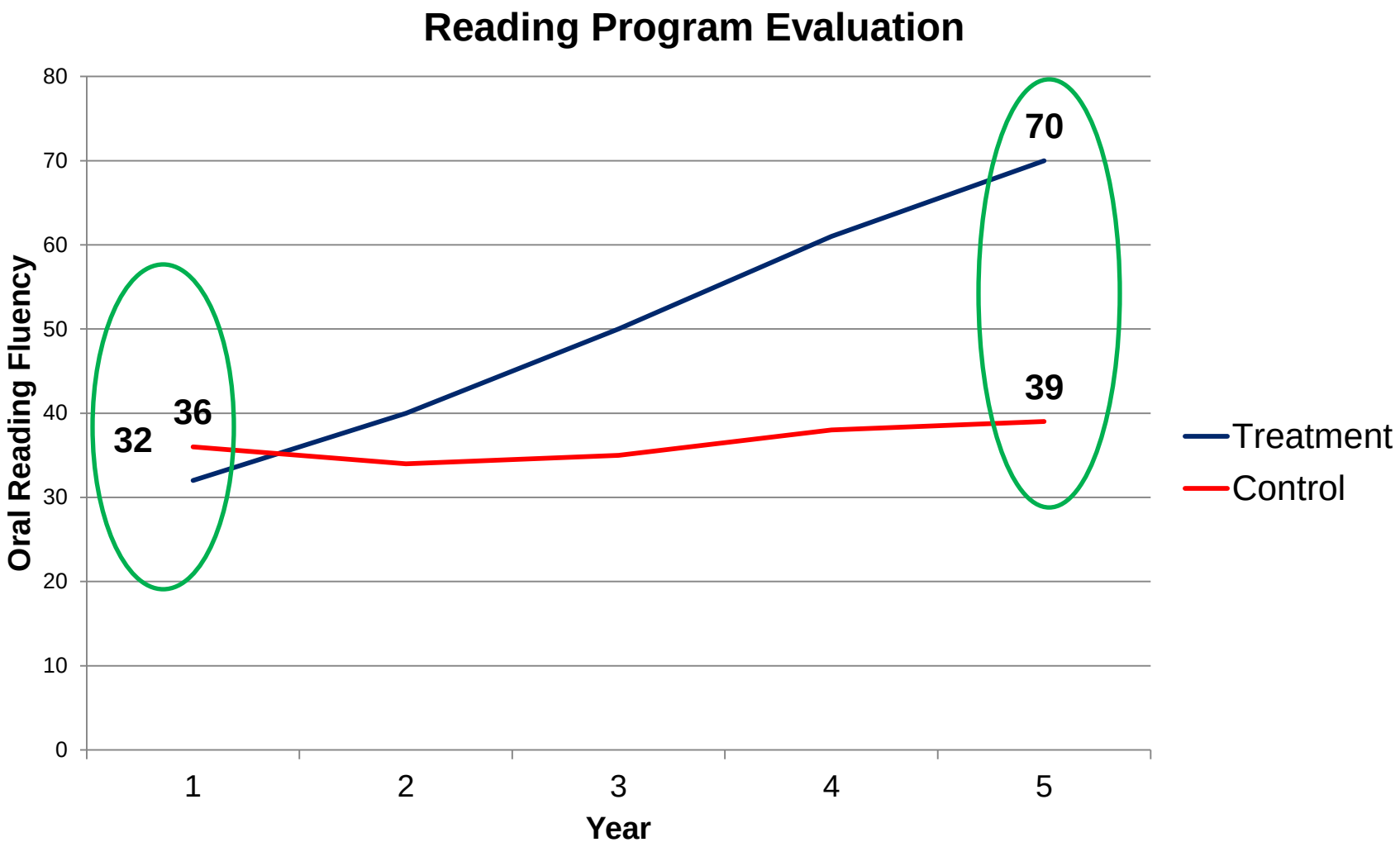
Let’s go back to sampling….

- To select our sample, we need to identify the appropriate population to sample from
- To verify we have an accurate control group, we look at our control (non-project) and treatment (project) groups’ outcome measures at baseline, as well as demographic characteristics, to see if there are major (statistically significant) differences between them

Because we can’t always be perfect…

- Difference-in-difference, or “double difference,” is a robust and common method for evaluating project impact because it takes into account both differences over TIME and differences BETWEEN a control and treatment group.
- It takes into consideration differences that may exist between the treatment and control groups.

“Difference-in-Difference”: Change over time and change between groups



Difference-in-Difference Measurement of Impact

	Control Group	Treatment Group
Baseline (Time=0)	36	32
Endline (Time=1)	39	70
Change (Time*Treatment)	+3	+38

So the actual impact is +35

*Calculation would be done by a statistical program and may not come out so perfectly, but this is the process; control for socio-economic status, geographic location, other factors if necessary

Number of students with improved reading skills

Steps in estimation, using cross-sectional samples:

1. Estimate total number of students reached by interventions

2. Estimate proportion of students demonstrating reading skill gains in representative sample, using baseline and endline data for **grade 2** and generalize to all intervention grades [To see exactly how this is done, refer to Kristi Fair's reference slides]

3. Multiply total number of students found in Step 1 by proportion showing gains (see Step 2), to obtain estimated number of students in intervention population with improved reading skills.

Estimate learners reached (denominator)

Grade	Number of learners reached, *counted only once*, by year of intervention			Total learners reached
	2013	2014	2015	
1	1,000,000	1,250,000	1,250,000	
2	750,000			
3	750,000			
Total	2,500,000	1,250,000	1,250,000	5,000,000

Measuring whether the goal is attained

- Assessments may be done in multiple grades, but the counting approach uses data from just one grade and generalizes to all intervention grades.

Rationale:

- Assessment data for every grade are unlikely to be available, so USAID chose a grade to measure change. This choice is in keeping with the GPE and USAID standard indicator measuring reading at grade 2.
- Even if we had data for every intervention grade, students may pass through several grades over time, and it is not known up front which grade students will show greatest progress in.
- And students can be counted only once, so it is not possible to look at changes in each grade in each year, and from those data, come up with a count.

DEVELOPING STANDARDS AND BENCHMARKS

Developing performance standards, benchmarks and targets

- *Step 1: Curriculum alignment*
- *Step 2: Discrepancy analysis*
- *Step 3: Reality Check*
- *Step 4: Establish Performance Standards and Benchmarks*
- *Step 5: Set targets*

Steps for performance standard setting

- ***Step 1: Curriculum alignment***
  - Does the existing early grade curriculum align with the 5 key reading competencies?
    - If ***yes***, determine if any if any enhancement required. Proceed to Step 2
    - If ***no***, re-align/adapt curriculum to emphasize/include reading, working with what is already there.

Steps for performance standard setting

- ***Step 2: Discrepancy analysis***
  - Are there gaps between existing curriculum and global or other relevant country standards?
  - Is the upper primary grade language curriculum consistent with reading requirement in other subjects, especially science and mathematics?
    - If ***yes,*** identify gaps and determine whether/which to address.
    - If ***no***, proceed to Step 3.

Steps for standard setting

- ***Step 3: Situational Analysis***
  - What are current student reading outcomes (EGRA, national exams, international tests)?
    - If unknown, conduct assessment, gather qualitative information (talk to teachers, students, parents, subject matter specialists, etc.)
  - What are the contextual factors that might affect your standards? (e.g., language complexity/transparency, L1 vs. foreign language)

→ Remember: Standards can be set "high" (i.e., children reading with comprehension by the end of grade 2)—it's your targets that will likely to be adjusted over time.

Steps for standard setting

- ***Step 4: Establish Performance Standards and Benchmarks***
  - What are the expectations for early grade reading achievement? (For example, oral reading fluency, vocabulary, levels of comprehension, etc.)
  - What constitutes mastery/proficiency? What are other relevant benchmarks?

Steps for performance standard setting

Identifying benchmarks

Method 1: Estimate the average oral reading fluency score for children who comprehend well (at least 80% comprehension or higher)

Method 2: Trend of scores of children who comprehend

Method 3: Average scores of the high-performing pupils/schools with low socio-economic status (i.e., poor pupils/schools who perform well in reading test)

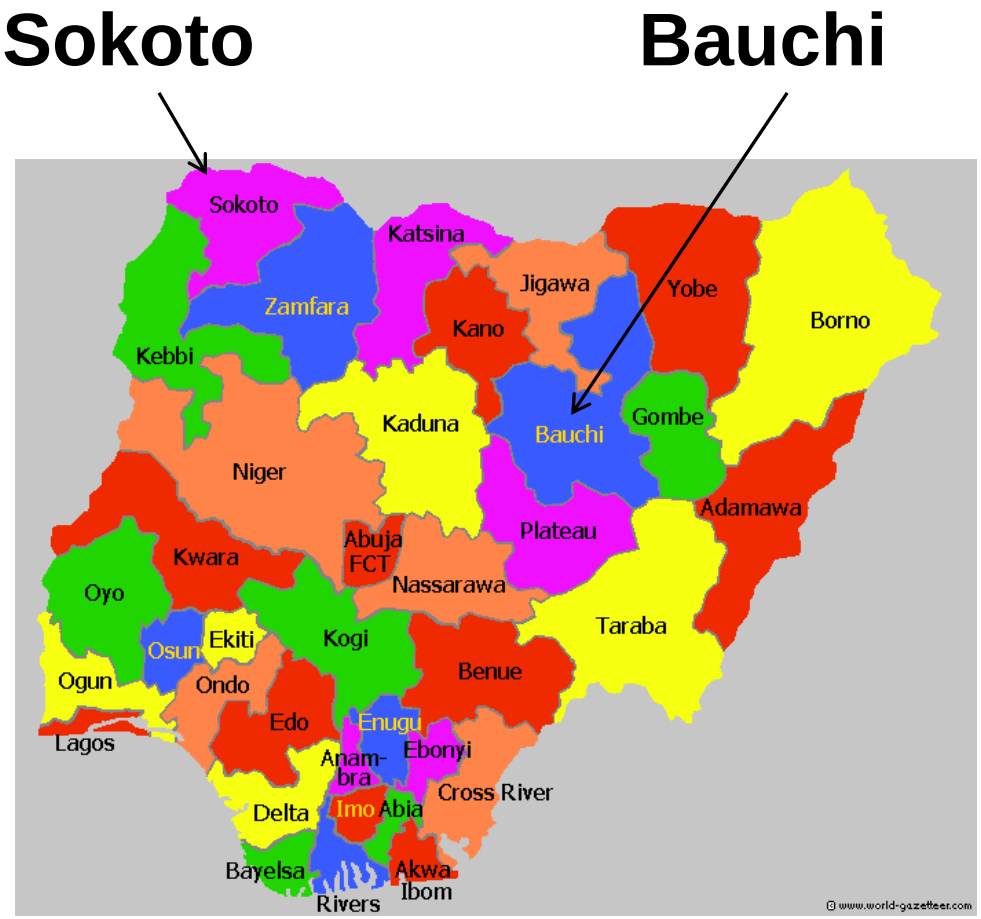
Steps for performance standard setting

- ***Step 5: Set targets***
  - What proportion of children will meet the standards' benchmarks by a certain point in time?
  - What is realistic given current levels of achievement, resources allocated and changes to the status quo that will be implemented?
    - Need to know where you currently are
    - "Easy gains" if large percentage of children with "0" scores – but then what?
    - What is the level of fidelity of implementation?
    - May need to first see what is possible

Standards setting and benchmarking: Northern Nigeria Case Study

- Northern Nigeria Education Initiative - USAID-supported program to strengthen government capacity to deliver basic education services (2009-2013)

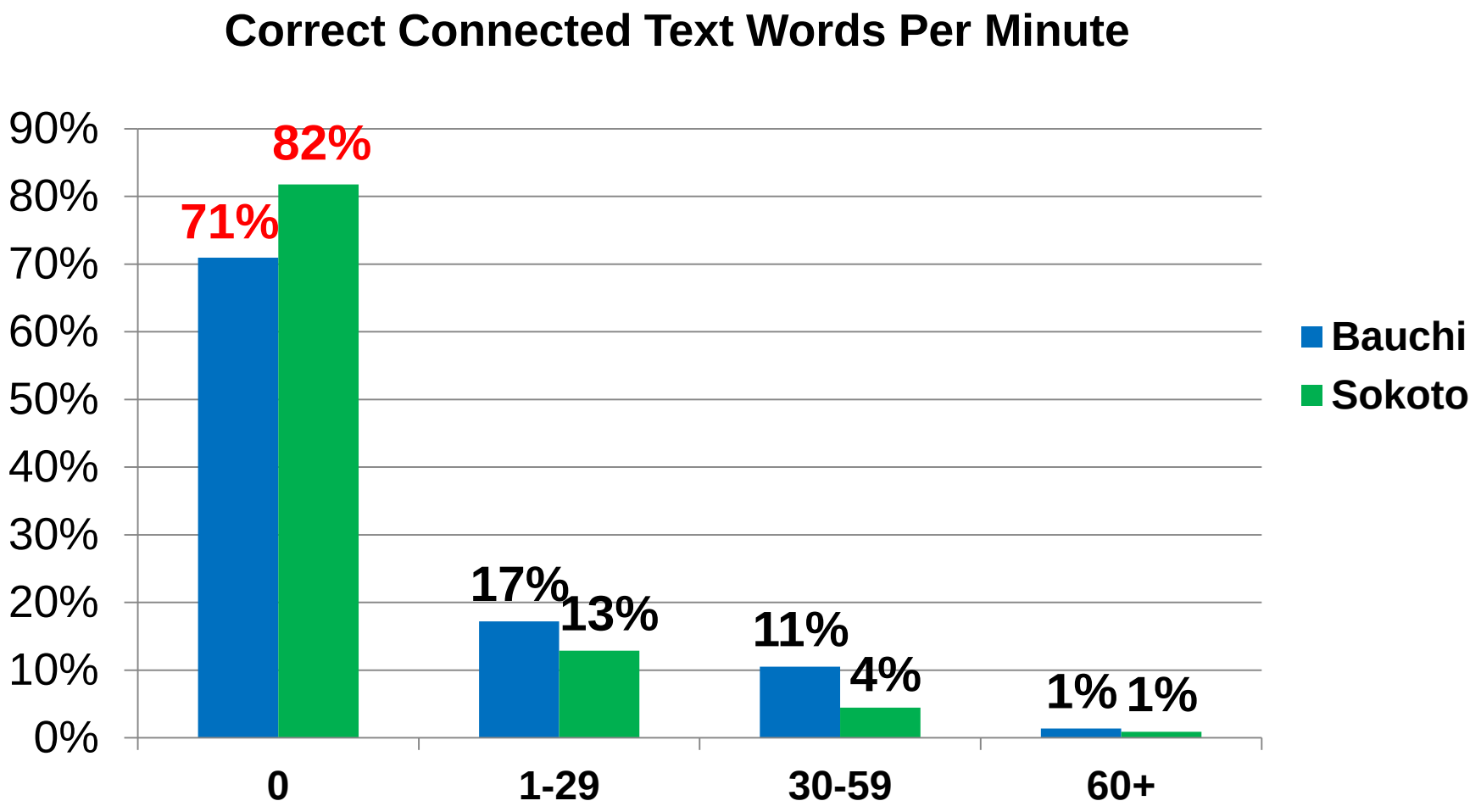
- Data from early grade reading and mathematics assessments used to inform education strategic planning and budgeting



Nigeria case study

- **Objective:** To develop contextually-specific indicators and benchmarks for early grade reading (Hausa) and math to be included in State education strategic plan M&E framework
  - Track progress over time
  - Way for the State to hold itself accountable
- **Process:** Workshop held with education officials (MOE, State Universal Basic Education Board) and other stakeholders (Colleges of Education)
  - Used data gathered from previously-conducted learning assessments to identify appropriate benchmarks

Snapshot of outcomes: Oral Reading Fluency (ORF) Results



Sample size: 4,023 pupils total in two States

Process of identifying reading benchmarks

- Identified average oral reading fluency scores of children who read the passage with at least 80% comprehension

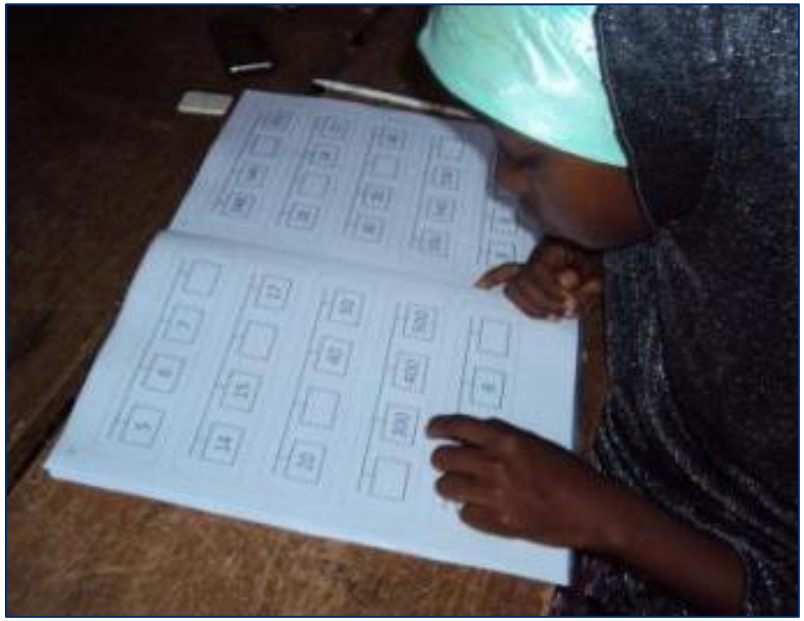
Average ORF scores of children reading with at least 80% comprehension		
Reading Skill	**Bauchi *(n=109)***	**Sokoto *(n=51)***
Oral reading fluency (average correct words per minute)	61.8	63.0

Key Indicator: Proportion of pupils who, by the end of two grades of primary schooling, demonstrate they can read and understand grade-level text in Hausa

End of P2 – Hausa Reading	Proposed Benchmark for Oral Reading Fluency (CWPM)	% at benchmark March 2011	% at benchmark May 2013	Proposed target for the end of 2015 academic year*
Non-Reader	0	Bauchi – 71% Sokoto – 82%	To be determined	50%
Emergent Reader	1-31	Bauchi – 18% Sokoto – 13%	To be determined	40%
Beginning Reader	32-61	Bauchi – 10% Sokoto – 4%	To be determined	5%
Reader	62 or higher	Both States – 1%	To be determined	5%

Math indicators and benchmarks

- Different processes used because there is no single measure (like oral reading fluency) that can be considered “the” defining indicator for measuring mathematics achievement
- Indicators and benchmarks identified for all skills measured in the EGMA; 3 were included in the State M&E plan (missing number, addition and subtraction levels 1 and 2)



Outcomes

- Empowering for government officials and education providers to be actively involved in identifying and agreeing on benchmarks
- Ownership of the results, which were adopted for inclusion into state strategic plan monitoring and evaluation framework
- Awareness of the lack of clear performance standards for reading and mathematics in the curriculum

Nigeria Case: Issues for discussion

- **Benchmarks may change**
  - Although today a child in Northern Nigeria appears to need to read an average of 62 words per minute to read Hausa with comprehension in Primary 3, this could change as teaching improves
- **Performance standards**
  - Should serve as the basis for indicators and benchmarks, but do not exist in many countries, particularly for reading.
- **Identification of performance targets**
  - Requires data over time. Need to know what is possible with improved instruction and support.

Group Exercise & Presentations

Selection of Goal 1 indicators, performance standards, targets, and the design for measuring change

1. Review the Goal 1 indicators on page 20 of the USAID Education Strategy Technical Notes. Which are applicable to your program?
2. Discuss any challenges to measuring these indicators in your country. How will they be addressed?
3. How will your country program identify benchmarks for each indicator? How will you identify targets? What is realistic and how will you know?
4. Develop (or discuss an existing) plan for measuring your program's impact, including sampling process, identification of a control group, and baseline and data collection schedule. Identify potential challenges and solutions.

Prepare a 10-minute presentation to share with the other teams.

Draw on technical guidance from Day 1, the reference materials for day 2, and the resource specialists as needed.