USAID
FROM THE AMERICAN PEOPLE

Day 2, Session 1

Measuring & Reporting Results Reference Materials

RTI International, Research Triangle Park, North Carolina, USA

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

Acknowledgments and Purpose

- These slides were 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 Aarnout Brombacher and Ash Hartwell. 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. 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 Objective

- Participants will understand how USAID's Education Goals 1 and 3 should be measured in terms of indicators, standards and targets and benchmarks.

- Understand the basic tenants of sampling and measurement.

- Understand how to track results by indicator across multiple data collection events in time and interpret the results for strategic programming purposes.

Contents

- Definitions: goal, indicator, performance standard . pp 5-8
- USAID Goal 1 and Goal 3 indicators. p9
- Measuring against these indicators. pp 10 - 13
- Steps for establishing performance standards with partner countries? pp 14 - 19
- How data should inform target setting pp 20-30
- Case Study: Nigeria. pp 31-41
- Statistics 101. pp 43-55
- Tracking change: improved reading. pp 56-79
- Reference Materials – USAID guidance, pp 80-92

Definitions: Goal

- **Objective that a program, system or agency plans to achieve.**
- Sample goals:
  - Eradicate extreme poverty in the next two decades
  - 90% of children in the country receive polio vaccine by age 2
  - 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
  - Others?

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

Putting all the terms together

- A <u>goal</u> 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 <u>indicator</u> of progress toward achieving that goal.
- The *basic level* <u>performance standard</u> 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."

Goal 1 and 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
- If trying to demonstrate change: results from at least two, preferably three points in time (B, M, E) from both treatment and control populations

Number of students with improved reading skills

Steps in estimation, using cross-sectional samples:

1. Estimate total number of students reached by interventions

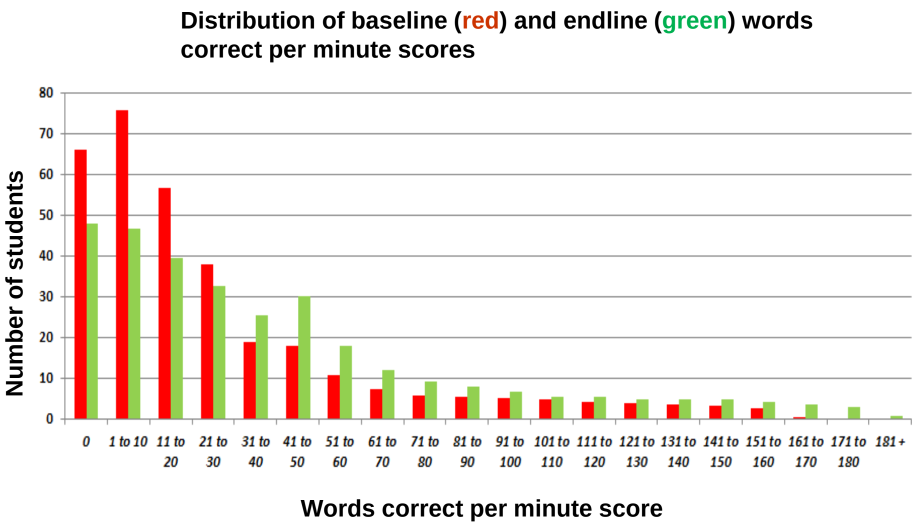
2. Estimate proportion of students demonstrating reading skill gains in representative sample, using baseline and endline data for one grade – usually grade 2 – and generalize to all intervention grades

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

Number of students with improved reading skills: Grade 2 data


Distribution of baseline (red) and endline (green) words correct per minute scores

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.
  - Is there mastery of phonics, decoding and comprehension in the upper primary grades?
    - If ***yes***, consider broadening review to reading-across-the-curriculum. Proceed to Step 2.
    - If ***no***, plan for remediation.

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: Reality Check***
  - How do student reading outcomes (EGRA, national exams, international tests) compare?
    - If no data, conduct rapid ASER-type assessment, gather qualitative information (talk to teachers, students, parents, subject matter specialists, etc.)
  - What are the contextual factors that affect reading? (E.g. language complexity/transparency, schooling resources, teachers, student language fluency and SES, etc.)
  - Are global standards appropriate for the context, are they do-able?
  - What is realistic and manageable for the country?

Steps for standard setting

- ***Step 4: Establish Performance Standards and Benchmarks***
  - What are the expectations for targeted grades? (E.g. early grade--wpm, vocabulary, punctuation, prosody, levels of comprehension and application, etc.)
  - What constitutes mastery/proficiency? What are other relevant cut points?
  - Where can country expect to be in two years, five years, etc.?

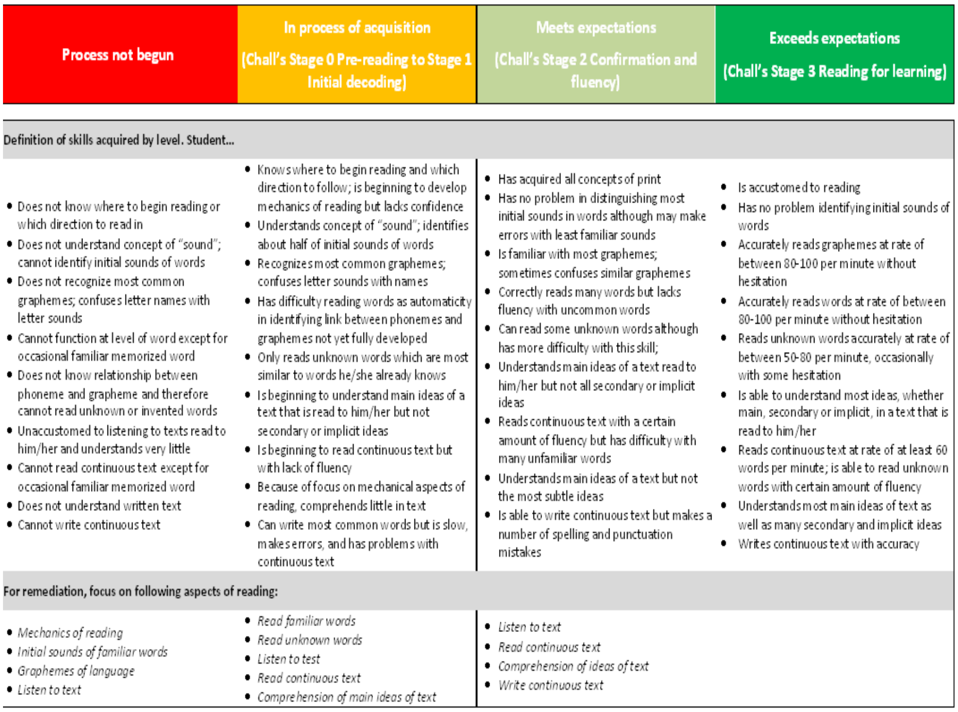
Steps for performance standard setting

- ***Step 5: Set targets***
  - What proportion of children will meet the standards' benchmarks?
  - What methods will inform target setting?

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)



Process not begun	In process of acquisition (Chall's Stage 0 Pre-reading to Stage 1 Initial decoding)	Meets expectations (Chall's Stage 2 Confirmation and fluency)	Exceeds expectations (Chall's Stage 3 Reading for learning)
Definition of skills acquired by level. Student...			
• Does not know where to begin reading or which direction to read in • Does not understand concept of "sound"; cannot identify initial sounds of words • Does not recognize most common graphemes; confuses letter names with letter sounds • Cannot function at level of word except for occasional familiar memorized word • Does not know relationship between phoneme and grapheme and therefore cannot read unknown or invented words • Unaccustomed to listening to texts read to him/her and understands very little • Cannot read continuous text except for occasional familiar memorized word • Does not understand written text • Cannot write continuous text	• Knows where to begin reading and which direction to follow; is beginning to develop mechanics of reading but lacks confidence • Understands concept of "sound"; identifies about half of initial sounds of words • Recognizes most common graphemes; confuses letter sounds with names • Has difficulty reading words as automaticity in identifying link between phonemes and graphemes not yet fully developed • Only reads unknown words which are most similar to words he/she already knows • Is beginning to understand main ideas of a text that is read to him/her but not secondary or implicit ideas • Is beginning to read continuous text but with lack of fluency • Because of focus on mechanical aspects of reading, comprehends little in text • Can write most common words but is slow, makes errors, and has problems with continuous text	• Has acquired all concepts of print • Has no problem in distinguishing most initial sounds in words although may make errors with least familiar sounds • Is familiar with most graphemes; sometimes confuses similar graphemes • Correctly reads many words but lacks fluency with uncommon words • Can read some unknown words although has more difficulty with this skill; • Understands main ideas of a text read to him/her but not all secondary or implicit ideas • Reads continuous text with a certain amount of fluency but has difficulty with many unfamiliar words • Understands main ideas of a text but not the most subtle ideas • Is able to write continuous text but makes a number of spelling and punctuation mistakes	• Is accustomed to reading • Has no problem identifying initial sounds of words • Accurately reads graphemes at rate of between 80-100 per minute without hesitation • Accurately reads words at rate of between 80-100 per minute without hesitation • Reads unknown words accurately at rate of between 50-80 per minute, occasionally with some hesitation • Is able to understand most ideas, whether main, secondary or implicit, in a text that is read to him/her • Reads continuous text at rate of at least 60 words per minute; is able to read unknown words with certain amount of fluency • Understands most main ideas of text as well as many secondary and implicit ideas • Writes continuous text with accuracy
For remediation, focus on following aspects of reading:			
• *Mechanics of reading* • *Initial sounds of familiar words* • *Graphemes of language* • *Listen to text*	• *Read familiar words* • *Read unknown words* • *Listen to test* • *Read continuous text* • *Comprehension of main ideas of text*	• *Listen to text* • *Read continuous text* • *Comprehension of ideas of text* • *Write continuous text*	

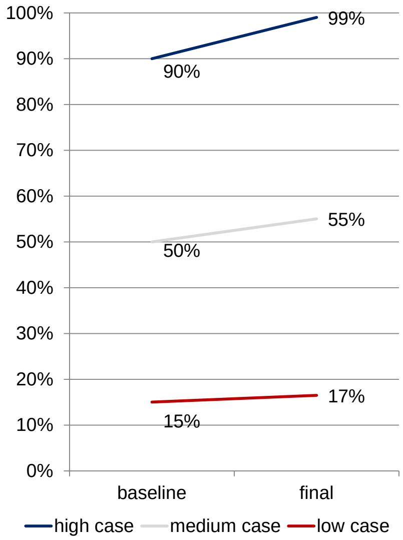
How data should inform target setting

Starting point

- Q: Under which scenario would it be easier to demonstrate a 10 percent improvement in the number of children meeting a benchmark?
  - – 15% meeting benchmark
  - – 90% meeting benchmark

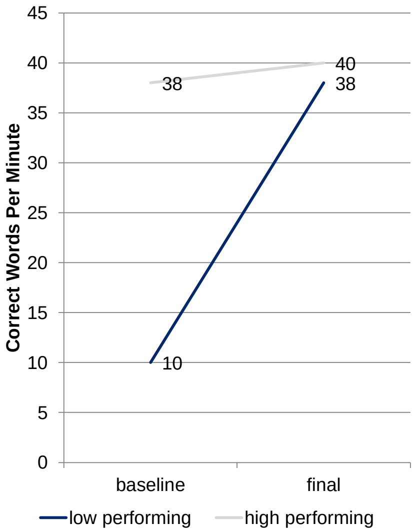
  - – HINT: be clear about the difference between percentage POINT or percent improvement

How data should inform target setting



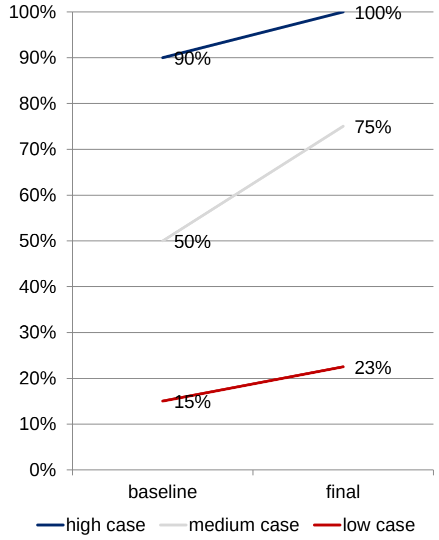
- A: If the base is 15%, 10% improvement is only 16.5%, vs. moving from 90% to 99%
- So can we set a global target of 10% improvement? NO!

How data should inform target setting



- A2: If children are scoring at a very low level, you can get dramatic gains in a short amount of time. So getting to a benchmark of 40 for a child who is at 10 is relatively easy.

How data should inform target setting



- A3: And if children are scoring at a very low level, getting at least 50% improvement is reasonable. You literally can’t get 50% improvement if 90% of kids are already at benchmark (>100%).

How data should inform target setting

Fidelity of Implementation

- Quiz: You have great results (300% improvement) from a pilot and are now ready to go to scale. However the implementation approach will change as the project will now use Ministry staff to coach teachers (in the pilot they hired their own staff)? Can you expect the same level of improvement in your scaled up program that you had under the pilot?

How data should inform target setting

Fidelity of Implementation

- A: No, because if you are not doing the same intervention, you cannot expect the same results. In this example the implementation approach has changed– Ministry staff may receive the exact same coaching but the partner does not control their salaries, transport, etc.
- How much should you lower your expectations?

How data should inform target setting

Measure

- Q: The implementing partner supplying massive quantities of books and has as its goal "to improve student motivation to read." Should EGRA be used to measure reading performance? How much gain can be expected?

How data should inform target setting

Measure

- A: No. The measure should be appropriate for the objective. (EGRA does not measure motivation).
- But if the goal of the intervention is to improve foundation reading skills, then EGRA is appropriate. The expected gain depends on the skill or competency being measured. For ORF we see typical gains between grades (in the absence of an intervention) of 14 wpm. So a reasonable goal might be a 50% (7 word) improvement above the control group.

How data should inform target setting

Intervention

- Q. You are designing a project and plan to put 60M into reading improvement with plans to work on pre-service teacher training and physical plant improvements, in-service training, support and materials for 2000 schools. How much gain can you expect given your investment?

How data should inform target setting

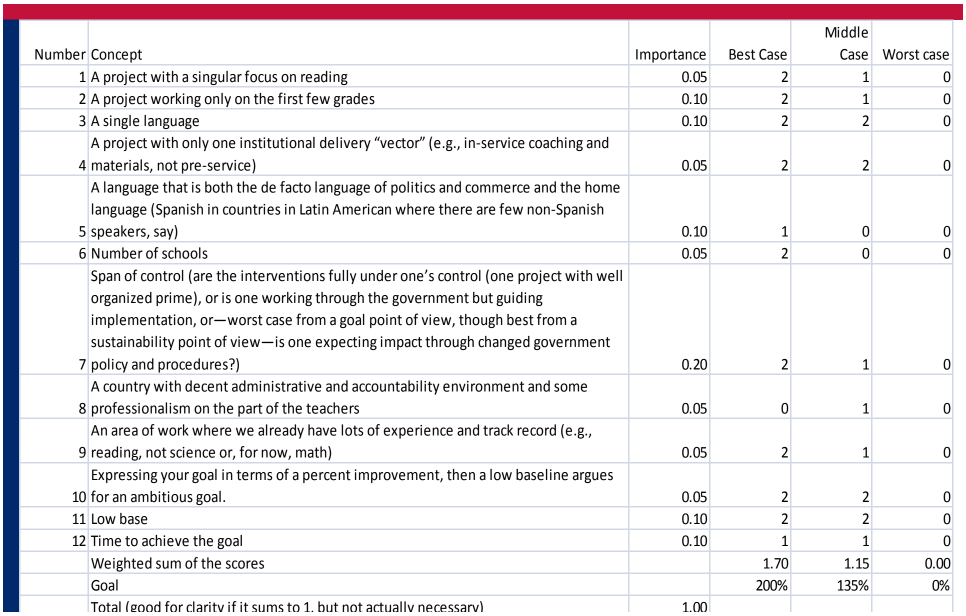
Intervention

- A. It depends on the relative weight of each component of the intervention, the starting point, and what share of the 60M will go towards moving the right levers to move reading improvement along.

- So what should go into the project plan? Let's do some calculations.

Micro soft Office
:el 2007 Workbo

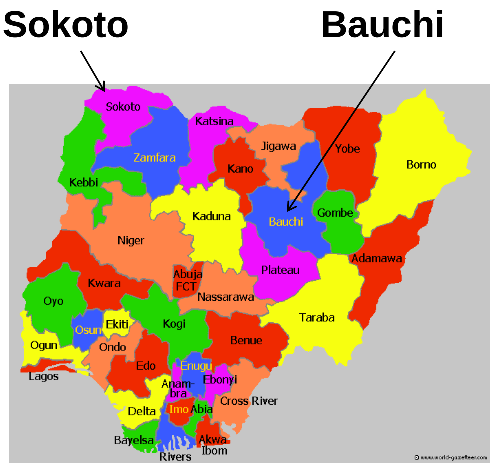
How data should inform target setting



Number	Concept	Importance	Best Case	Middle Case	Worst case
1	A project with a singular focus on reading	0.05	2	1	0
2	A project working only on the first few grades	0.10	2	1	0
3	A single language	0.10	2	2	0
4	A project with only one institutional delivery "vector" (e.g., in-service coaching and materials, not pre-service)	0.05	2	2	0
5	A language that is both the de facto language of politics and commerce and the home language (Spanish in countries in Latin American where there are few non-Spanish speakers, say)	0.10	1	0	0
6	Number of schools	0.05	2	0	0
7	Span of control (are the interventions fully under one's control (one project with well organized prime), or is one working through the government but guiding implementation, or—worst case from a goal point of view, though best from a sustainability point of view—is one expecting impact through changed government policy and procedures?)	0.20	2	1	0
8	A country with decent administrative and accountability environment and some professionalism on the part of the teachers	0.05	0	1	0
9	An area of work where we already have lots of experience and track record (e.g., reading, not science or, for now, math)	0.05	2	1	0
10	Expressing your goal in terms of a percent improvement, then a low baseline argues for an ambitious goal.	0.05	2	2	0
11	Low base	0.10	2	2	0
12	Time to achieve the goal	0.10	1	1	0
	Weighted sum of the scores		1.70	1.15	0.00
	Goal		200%	135%	0%
	Total (good for clarity if it sums to 1, but not actually necessary)	1.00			

CASE: Northern Nigeria Education Initiative - NEI

- USAID-supported initiative 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


Sokoto

Bauchi

1

Learning assessments in Northern Nigeria

- **Early Grade Reading Assessment (EGRA) in Hausa** (March 2011)
  - EGRA assesses foundational reading skills shown to be predictive of later reading achievement
  - Primary 3 learners in government schools (secular and Islamiyya)
- **Early Grade Mathematics Assessment (EGMA)** (May 2012)
  - Key numeracy skills including number identification, quantity discrimination, missing number, addition and subtraction
  - Primary 2 and Primary 3 pupils in secular and Islamiyya schools

→ Focus throughout has been on capacity-development, sustainability, and using data to improve decision-making

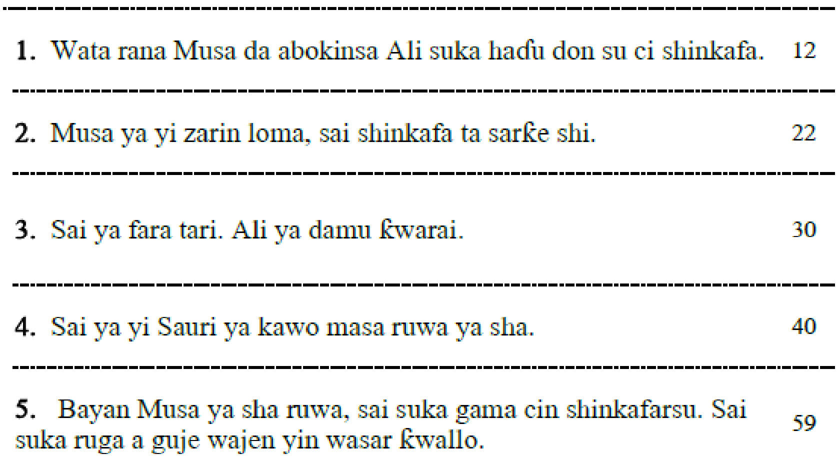
2

Developing learning benchmarks: 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

Process of identifying reading benchmarks

- Focused on oral reading fluency and related reading comprehension



1. Wata rana Musa da abokinsa Ali suka haɗu don su ci shinkafa. 12

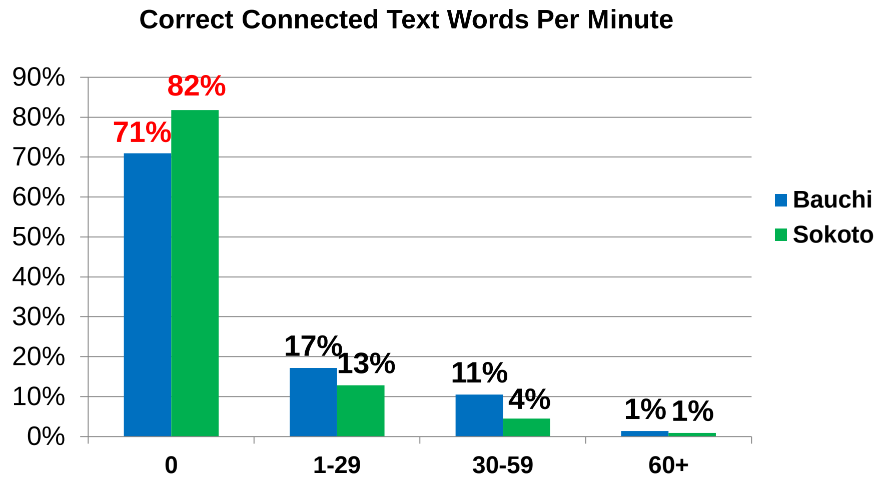
2. Musa ya yi zarin loma, sai shinkafa ta sarƙe shi. 22

3. Sai ya fara tari. Ali ya damu ƙwarai. 30

4. Sai ya yi Sauri ya kawo masa ruwa ya sha. 40

5. Bayan Musa ya sha ruwa, sai suka gama cin shinkafarsu. Sai suka ruga a guje wajen yin wasar ƙwallo. 59

Snapshot of outcomes: Oral Reading Fluency (ORF) Results



Sample size: 4,023 pupils total in two States

Process of identifying reading benchmarks

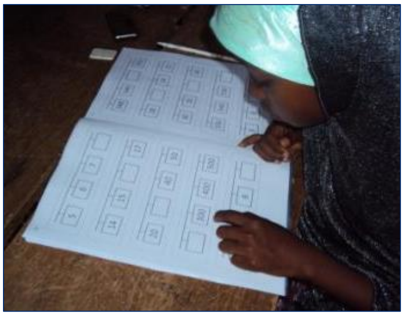
- Identified average oral reading fluency scores of children who read the passage with at least 80% comprehension

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	1-31s	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)

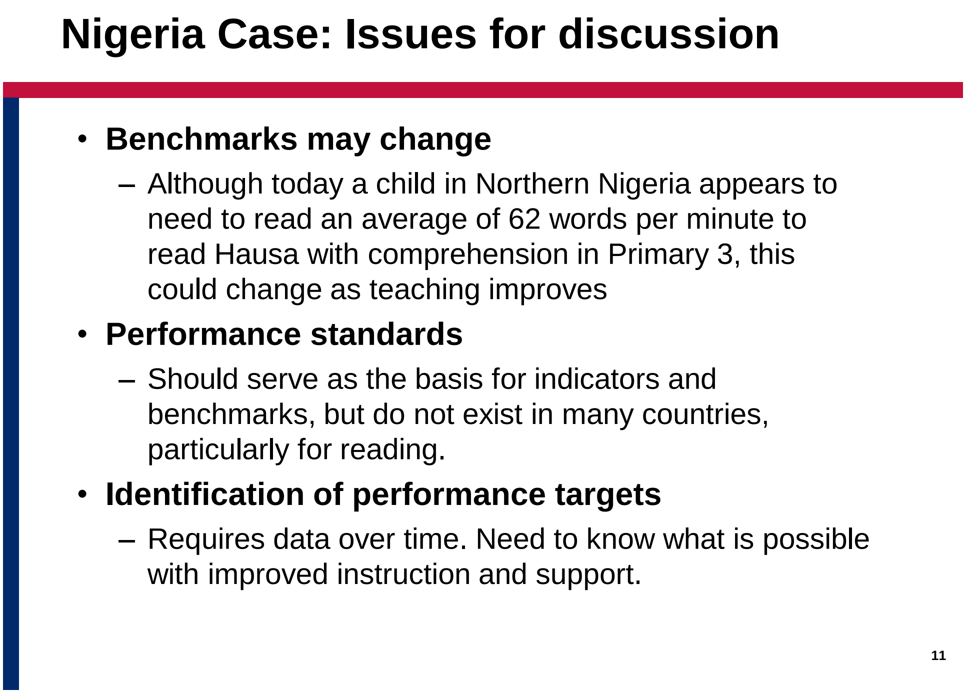


Outcomes

- 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

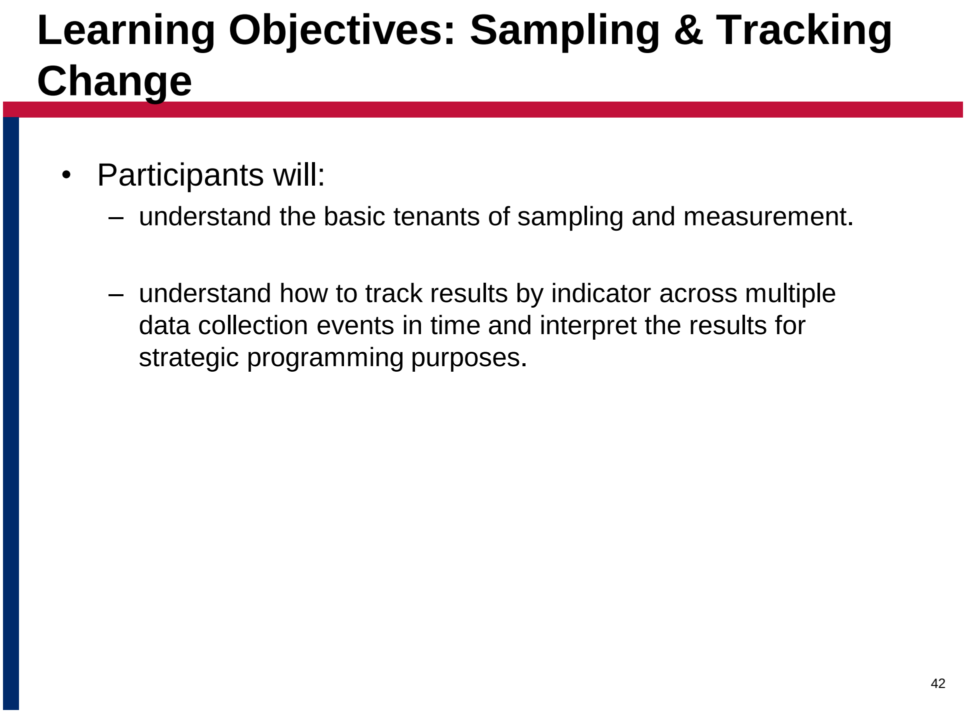
Issues for discussion

- **Reading benchmarks**
  - How can we measure comprehension more comprehensively – yet still gather data in a timely and efficient manner at the lower grades?
  - For many languages, we do not yet know what an “appropriate grade-level text” is
- **Mathematics benchmarks**
  - No one skill can serve as *the* indicator against which to measure progress.
  - What are the pros/cons of using many indicators? Of a composite score?



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.



Learning Objectives: Sampling & Tracking Change

- Participants will:
  - understand the basic tenants of sampling and measurement.
  - understand how to track results by indicator across multiple data collection events in time and interpret the results for strategic programming purposes.

Sampling 101

- Why sampling?
- Do you have to drink whole 2-gallon pot to know how salty the soup is?
- Lower 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 so as to allow generalization back to the total population
- If cannot generalize (not representative), no good

Sampling

- 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.

Sampling

- Why a mistake to say “make sure you include 3 special-needs schools, or 3 project schools, or 3 private schools, in the sample?”
  - First, you cannot generalize from the 3 to all of them as a subset of the population, anyway
  - Second, if the sample is pretty large, say 100, the characteristics of the special schools will not affect the total much
  - Third, you don’t know that the right proportion is 3 out of 100, so if you include special categories out of proportion, you have to “weight” them in creating the total averages—it is more hassle
  - Fourth, for all but the smallest categories, a large enough random sample will include SOME of them anyway, but you STILL cannot generalize to the subset unless the sample is huge

Sampling

- Most of the time random is best
  - Surest way to ensure representativeness
  - Pure random is always proportional to the population, so it is representative
  - It is the best way to make sure the sample “stands for” the population, if sample is large and truly random
- There are valid exceptions
  - Case studies for HOW things are happening (not HOW MUCH) is happening
  - Can force certain categories of interest (e.g., project schools) to be included BUT:
    - Need to include enough to be able to generalize about THEM or else it is totally pointless
    - Need to weight them in proportion to their population, when you construct total averages

What is wrong with % rules of thumb

- They are incorrect because they do not guarantee representativeness OR cost too much
  - All fixed % levels will generate under-representativeness or over-expense
  - Example: you KNOW 50% of teachers in a population of 10 are untrained, 50% are trained
    - If reason by %, what is the sample % that will generate a proper estimate most of the time? What if you do a 50% sample? What extremes do you get? How often?
    - Really need about 90%!!
  - But if you have 100 teachers, sampling 90 is overkill
  - Think through the intuition

"Real" way to figure out sample sizes - 1

- It only depends on two things:
  - How sure ("confident") you want to be of what you are saying
  - How variable the actual population is
  - And does NOT depend on population, except for small populations
- First: how sure: "we are 95% sure that the real or population fluency is 50 words per minute, plus or minus 4"
  - So, do you want to be 90%, 95%, or 99% sure?
  - And, do you want the plus or minus to be1 word or 2 or 3 or 10?
  - The more sure you want to be, the larger the sample
  - The smaller the "plus or minus," the larger the sample

“Real” way to figure out sample sizes - 2

- Second: how variable is the population?
- Intuition by going to extremes:
  - In a “low variance” pop of 100, ½ the kids read at 24 cwpm, ½ at 26
    - What’s the average?
    - What’s the farthest off you’ll be even with a sample of 10? **Ever**?
  - In another, “hi variance” pop of 100, ½ the kids read at 15, ½ read at 35…
    - What’s the average?
    - What’s the farthest off you can be?
    - How far off could you be with a sample of 10?
      - Remember you have 50 kids reading at 16, and 50 reading at 35
      - There is a very strong chance that a LOT of your kids would be reading only at 15, if your sample is only 10

No real universal rules of thumb with the “real way”

- Remember: no % rules of thumb, EVER
- There are SOME rules of thumb for special cases
- There are some equations that are fairly simple for 95% confidence
- For assessing proportions or percentages (e.g., % of teachers that are trained, % of kids with books):
  - Sample size ≈ 4 (guess prop trained * (1 – guess proportion trained)) / margin^2
  - How to guess the prop trained? Worst case (safety) is 50%
  - If your margin is 5 percentage points, then
  - Sample size = 4 * (0.5 * (1-0.5)) / .0025 = 400
  - If your margin is 10 percentage points, then
  - Sample size = 4 * (0.5 * (1-0.5)) / .01 = 100

However… Some cautions

- This applies only if you pick the teachers or kids completely at random from a national list
- That is impractical (such lists often don’t exist) and costly (may have to go to a school just for 1 kid)
- So use “clusters”: pick schools at random and then pick kids at random within the school
- But 100 teachers chosen completely at random are more representative of the teachers than if you choose 25 schools and THEN 4 teachers per school
  - Because teachers vary a lot between schools; by limiting the number of schools you limit the representativity
- To correct for this we recommend doubling or tripling the sample size, say (to follow above example) 300 teachers, 5 per school in 60 schools

Another caution

- You need the same sized sample for ANY sub-population you care about
- So if you need 50 schools for a national sample…
  - And you want same confidence at province level that you’d have at national level, and you have 10 provinces, you need 500 schools
- Soup analogy
  - Yes, a sip will do, if soup is stirred (sample is random), no matter how large the pot
  - But if you have 30 small pots (30 districts), not one large pot (just the nation), then you need an equal-sized sip from EACH pot, no matter how small the pot (unless VERY small)

How about for averages rather than proportions? - 1

- There are no rules of thumb based on standard formulas, like for proportions
  - In proportions we can use a standard formula because the variability is known if you guesstimate the proportion (discuss example)
  - With things like fluency (or pupil-teacher ratios—things that are continuous numbers), we don't know the variability ahead of time
- But, we have enough <u>experience</u> that we can make some recommendations; the variability is pretty similar across countries
  - Thus, you use prior information to make a good estimate of the sample size needed

How about for averages rather than proportions? – 2 National diagnostic

- We recommend about 500 students per "cell"
  - About 10 students in 50 schools
    - Already takes into account clustering, in most countries
- If you just want a national sample of all kids in one grade, both genders, one language, then 500
  - This will enable you to say: "We are 95% sure the fluency is X, plus or minus 6."
    - Since inter-grade gain is around 12 to 15, a "plus or minus" factor of 6 makes sense
- But noting soup analogy, you will need 500 for ANY cell of interest. So, if want gender accuracy, need 1000. If want gender AND two grades, need 2000, etc. If want gender AND 5 regions AND 2 grades, need 500 X 2 X 5 X 2 = 10,000. And so on.
    - This is what is needed if you want to make the have the same confidence in each cell as in the national example above

How about for averages rather than proportions? – 3 Project baseline

- About same
- BUT: need about 500 in the control and 500 in the treatment groups! (10 students per school in 50 treatment and in 50 control schools.)
- This assumes that the project will be able to have an impact of at least ½ of a normal inter-grade gain, and you want to be able to prove that
- If your likely gain is less, you can still prove it, but will need larger and larger size
  - But a project improvement of less than ½ of the normal inter-grade gain is not substantively interesting, so why bother proving you had that impact?
  - Thus, setting the sample size at around 500 makes sense

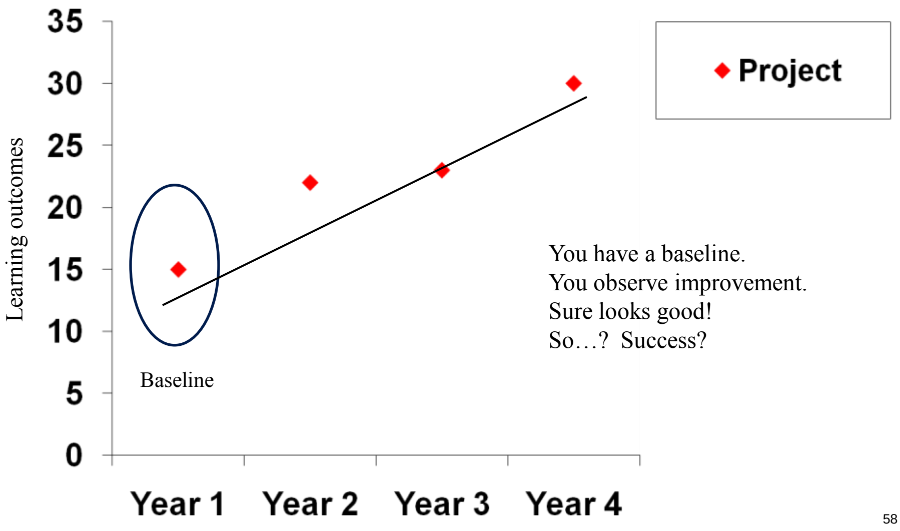
Tracking Change: Improved Reading Do a baseline

- Do a baseline
- Do a baseline
- Do a baseline
- Do a baseline
- Do a baseline
- Do a baseline
- What is the message in this slide?
- However, doing a baseline is NOT enough, especially to prove impact

Tracking Change: The only true key question in impact evaluation is:

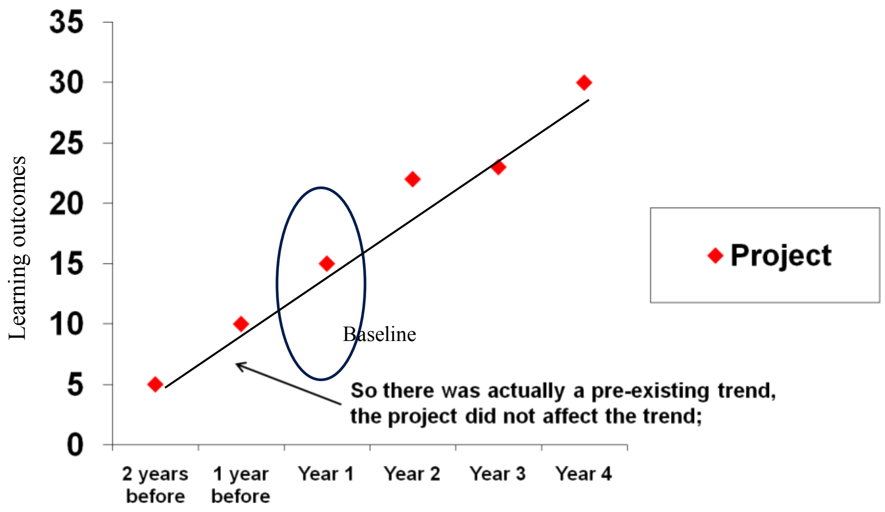
- “what would have happened without our project, in schools that are otherwise exactly the same?”
  - If you can answer that honestly and rigorously you have a good design
  - And if you can’t, a baseline is, in any case, not very good
    - E.g., if no control group
    - Or a bad control group
  - Let’s see why
  - (Lingo: all this is equivalent to saying you need a “counterfactual”: the counter to the treatment)

Case: no control schools: was this a good project?



You have a baseline.
You observe improvement.
Sure looks good!
So…? Success?

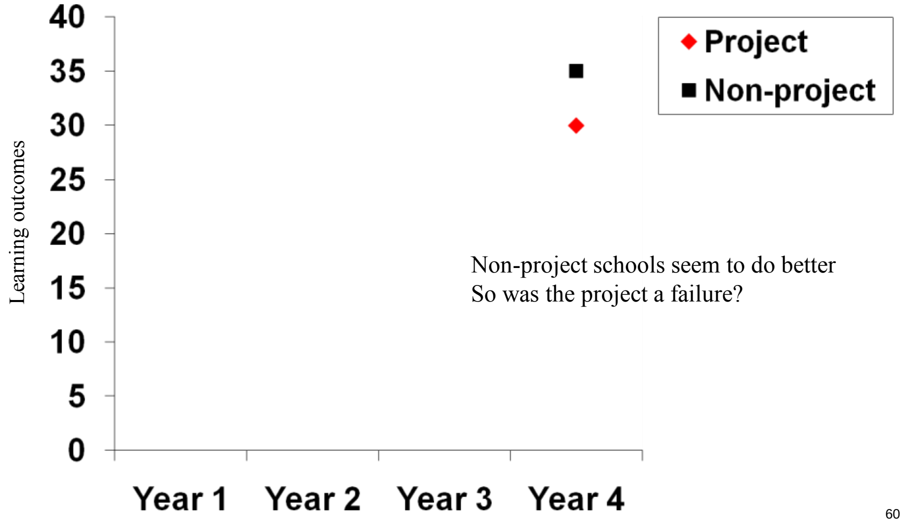
But what if this was really the case?



Baseline

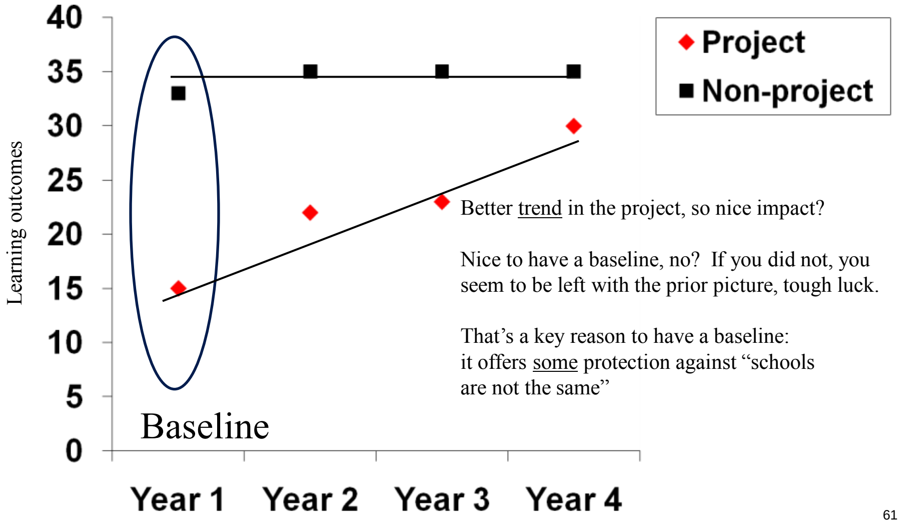
So there was actually a pre-existing trend, the project did not affect the trend;

Take this case: Project no good?



Non-project schools seem to do better
So was the project a failure?

How about now? Sure seems good!

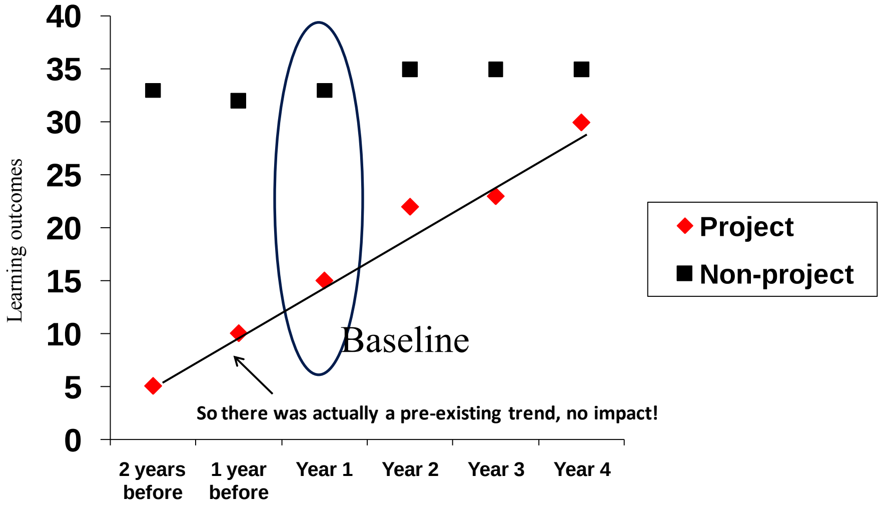


Better trend in the project, so nice impact?

Nice to have a baseline, no? If you did not, you seem to be left with the prior picture, tough luck.

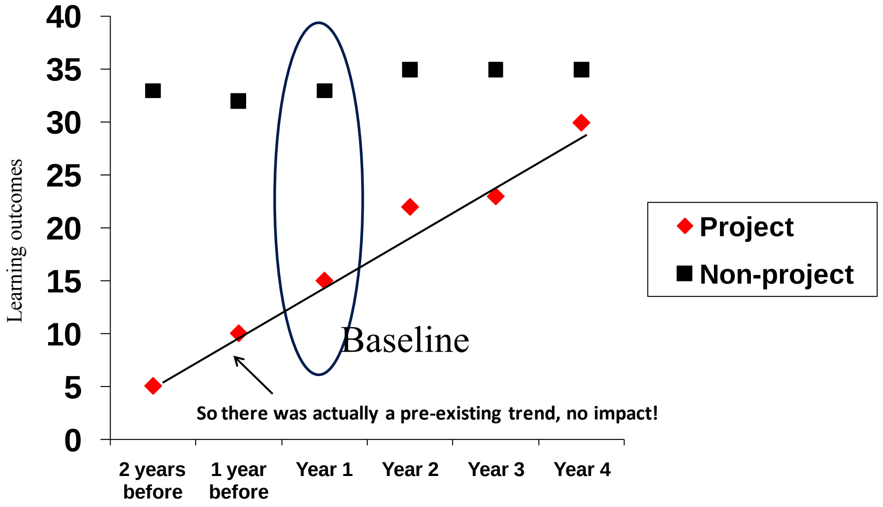
That’s a key reason to have a baseline: it offers some protection against “schools are not the same”

But, wait!



So there was actually a pre-existing trend, no impact!

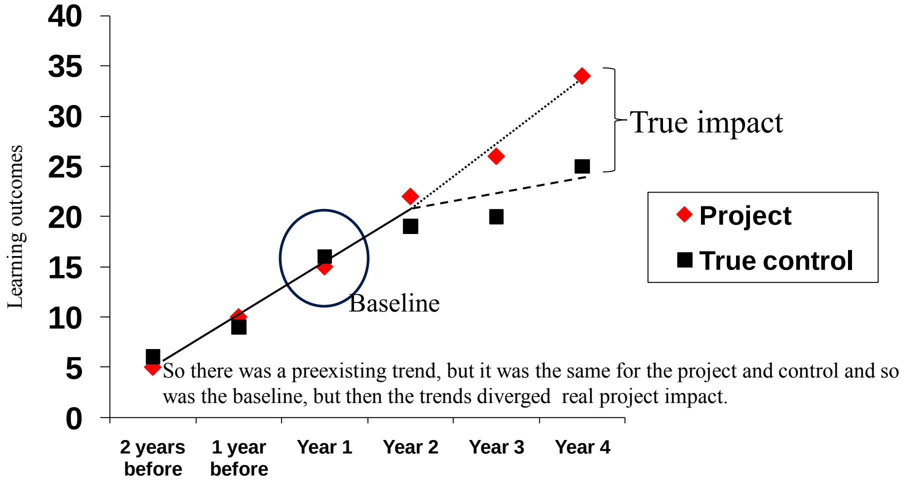
Had there been proper control (“all other things truly equal”) this would not be possible



So there was actually a pre-existing trend, no impact!

63

Instead, with proper controls, you’d see something like this if there was impact, and the size of the impact is noted…



So there was a preexisting trend, but it was the same for the project and control and so was the baseline, but then the trends diverged real project impact.

64

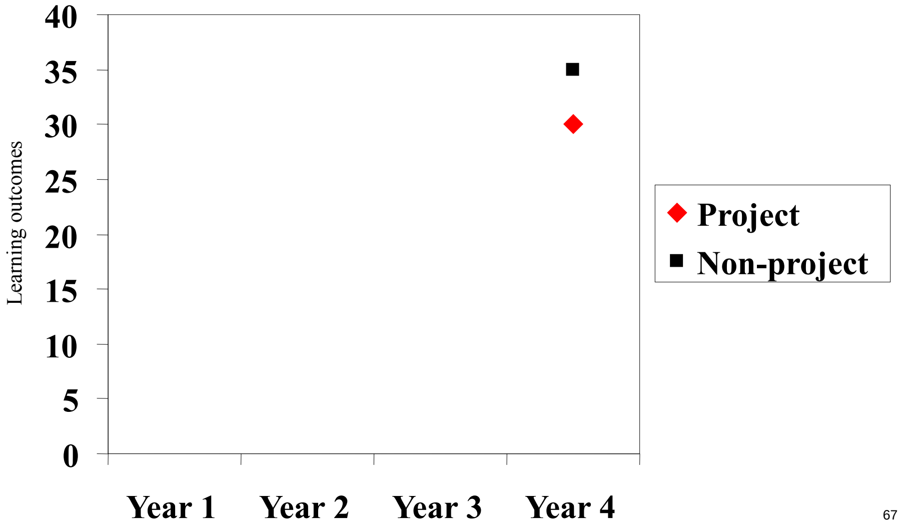
So we can have baselines… But…

  - If there are no control schools, or “bad” controls (good controls: “exactly the same except for the intervention”), then baseline can be misleading
  - We want proof of causality not just correlation… baseline without good controls will not do that…
  - Problem is that the other schools were not exactly the same as the project schools
    - They were probably richer (thus they had higher scores all the time)
    - But they were less enthused (so their scores were not improving)
  - The project schools perhaps volunteered (or got chosen) into the project because they were already go-getters on an improvement trend
  - Note that, thus, having a baseline is not good enough; there may have been pre-trends… but no one can collect pre-trends…!
- And that is why having baselines is good, but is not good enough

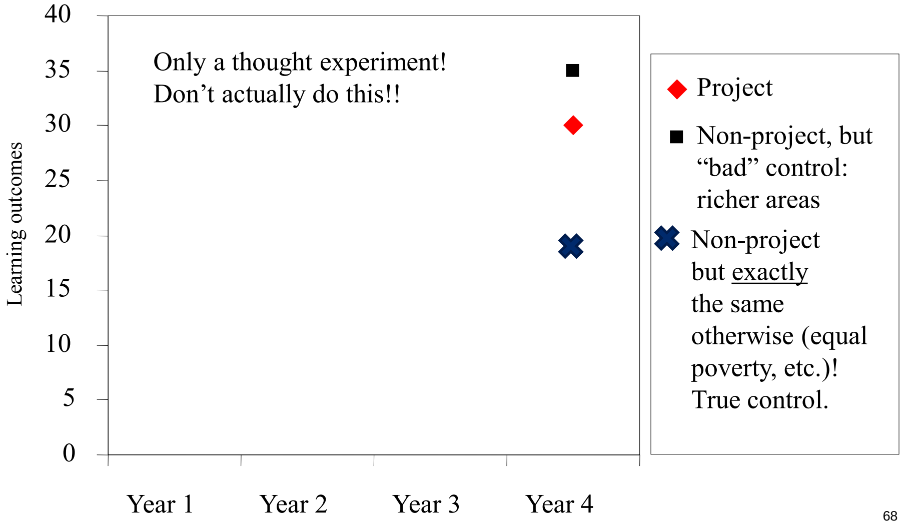
Controls

Having proper controls is just as important (maybe more so?) than having a baseline…

Ok, NOW what? : Back to “project no good?”



Ok, NOW what? Think again: project good!


Only a thought experiment!
Don’t actually do this!!

What’s the difference?

- You are now comparing the project schools against EXACTLY similar schools
- In some sense it does not matter where they started, so no need for a baseline if the schools are truly, exactly the same OTHER than the project
- But you still want to do a baseline because…
- TALK, but “answers” in the next slide

Importance of baselines

- Important for tracking and monitoring if not as much for proof of impact
  - ■ Most importantly, gives you evidence of how to fix things: baseline of associated factors
    - E.g., if kids aren’t reading, do those with books read more?
    - Do teachers with generic in-service teach better, or do those with specific in-service and in-school teach better? That justifies materials provision, or shifting emphasis to in-service, for instance
- Adds to the standard of proof
  - can show that project schools that are exactly similar not just ended higher but improved faster
  - Extra insurance: to not have baseline you have to be awfully sure your control schools are “otherwise exactly the same”
- Can track if there was leakage (discuss—important!)
- Upshot: baselines are key, but not enough… And (below) we will also see that you can salvage things somewhat if you did not do one

What’s the lingo mean?

- How do you find schools that are exactly the same?
- All the lingo is about how to do that
- Well, you can’t… quite…
- It is impossible to make project and non-project schools the same on EVERY feature (income, gender of principal, etc.) other than the intervention
- So what do you do?
- All these will also show that there are some ways to create control schools <u>even if you were not able to measure a baseline, under certain conditions</u>

What’s the lingo mean?

Now for some lingo… about types of impact evaluation, if there is time…

Ways to “make schools the same”

- Randomization
  - The best, no two ways; but still not perfect
  - Assign schools (or teachers or kids) to project and control at random
  - Make them the same by having lots of them, luck of the draw and law of large numbers will make sure both groups “the same” on any criteria, not just the ones you can think of and observe
  - Can do baseline, but need not
  - But if you don’t, you do still have to assign the schools to the project at random at startup
  - Objections? Ethics? NGOs, contractors say: “Please don’t give us schools at random?”
    - Randomize ‘within’ – randomize entry order – take advantage of natural lotteries

Ways to “make schools the same”

- Case-matching
  - Maybe project schools already known (maybe even forgot baseline)
  - So, find other schools that are “as similar as possible”
  - Look at ALL of their characteristics systematically, find the non-project schools that are statistically closest to the project schools
    - You cannot just control for the things that will make you look good
    - You cannot just control for 1-2 things that make intuitive sense
  - Then compare their test results
  - Advantage: Need not be done in advance, need not (but can) use a baseline
  - Pitfall: not always possible; if project schools are the 400 worst in the country then there is no match; or if they were originally chosen with some fairly un-matchable criterion (schools the NGO implementer already knew)
  - Imperfect matching is a pitfall, can’t match for unobserved things, but randomization can match

Ways to “make schools the same”

- Discontinuity studies
  - Project schools already known, were chosen based on some criterion such as poverty, e.g., “all schools below poverty line”
    - Must have selected based on that criterion only
  - Or “all the schools with worst test scores”
  - Solution: pick a random sample of the schools right below and right above the cutoff (these will be project and control)
    - The schools will be quite similar to each other, because they differ slightly along the “targeting” variable; need not use baseline (but better with)
  - Then compare their outcomes
  - Advantages: need not be done in advance, is more solid than matching cases, because closer to random
  - Disadvantage: not so easy to find such cases, measures impact only for those close to the cutoff

“Differences in differences”

- Difference between project (pre- and post) and control (also pre- and post); so “differences in differences”; it is the two differences that are the “real” program effect
- Can be applied to any of the previous approaches
- Allows you to estimate trends, not just final impact, which is more powerful
- And you should have a baseline anyway, so may as well use it

Monitoring vs. impact (“performance vs. impact”?)

- Baselines and ongoing measurement are important to monitoring and management, even if not so much for proof of impact
- And these are important; otherwise nothing happens anyway
- And when a “technology” (true technique or just an approach) is already proven, but simply needs to be scaled-up, you don’t need much rigorous proof of impact (e.g., vaccination); you know there will be impact, so you just monitor and manage
- This is when simple “bean-counting” of outputs (not outcomes) or even inputs (shots given) is quite justified
- But the Agency and the int. community need causal proof especially of new things; that is an impact issue, not a management or monitoring issue
- In education, since interventions are not as guaranteed as in health, may need to keep proving, not just monitoring
- Plus, because implementers make a difference, even as a manager, you may want proof that your implementers are effective

What about real life?!

- Sometimes things get messed up
  - The experiment gets all messed up as too many uncontrolled factors come in
  - Solution: measure all the other things as much as possible; there are statistical techniques to control for these factors, though not perfect
- There are unexpected consequences
  - Emphasis on impact might narrow focus away from important things that are harder to measure, such as policy reforms
    - The lesson is not to shy away from things such as policy reform, but to raise the bar on that as well
    - Remember the motto: get as close to outcomes as possible; challenge yourselves and your contractors
    - But make it appropriate
    - Also, a focus on rigor might help appropriately narrow the interventions to where they have a chance of working; don’t do sophisticated policy reforms in a war-torn country…

What about real life?!

- There are still severe measurement issues
  - Are you measuring the intervention or the intervener?
    - Lack of placebo in any of these methods prevents one from assessing this
    - Yet knowing who is a good intervener (a good NGO or a good contractor) has value—but it is a different value from knowing that the technique has impact
  - Control might self-treat: leakage
    - If techniques are easy to copy, may "leak" from treatment to control
    - No good way to address this issue because clustering is good (discuss)
    - Confounds the research but is actually evidence of effectiveness
    - Very hard to detect without baseline: a key reason for baselines
  - Treatment groups may "crowd in" other inputs; control group might demoralize; all this biases the results even of the "gold standard" measures

REFERENCE SLIDES

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

Counting against the indicators

- The following slides were provided by Kristi Fair from the January 2013 Data and Metrics Community of Practice meeting, held in Washington, DC.

- These slides are provided as reference should participants want to get into the details of how a multiple thresholds approach might be used for calculating against the Goal 1 indicator.

Proportion and number of students with improved reading skills

So, if denominator = 5,000,000, we can say that 28%, or 1,400,000 children demonstrated reading gains

Year	% of non-readers (0 wcpm)	% reading 1 - 40 wcpm	% reading 41 - 70 wcpm	% reading 71+ wcpm	% with reading gains
2012	22	63	12	3	
2015	16	47	20	17	
Net pct. point change	+ 6	NA	+ 8	+ 14	**+ 28**

Adjustment to remove double-counting when “Non-readers” become “Level 2 readers”

Year	Non-readers	Level 1 readers	Level 2 readers	Reading gains
Baseline 2012	70%	20%	10%	
Endline 2015	20%	35%	45%	
Improvement	+ 50 percentage points: FEWER non-readers		+ 35 percentage points: MORE Level 2 Readers	~~+ 85 points?~~
Adjustment : 2015 Level 2 – (2012 Level 1 + 2012 Level 2) = 45% - (20% + 10%) = 15 points double-counted				+ 85 points – 15 points double counted = **+ 70 points**

Limitations of the approach

- Two thresholds offer greater precision than just one, and less complexity of calculation than three or more thresholds. Greater precision and “capture” of improvement is gained with more thresholds, but at considerably greater complexity.
- Cross-sectional nature of samples results in a strictly gross estimate of “percentage point gain”.
- Extrapolation from a single grade sample to represent a broader range of grades in the intervention population is acknowledged as a “necessary leap” to control evaluation costs and reduce estimation complexity.
- The method is sensitive to the location of thresholds and definition of levels: A different result will obtain if thresholds or category (level) criteria are changed.

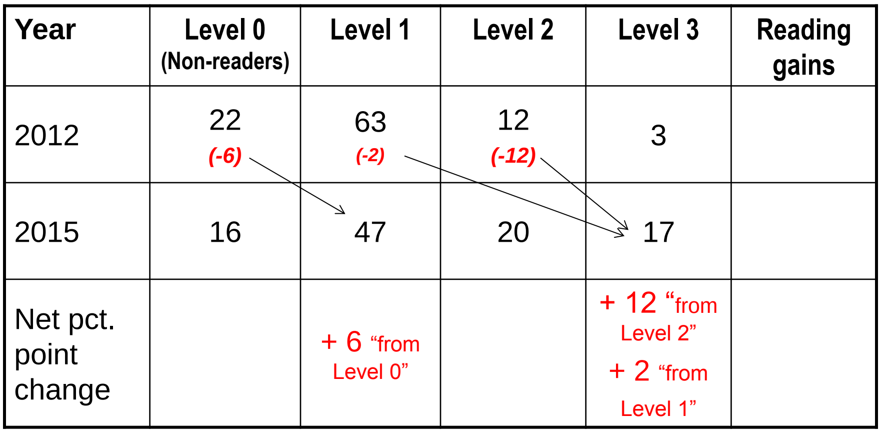
Application of adjustment to more than 2 thresholds

Year	Level 0 (Non-readers)	Level 1	Level 2	Level 3	Reading gains
2012	22	63	12 *(-12)*	3	
2015	16	47	20	17	
Net pct. point change				+ 12 “from Level 2”	

Application of adjustment to more than 2 thresholds

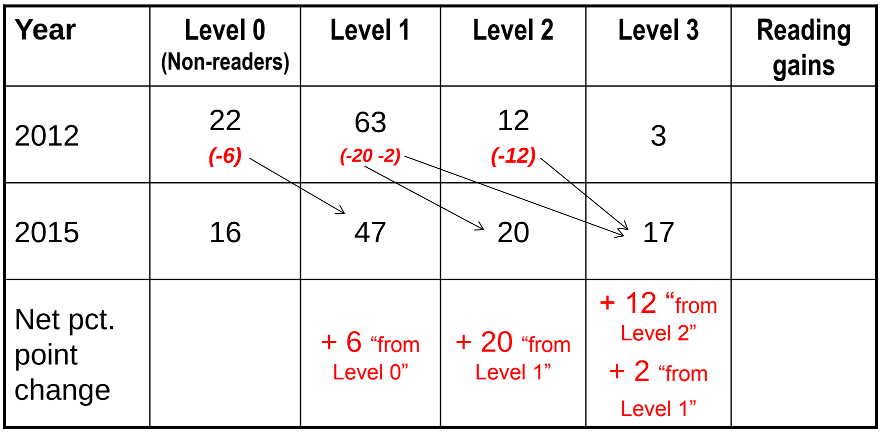
Year	Level 0 (Non-readers)	Level 1	Level 2	Level 3	Reading gains
2012	22 *(-6)*	63	12 *(-12)*	3	
2015	16	47	20	17	
Net pct. point change		+ 6 “from Level 0”		+ 12 “from Level 2”	

Application of adjustment to more than 2 thresholds



Year	Level 0 (Non-readers)	Level 1	Level 2	Level 3	Reading gains
2012	22 (-6)	63 (-2)	12 (-12)	3	
2015	16	47	20	17	
Net pct. point change		+ 6 "from Level 0"		+ 12 "from Level 2" + 2 "from Level 1"	

Application of adjustment to more than 2 thresholds



Year	Level 0 (Non-readers)	Level 1	Level 2	Level 3	Reading gains
2012	22 (-6)	63 (-20 -2)	12 (-12)	3	
2015	16	47	20	17	
Net pct. point change		+ 6 "from Level 0"	+ 20 "from Level 1"	+ 12 "from Level 2" + 2 "from Level 1"	

Application of adjustment to more than 2 thresholds

Year	Level 0 (Non-readers)	Level 1	Level 2	Level 3	Reading gains
2012	22 (-6)	63 (-20 -2)	12 (-12)	3	
2015	16	47	20	17	
Net pct. point change		+ 6 “from Level 0”	+ 20 “from Level 1”	+ 12 “from Level 2” + 2 “from Level 1”	+ 38 percentage points

Application of adjustment to more than 2 thresholds

Year	Level 0 (Non-readers)	Level 1	Level 2	Level 3	Reading gains
2012	22 (-6)	63 (-20 -2)	12 (-12)	3	
2015	16	47	20	17	
Net pct. point change		+ 6 “from Level 0”	+ 20 “from Level 1”	+ 12 “from Level 2” + 2 “from Level 1”	+ 38 percentage points