

Going to scale with oral assessments

READ INDIA CONFERENCE

University of Pennsylvania - South Asia Center
Pratham
October 3-4, 2008

Luis Crouch
lcrouch@rti.org

Oral and large scale

- Oral and large scale: possible
- Oral and large scale and high-quality and inexpensive? Harder
- Problems
- Non-problems?
- Solutions to the problems

Stop and recall: why oral?

- Need for quality/achievement indicator
- International donors lame on quality
- Quality proxied by learning
- Quality and access traditionally track but maybe only under “normal” conditions...Evidence: qual can go off the track.

Lame donors

Education

Maternal Health

Target	Ensure that, by 2015, children everywhere, boys and girls alike, will be able to complete a full course of primary schooling	Reduce by three quarters, between 1990 and 2015, the maternal mortality ratio
	Net enrolment ratio in primary education	Maternal mortality ratio
Indicators	Proportion of pupils starting grade 1 who reach grade 5	Proportion of births attended by skilled health personnel
	Literacy rate of 15-24 year-olds	

Lame donors

Proportion attended

	Education	Maternal Health
Target	Ensure that, by 2015, children everywhere, boys and girls alike, will be able to complete a full course of primary schooling	Reduce by three quarters, between 1990 and 2015, the maternal mortality ratio
	Net enrolment ratio in primary education	Maternal mortality ratio
Indicators	Proportion of pupils starting grade 1 who reach grade 5	Proportion of births attended by skilled health personnel
	Literacy rate of 15-24 year-olds	

Lame donors

Education

Maternal Health

Target	Ensure that, by 2015, children everywhere, boys and girls alike, will be able to complete a full course of primary schooling	Reduce by three quarters , between 1990 and 2015, the maternal mortality ratio
Indicators	Net enrolment ratio in primary education	Maternal mortality ratio
	Proportion of pupils starting grade 1 who reach grade 5	Proportion of births attended by skilled health personnel
	Literacy rate of 15-24 year-olds	

Actual outcome



Lame donors

Education

Infant mortality

Target	Ensure that, by 2015, children everywhere, boys and girls alike, will be able to complete a full course of primary schooling	Reduce by two thirds, between 1990 and 2015, the under-five mortality rate
Indicators	Net enrolment ratio in primary education	Under-five mortality rate
	Proportion of pupils starting grade 1 who reach grade 5	Infant mortality rate
	Literacy rate of 15-24 year-olds	Proportion of 1 year-old children immunized against measles

- So, some of us feel strong need for index of quality
- Some (Filmer, Hassan, Pritchett) have called for PISA or later
- Something to be said for that
- But very late for improving
- Earlier in grade structure: can do something
- But what? Reading obvious skill.
- But how?
- Oral: doable before children can do paper and pencil
- Power: so obvious



Need to click out to get fluent and non-fluent readers.

John had a little dog. The little dog was fat. One day John and the dog went out to play. The little dog got lost. But after a while the dog came back. John took the dog home. When they got home John gave the dog a big bone. The little dog was happy so he slept. John also went to sleep.

- So, something powerful and useful about oral reading
- But also:
 - Sequential, therefore can be time-consuming
 - Assessing a real performance task raises issues of inter-rater reliability
- If use is policy awareness: does not matter much
- But moving to intervention and evaluation: it does matter
- What are some solutions?
- Also, what other things are we finding out?

Basis of experience

- USAID, World Bank: paid for “codifying” an approach to oral testing being called EGRA (Early Grade Reading Assessment)
- Have developed “toolkit”
 - Includes assessment components, why each component, training guides, sample size guides, etc.
- Doing pilot experiences, case studies in:
 - The Gambia, Senegal, Haiti, Honduras, Guyana, Kenya, Liberia, Nicaragua, earlier Peru, re-doing Peru, etc.
- Others on their own, but related:
 - South Africa, Afghanistan, Bangladesh, Mali, Niger
- Some 20 languages?

Questions related to “large-scale”

- What is inherently large?
- Does eval of large scale programs need to be large scale?
 - Take ORF in connected text as a marker variable (measured in cwpm) (can discuss why)
 - Most between-grade differences are around 14
 - We now know most positive “associated factors”, even singly, are associated with an inter-grade diff of about 10
 - We’d therefore hope that a successful purposeful campaign should improve performance by 10-15 cwpm, ideally much more



some results



Some results

- Also know Std Dev around 20-30, but is lower in the lower grades (hence lower with the mean)
- Thus, CIs similar to the following are observed:

(Liberia)	Grade 2	Grade 3
Sample size (47 schools, 10 kids/grade)	429	407
Mean ORF-connected text (in correct words per minute)	17.7	27.8
<u>Assuming simple random sampling</u>		
Std Dev	18.7	21.9
Std Err	0.9	1.1
Lower bound 95% CI	15.9	25.6
Upper bound 95% CI	19.4	29.9
<u>Assuming children are clustered into schools</u>		
Std Dev	33.8	41.4
Std Err	1.6	2.1
Lower bound 95% CI	14.4	23.6
Upper bound 95% CI	20.9	31.9

- So maybe eval of large scale programs need not be that large?
- One problem posed by intervention programs: clustering and design effect (discuss?)
- For awareness and policy dialogue uses, variance does not matter too much
 - In evaluation, it does, since we want to test hypotheses
 - So need to worry about design effect
 - Especially if there is clustering of schools for reasons of cost-saving in implementation: double clustering (clusters of schools, schools within clusters)

- Other issues of reliability, internal correlations, etc., are not directly relevant to “large” but larger – stakes are higher, so want to make sure characteristics are good

In fact, in the end, there is only one issue that I think creates inherent “largeness” problems...

What if assessment **is** the intervention?

- Localized accountability model
- E.g., current USAID experiment in Liberia:
 1. Control schools
 2. Treatment one: measure, inform, **only**
 3. Treatment two: measure, inform, PLUS teacher training in reading, reading kits, reading books, supervision and support
- In Treatment one, measurement is the intervention
- In Treatment two, measurement is still integral

What if assessment is the intervention?

- Need to do all schools, should do with all kids, not sample, otherwise no real accountability
- Could have school-level accountability, but a sample of 20-30 does not characterize a school anyway...
- So here there is a dilemma with scale
- Doing with outsiders impossibly expensive
- Therefore only conclusion is to do with teachers themselves
- Or unpaid volunteers
- Issue of inter-rater reliability
- Using paid workers we have been able to get beyond 95%
- What about with untrained volunteers? Teachers? Who knows?
- Option with teachers: random audit (both to audit and to improve)

One promising technique not used much in education is Lot Quality Assurance Sampling

Allows use of samples as low as 20, but it is not a localized parental accountability model

And only if result is binary (can read, cannot read) and if you do not care about degree of deviation from non-compliance, only if a school or unit is “compliant” (kids reading) or not

Basic idea: easier to tell a coin is not fair than to tell how biased it is... Can use smaller samples.

But all those are the “nerdy” issues

Issues of political economy and mass psych are also interesting:

- What is the tipping point?
- Does scale matter in getting gov't to “do the right thing”
- Do we know? This would be evaluation in the service of determining scale, not so much evaluation of large scale, but worth thinking about