From Impact Evaluation to Policy Design:

Evidence, External Validity, and Effective Policymaking

Martin J. Williams

Blavatnik School of Government, University of Oxford

ENAP, 4-8 June, 2018

Reducing Teenage Pregnancy in Australia

Teenage pregnancy creates social and economic costs for the mother and for society

Australia has one of the highest teen pregnancy rates in the OECD

Policymakers in Western Australia implemented a program that was in use in 89 other countries...

Infant simulators

Electronic babies that

cry for food, changing, burping

Given to girls in public

schools (age 13-15) for one week

Administered by a school nurse

Expensive: AUS$1200 each

What Happened?

The program actually *doubled* pregnancy rates:

- 8% of girls in the intervention programme had at least one birth before age 20
- compared to 4% of girls in a control group

Widely used program (89 countries!) turned out to be ineffective

Brinkman, Sally, Sarah E Johnson, James P Codde, Michael B Hart, Judith A Straton, Murthy N Mittinty, and Sven R Silburn. 2016, "Efficacy of infant simulator programmes to prevent teenage pregnancy: a school-based cluster randomised controlled trial in Western Australia." The Lancet 388: 2264 - 71. Photo: http://www.realityworks.com/products/realcare-baby

The Importance of Evidence

Evidence matters because many policies do not achieve the effects we expect

- Even when many people expect them to work
- Even when they are widely used

Our instincts are not perfect, so we need to:

- Base our policies on evidence
- Evaluate them to see if they are working

"Science is what we do to keep from lying to ourselves."

- Richard Feynman

Test Your Policy Instincts

Maybe you think baby dolls were obviously a bad idea

So let's test some other policies

For each policy, write down whether you think it worked

- Yes
- No
- Mixed

The Policies

1. Getting Teachers to Come to School

The problem: Recent improvements in primary school access have not been accompanied by improvements in school quality. Poor learning outcomes may be due, in part, to teacher absenteeism: a nationally representative survey found that 24 percent of teachers in India were absent during normal school hours.

The policy: Teachers received a camera and had to ask a student to take a photograph of the teacher and the other students at the start and end of each day. The cameras had a tamper-proof system that made it possible to precisely track each school's openings and closings. Teachers received a Rs 50 ($1.15) bonus for day they attended in excess of the minimum required.

Outcome of interest: Teacher absenteeism and student test scores.

Where: rural schools in Rajasthan, India.

Did it work? Yes? No? Mixed?

2. Getting Nurses to Come to Health Clinics

The problem: Public health care provision in India is plagued by high staff absence, low effort by providers, and limited use by potential beneficiaries who prefer private alternatives. Public health facilities are often closed (about 56% of the time) due to high absenteeism rates.

The policy: The same monitoring/incentive program as the one implemented in schools.

Outcome of interest: Nurse absenteeism.

Where: Health clinics in Rajasthan, India.

Did it work? Yes? No? Mixed?

3. Job placement assistance

The problem: Long-term unemployment among young people

The policy: Government contracts private companies to provide job search help to young graduates who have been unemployed for six months.

Outcome of interest: Whether a worker has a contract with a length of more than six months, after eight months of search.

Where: 235 cities in France.

Did it work? Yes? No? Mixed?

4. Early childhood education and service provision

The problem: Low-income children face a variety of disadvantages that affect their life opportunities and impose costs on society.

The policy: The Head Start programme provides comprehensive educational, medical, and nutritional support to 3- and 4-year olds from low-income families.

Outcome of interest: Various educational, medical, and nutritional indicators.

Where: USA

Did it work? Yes? No? Mixed?

5. Moving to Opportunity

The problem: Children who grow up in high-poverty neighborhoods have worse socio-economic and health indicators as adults, even controlling for individual characteristics.

The policy: Moving to Opportunity provided housing vouchers to a randomly selected group of poor families, with the requirement that the families move to a low-poverty neighborhood.

Outcome of interest: Employment and earnings of the children as adults.

Where: Baltimore, Boston, Chicago, Los Angeles, New York City

Did it work? Yes? No? Mixed?

6. Eyeglasses and Academic Performance

The problem: Many primary school students in developing countries have vision problems. In almost all cases their vision can be corrected with properly fitted eyeglasses, but very few of them have eyeglasses.

The policy: A randomized trial that offered free eyeglasses to children in grades 4, 5 and 6.

Outcome of interest: Test scores

Where: Rural primary schools in China

Did it work? Yes? No? Mixed?

The limits of evidence-based policymaking

Tamil Nadu Integrated Nutrition Project (TINP)

- Nutritional counselling for mothers + supplementary food for infants
- Rural areas of poor districts; 9 million people
- Improved nutritional knowledge in mothers, significant decline in child malnutrition

Bangladesh Integrated Nutrition Project (BINP)

- Program design copied from TINP; covered $1/8^{th}$ of Bangladesh
- Improved nutritional knowledge in mothers
- No impact on child malnutrition

What happened?

In rural Tamil Nadu, women make household food decisions

- Improved nutritional knowledge in mothers + additional food $\rightarrow$ mothers and children eat more food

In rural Bangladesh, men and mothers-in-law make household food decisions

- Improved nutritional knowledge in mothers + additional food $\rightarrow$ mothers know more, but additional food for infants is treated as a substitute rather than a supplement

Example from Cartwright and Hardie, (2012), *Evidence-Based Policy: A Practical Guide to Doing it Better*, OUP.

External validity

This is the problem of external validity of evaluation results

- Impact evaluation asks: "did it work there?"
- What we need to know for policymaking is :"will it work here?"

Learning how to think about this question is our main goal for this week

Framing questions from Cartwright and Hardie, (2012), *Evidence-Based Policy: A Practical Guide to Doing it Better*, OUP.

Goals of this Course

1. Review impact evaluation methods, critical appraisal, sources of evidence

   - Did the policy work?

2. Understand why the same policy can have different effects in different contexts

   - Will it work the same here as it did there?

3. Think about how to adapt policies to make them suit your context

   - How can we make it work better here?

Structure of the Course

- Monday: review of quantitative impact evaluation methods
  - How to evaluate a particular policy in one or more contexts
- Thursday: external validity and mechanism mapping
  - How to predict whether a policy will have the same effect in this context as it did elsewhere
- Friday: policy adaptation and group presentations
  - How to optimize a policy for a particular context

Small Group Projects

The best way to learn new skills is to apply them!

Each person will join a group based on policy area interest, and:

- Assess the evidence about that policy's effectiveness in other contexts
- Predict what the effects of the policy would be in Brazil
- Propose adaptations to the policy to make it better suited to Brazil

Each group will present their analysis and recommendations on Friday

Quantitative Methods

Some people are more knowledgeable about quantitative methods than others

This course focuses on teaching **intuition** and **interpretation**, not on technicalities of doing an impact evaluation

Aim is to help you be able to read a study, understand its findings, and have a sense of the reliability of these findings

If you are very familiar with these methods, please use your expertise to help your colleagues understand

- Monday may be review for some of you

Course expectations

This course relies on your sharing your experience and policy expertise

- You are the experts on your own context

If you do not usually speak much, challenge yourself to speak up and share your experience and questions

- If you are someone who speaks a lot, make sure others also have space to speak

There is no homework, but please use in-class working time well

- Otherwise Friday will be very boring!

Any others?

Questions?

Next step: reviewing quantitative methods for impact evaluation

Outline

1. The fundamental problem of evaluation: **counterfactuals**
2. Experimental evaluation methods: **randomized controlled trials (RCTs)**
3. Observational evaluation methods: **difference-in-difference**

Counterfactuals

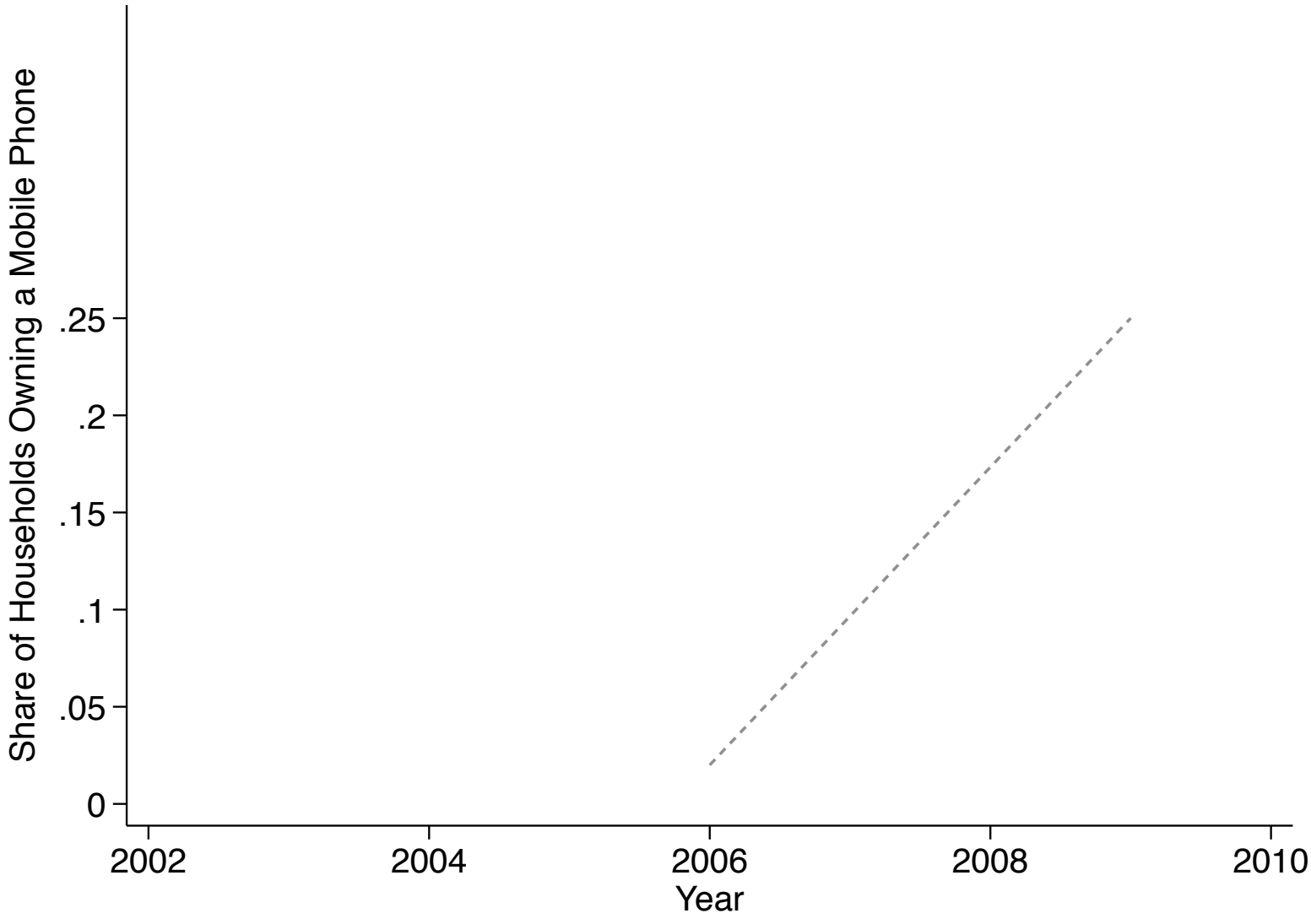
Impact of the Millennium Villages Project

The Millennium Villages Project (MVP) is an anti-poverty intervention in villages across Africa, championed by Jeffrey Sachs

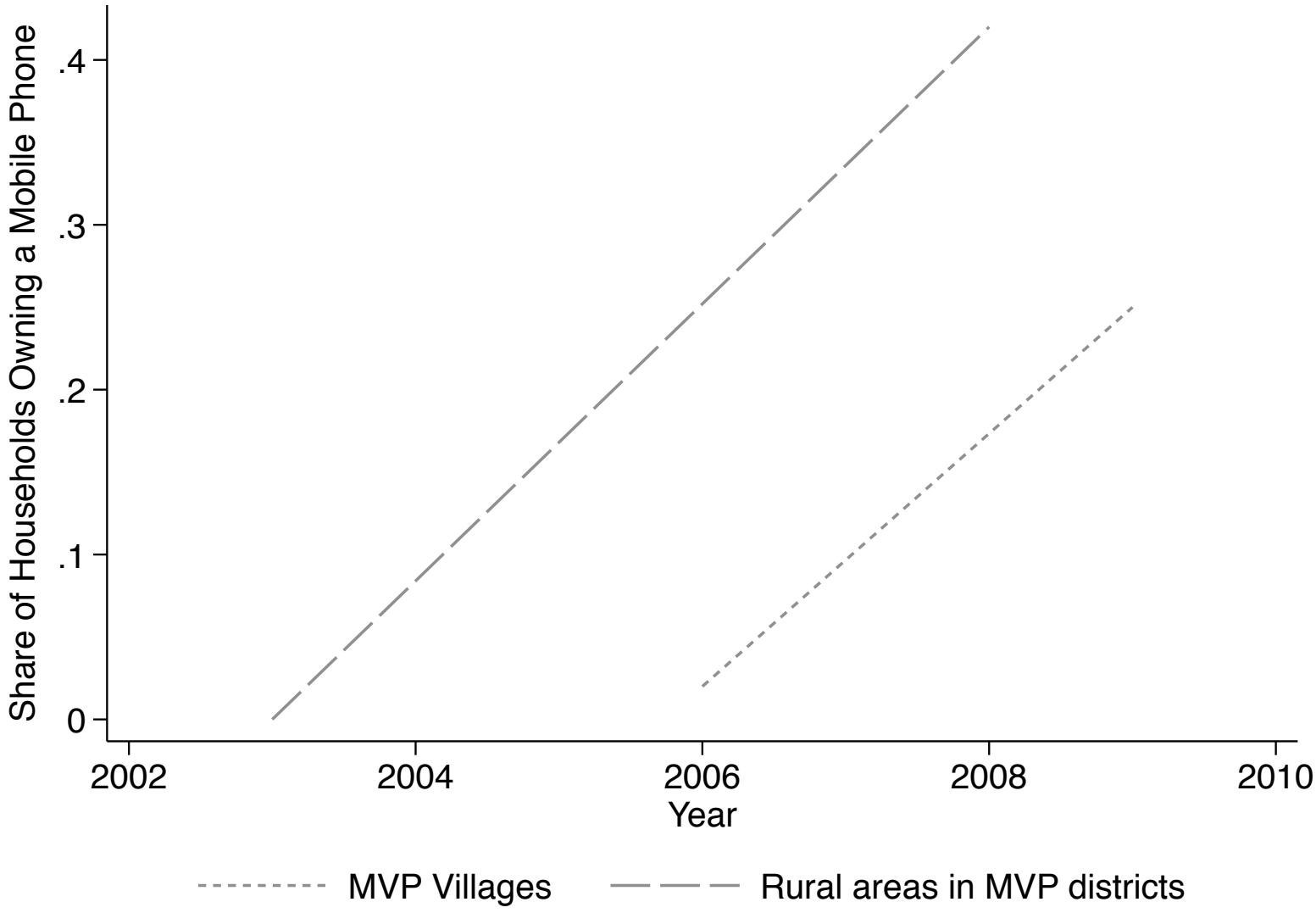
A key finding from their June 2010 report *Harvests of Development* is that the project increased ownership of mobile phone from 4% to 30% from 2006 to 2009

That seems like a huge impact!

Phone Ownership Increased in MVP villages. . .



. . . but increased the same amount in non-MVP villages



Impact Evaluation Answers Causal Questions

Many policy questions are causal questions:

- Do baby dolls reduce teen pregnancy?
- Will reducing class sizes improve educational outcomes?
- Does incarceration reduce subsequent criminality?
- Do privatised industries out-perform nationalised ones?
- Do scholarships improve academic performance?

The fundamental problem of evaluation

This is the fundamental problem of evaluation:

- How do we measure a policy's impact on an individual, if we do not observe what would have happened to that individual without the policy?

A **counterfactual** is what would have happened to an individual in the absence of a policy

This is often referred to as a **control group** which can be compared against a **treatment group**

Example 1: Class size

What is the effect of class sizes on educational outcomes?

E.g. suppose we observe the educational outcomes for children taught in large classes

The counterfactual: how would they have done in smaller classes?

Example 2: Incarceration and criminality

What is the effect of incarceration on subsequent criminality?

E.g. suppose we observe the criminal records of convicted criminals post-incarceration

The counterfactual: how would they have done had they not been incarcerated?

Example 3: Privatisation and Economic Performance

Do privatised industries out-perform nationalised ones?

E.g. suppose we observe the productivity of privatised industries.

The counterfactual: how would they have done had they remained nationalised?

The Challenge of Counterfactuals

But we never actually observe the counterfactual - that would be an alternate reality!

What we actually observe is usually:

- **Before-after comparison**: the level of a key outcome for the same individual, before and after "treatment"

- **Treated-untreated comparison**: the level of a key outcome for different individuals, after a policy

But before-after comparisons can be misleading, as with the MVP villages

Treated-untreated comparisons can also be misleading. . .

Treated-untreated comparison: Do hospitals make people sick?

The US National Health Interview Survey (NHIS) contains the questions:

1. "During the past 12 months, was the respondent a patient in a hospital overnight?"
2. "Would you say your health in general is excellent, very good, good, fair, poor?" [5 = excellent health, 4, 3, 2, 1= poor health]

Treatment: hospitalisation.

Outcome: health status

Angrist and Pischke. 2009. "Mostly Harmless Econometrics"

Treated-untreated comparison: Do hospitals make people sick?

Group	Size	Mean health status
Hospital	7,774	2.21
No Hospital	90,049	2.93

The difference in mean health status is: -0.72.

So. . . does going to the hospital make people sicker?

What's the issue?

Ways to Estimate Counterfactuals

To estimate the counterfactual for a policy, we must find a **control group**

This control group should be as similar as possible to the **treatment group** (those that were directly affected by the policy)...

- ...but NOT have been affected by the policy

There are several ways to use control groups to estimate the causal effects of a policy. We will review:

- Randomized controlled trials (RCTs)
- Difference-in-difference estimation

Randomized controlled trials (RCTs)

The objective and the problem

We're interested in the causal effect of a given policy

We want to measure the changes in the well-being of individuals that can be attributed to a particular policy

But we can't just:

- measure the difference between treated and untreated individuals
  - because treated and untreated individuals may be systematically different
- measure the difference in treated individuals before vs after
  - because the change may not be attributable to the policy

Random assignment

What if we randomly selected who got the treatment?

- People who were not selected would then be a control group

This solves both problems:

- There are unlikely to be systematic differences between groups
- Any population-wide trends would affect both groups equally

In this case:

1. Control group outcomes are a valid counterfactual
2. Comparing the outcomes of the treatment and control groups will deliver an estimate of the causal effect

An Example

Progresa was a conditional cash transfer (CCT) in Mexico. It was randomly rolled out across communities. One outcome was to increase utilization of health services (visits to clinics)

- Before the program, treatment and control communities had the same number of consultations per day

After the program, treatment communities had an average of 12.8 consultations per day

- Control communities had an average of 11.5 consultations per day
- The causal **treatment effect** is 12.8 - 11.5 = 1.3 additional consultations per day

Source: Gertler and Boyce (2001)

Public Financial Management Reform in India

Payment approval in multi-level bureaucracies can be slow

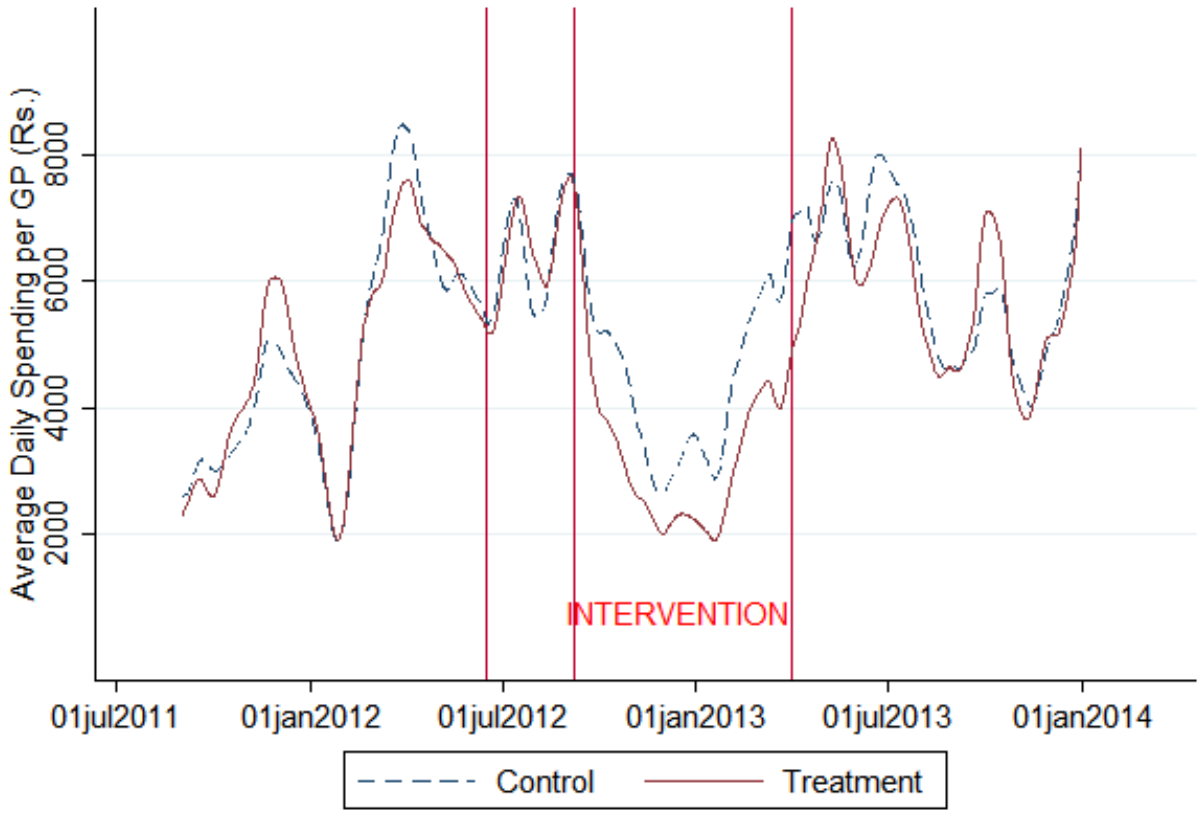
- Alternative: electronic platforms to directly link lowest level of government to fund disbursement from state government
- Can these improve performance and reduce leakage?

Researchers worked with Bihar state in India to randomly implement electronic platform in 12 randomly selected districts

E-transfer reduced leakage (ghost workers) and corruption (officials' wealth declarations), and improved some measures of efficiency

Impact of e-transfer system

Figure 3: GP daily Expenditures on MGNREGS during the Study Period



Source: CPSMS data on GP savings accounts.

RCTs in Practice

1. Define the population that is eligible for the program
2. Randomly assign which units are in the treatment and the control groups
3. Implement the intervention only for units in the treatment group
4. Collect data in both groups and compare the outcomes of treated and control groups.

What can go wrong?

Some problems that can make RCTs invalid

1. Contamination/Spillovers
2. Non-compliance
3. Small samples

Threat 1: Contamination/Spillovers

People in the control group benefit from the treatment too.

This can happen for two reasons:

1. They access the treatment directly.
2. They benefit from the treatment group being treated.

Detecting the presence of spillovers: the benefits of deworming

Intestinal worms are believed to have a negative impact on education

Two kinds of RCTs were implemented:

- Randomly select some students within each school to deworm
  - Compare treated/untreated kids in the same schools -> small effect
- Randomly select some schools, and deworm all students within the school
  - Compare treated/untreated kids in different schools -> large effects

Why are the results different?

Treatment reduces the transmission of infections to other kids

Miguel and Kremer (2004), "Worms"

Threat 2: Non-compliance

Individuals who are offered a treatment refuse to take it. In other words, some of the treated turn out to be untreated.

Microfinance in Ghana

Opportunity International Savings and Loans, Ltd. designed a door-to-door marketing campaign with micro-entrepreneurs.

Out of every 100 business owners that were targeted:

- 15 visited a branch
- 5 started an application
- 2 completed it
- 1 took a loan...

Karlan and Appel (2016) Failing in the Field. Chapter 7.

Threat 3: Small samples

If the number of units participating in the trial is small, then an effect may appear just by statistical chance

- With small numbers, randomization might still create unbalanced groups
- A seemingly large effect may be driven by just a few units

What is too small? It depends. Some VERY rough guidelines:

- Ignore any study with less than 30 units
- Be very skeptical of any study with less than 200 units
- We'll talk tomorrow about the statistical uncertainty associated with studies

Opportunities for RCTs in Policy

What can be randomized:

1. Access: Which people will be offered access to the program.
2. Timing: When to provide access to the program.
3. Encouragement: Which people will be given encouragement to participate in the program.

When the program is:

1. being pilot-tested
2. adding new services
3. adding new people/locations
4. over/under-subscribed

Source: Glennerster and Takavarasha (2013)

Recap on RCTs

RCTs:

- make it possible to estimate a counterfactual outcome for the policy
- allow us to compare outcomes of treatment versus control group
- are subject to a number of threats: contamination/spillovers, non-compliance, small samples

Questions?

But what about when randomization is not possible?

Difference-in-Difference Estimation (DiD)

Difference-in-Difference

Recall that we needed RCTs because simple treated-untreated or before-after comparisons might not give us the true causal effect of the policy, because:

- There may be systematic differences between treatment and control groups
- There might be general, population-wide trends that affect both groups

The idea behind DiD is that we can combine a treated-untreated comparison with a before-after comparison to solve these problems

- Instead of comparing treatment and control group outcomes after the policy has been implemented)...
- ...we compare treatment and control before AND after implementation...
- ...and look at treatment group change compared to control group change

An example

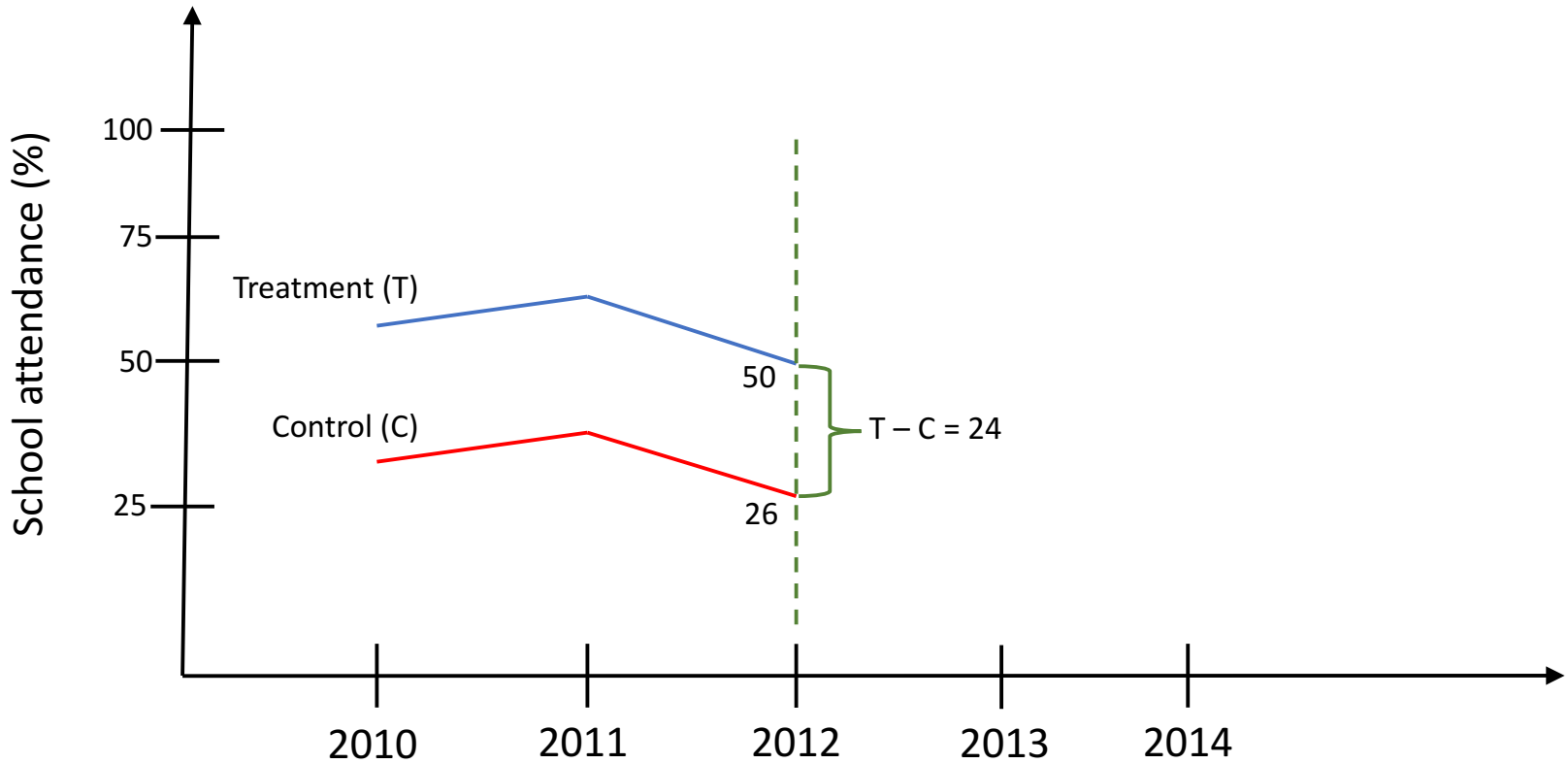
Hypothetical policy: do conditional cash transfers raise school attendance?

Suppose CCTs were implemented in 2012 in some municipalities, but not implemented in others

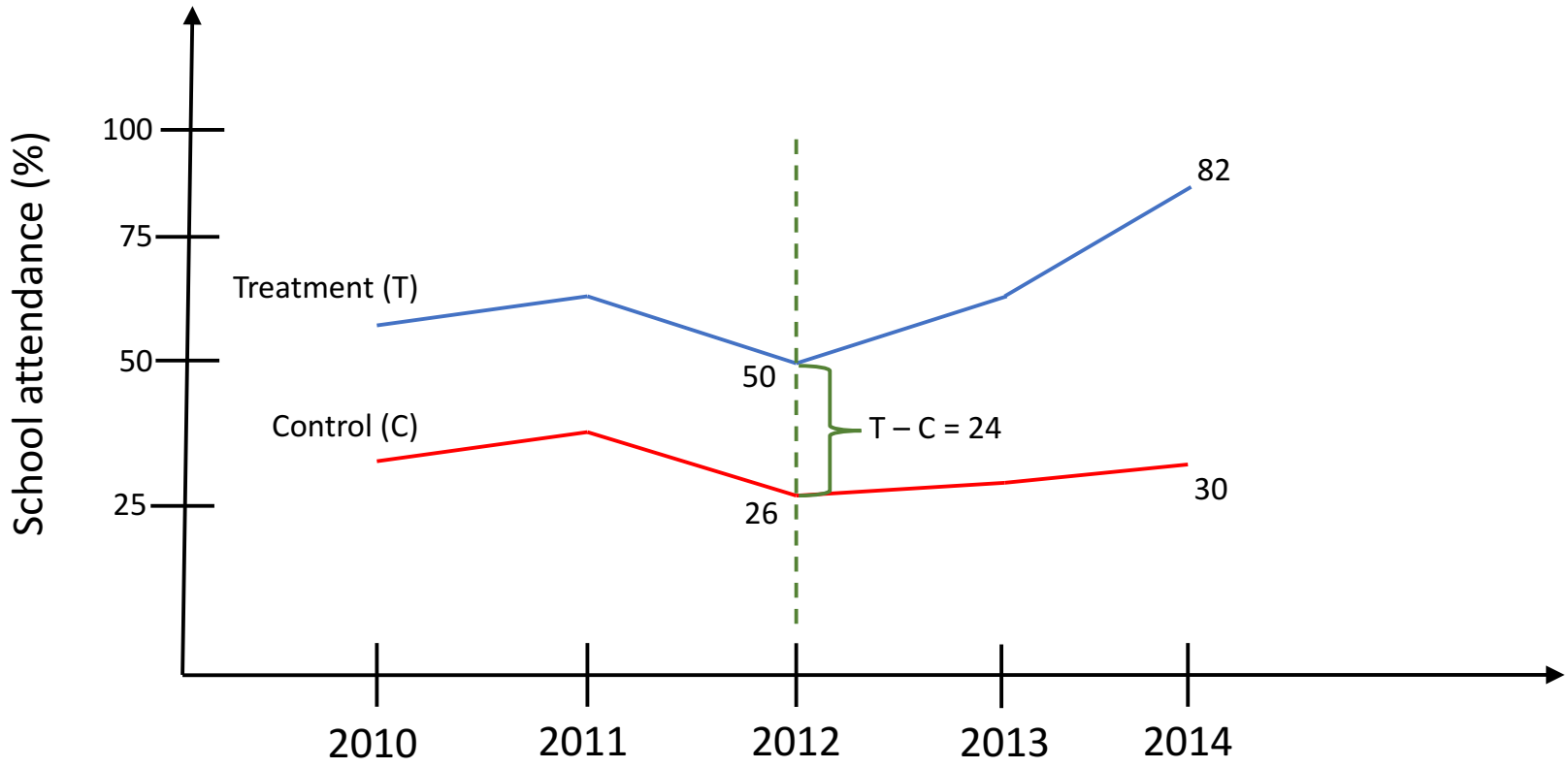
- But the municipalities were NOT randomly selected

Let's use difference-in-difference to estimate the causal impact of the policy

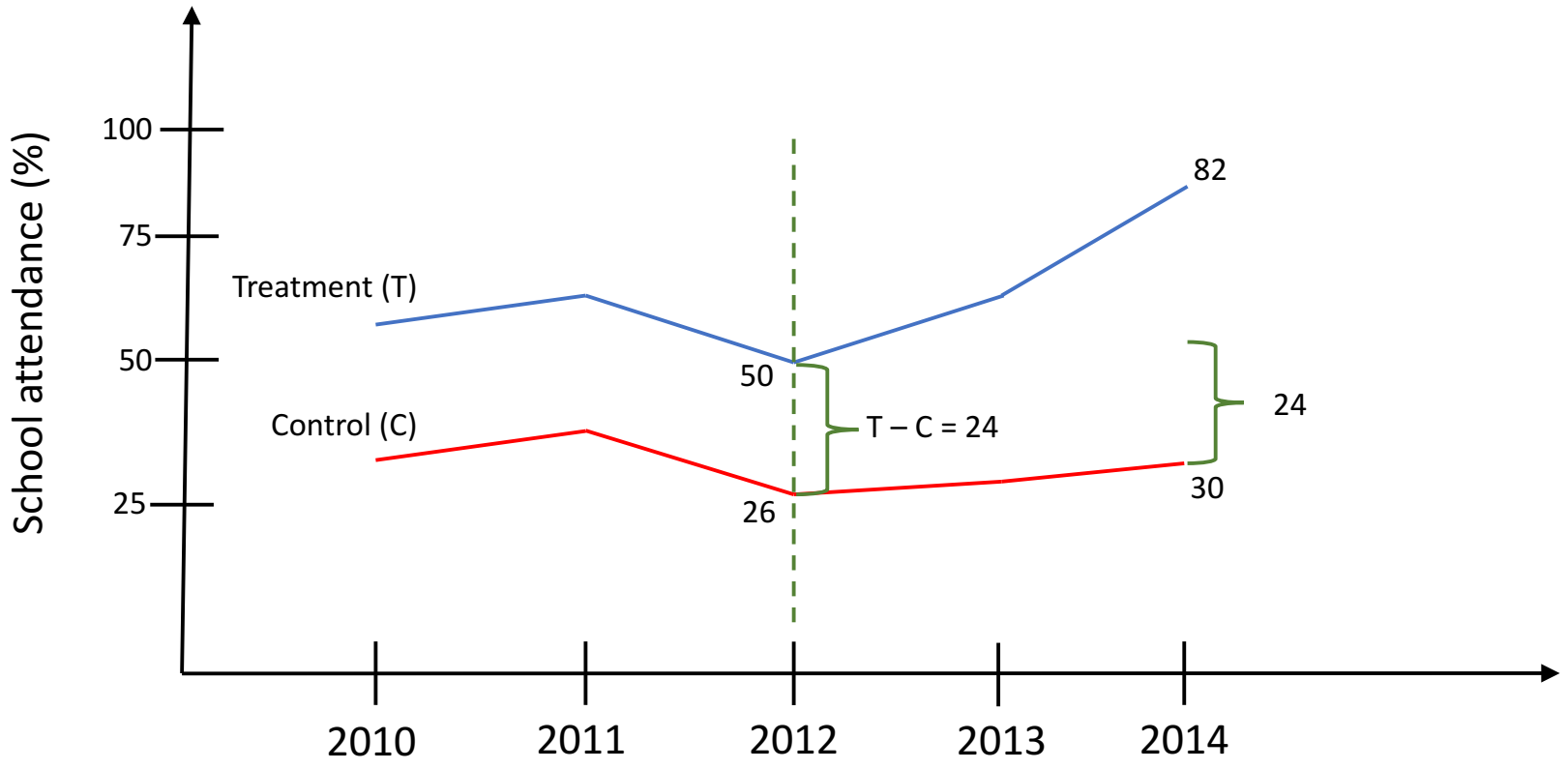
Before implementation, treatment is 24 higher than control



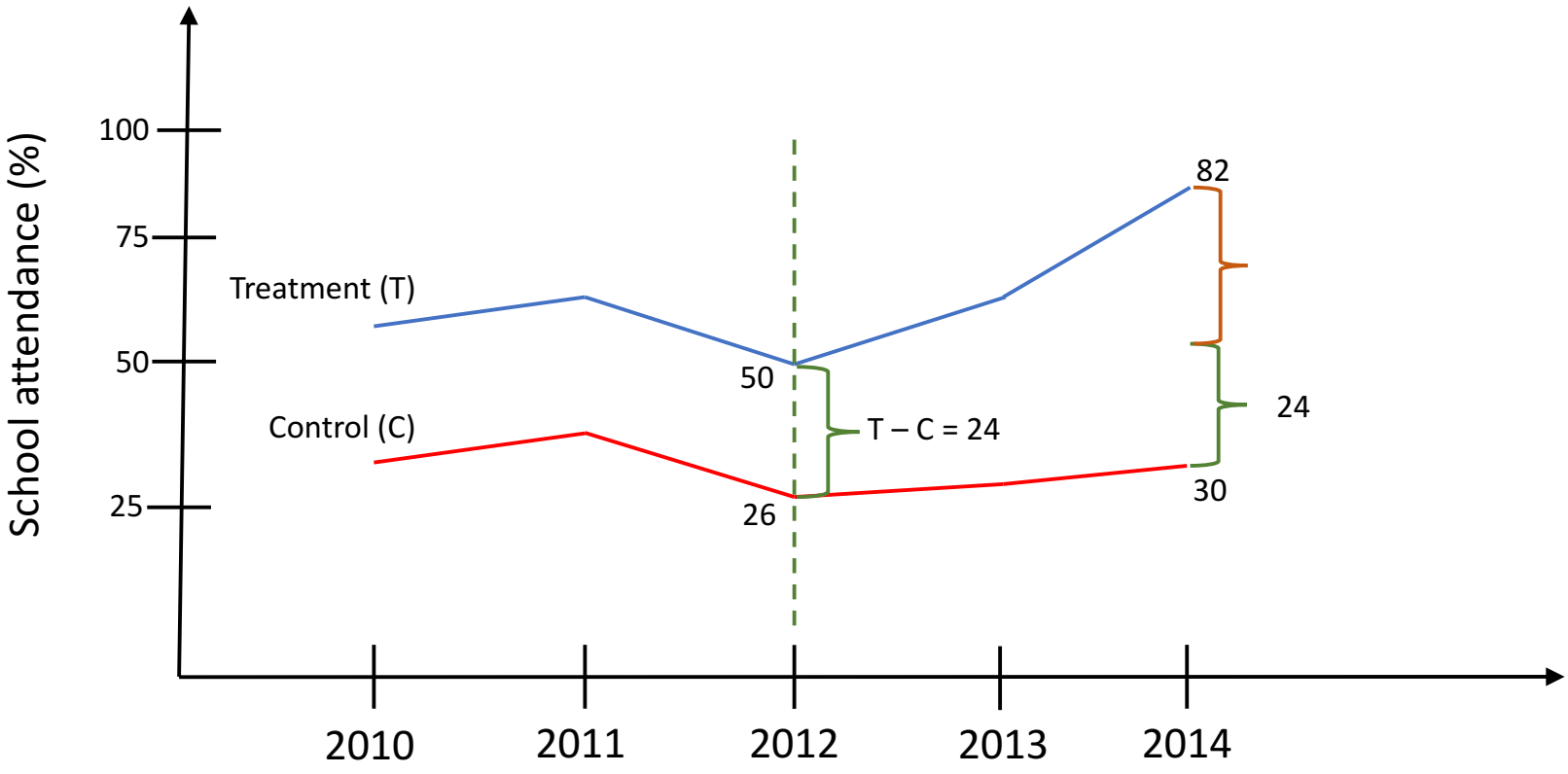
After implementation, treatment is 50 higher than control



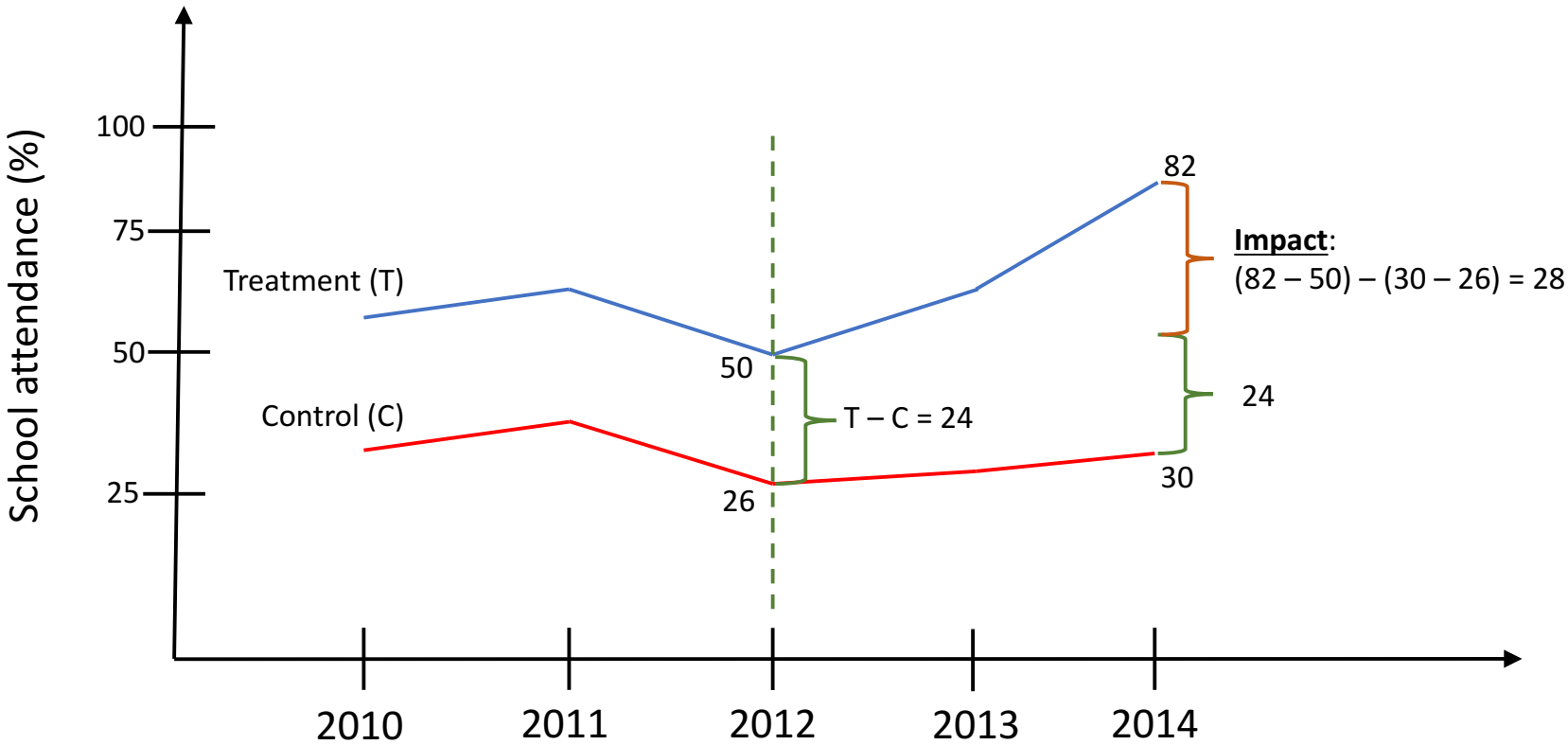
Baseline difference in outcomes is 24 percentage points



What is causal impact of the policy?



Compare CHANGES between treatment and control



When can we use difference-in-difference?

DiD assumes the only thing that could have affected the outcome in the treatment group (but not the control group) was the policy

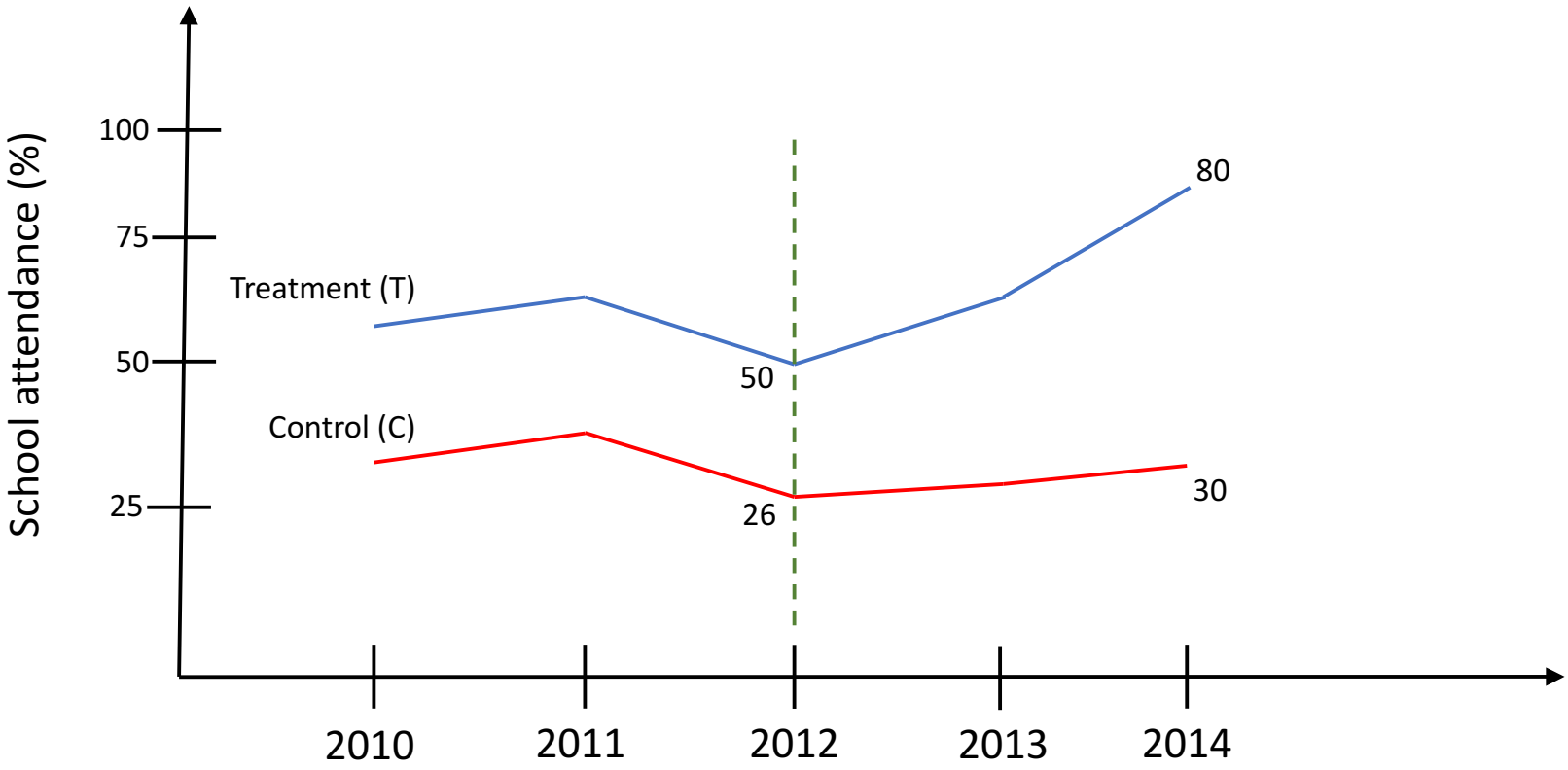
But what if there were other differences?

- e.g. the CCT was implemented by municipalities that were also doing other things to improve school attendance

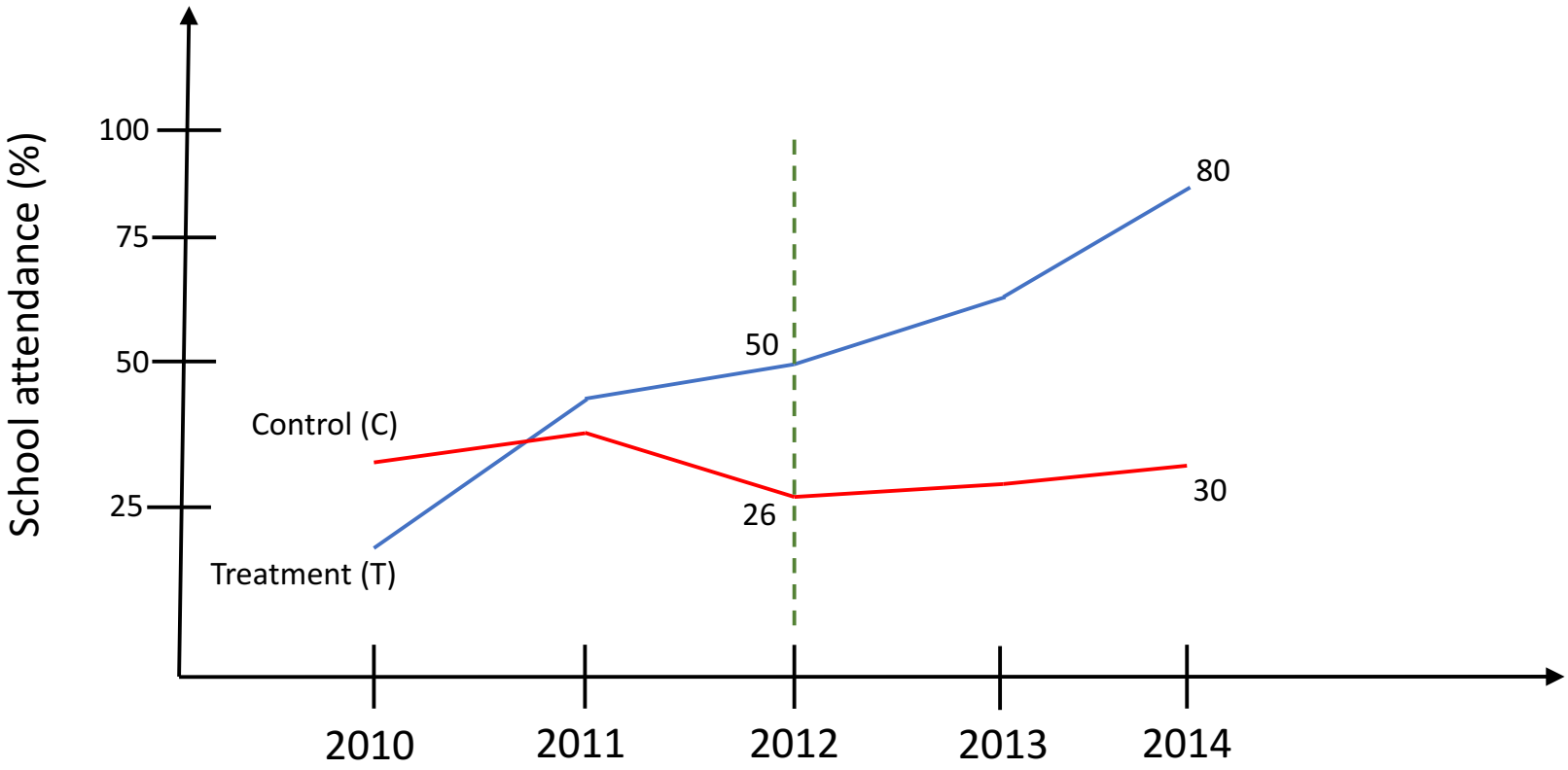
To test this, examine whether the trends in school attendance were the same before the policy

- **Parallel trend** assumption
- If not, then may be other differences between treatment and control groups that could be driving school attendance

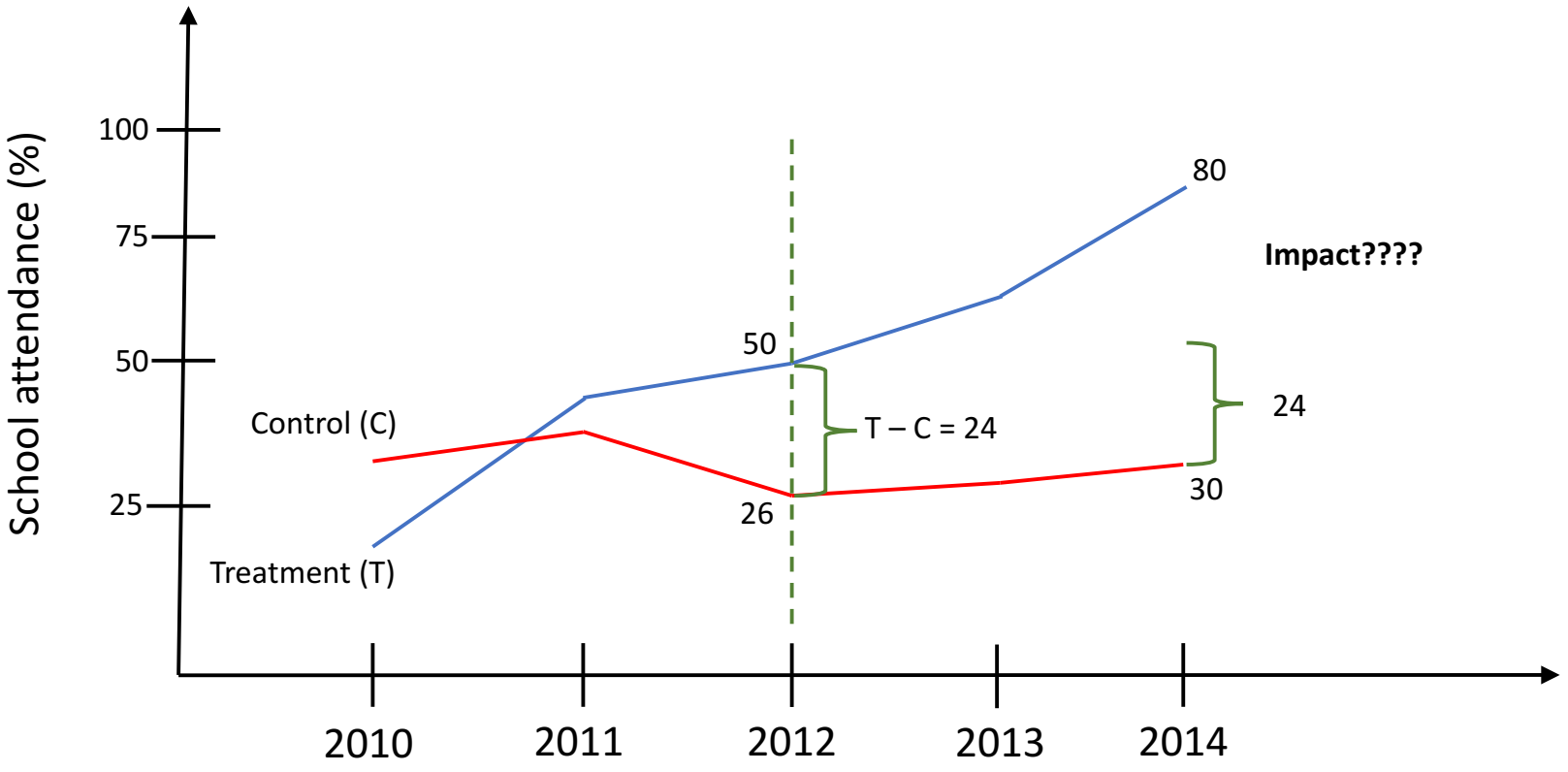
Trends prior to implementation are parallel



Trends prior to implementation are NOT parallel



Without parallel trends, can't identify causal impact



Another example: "Ban the Box" (BTB) legislation

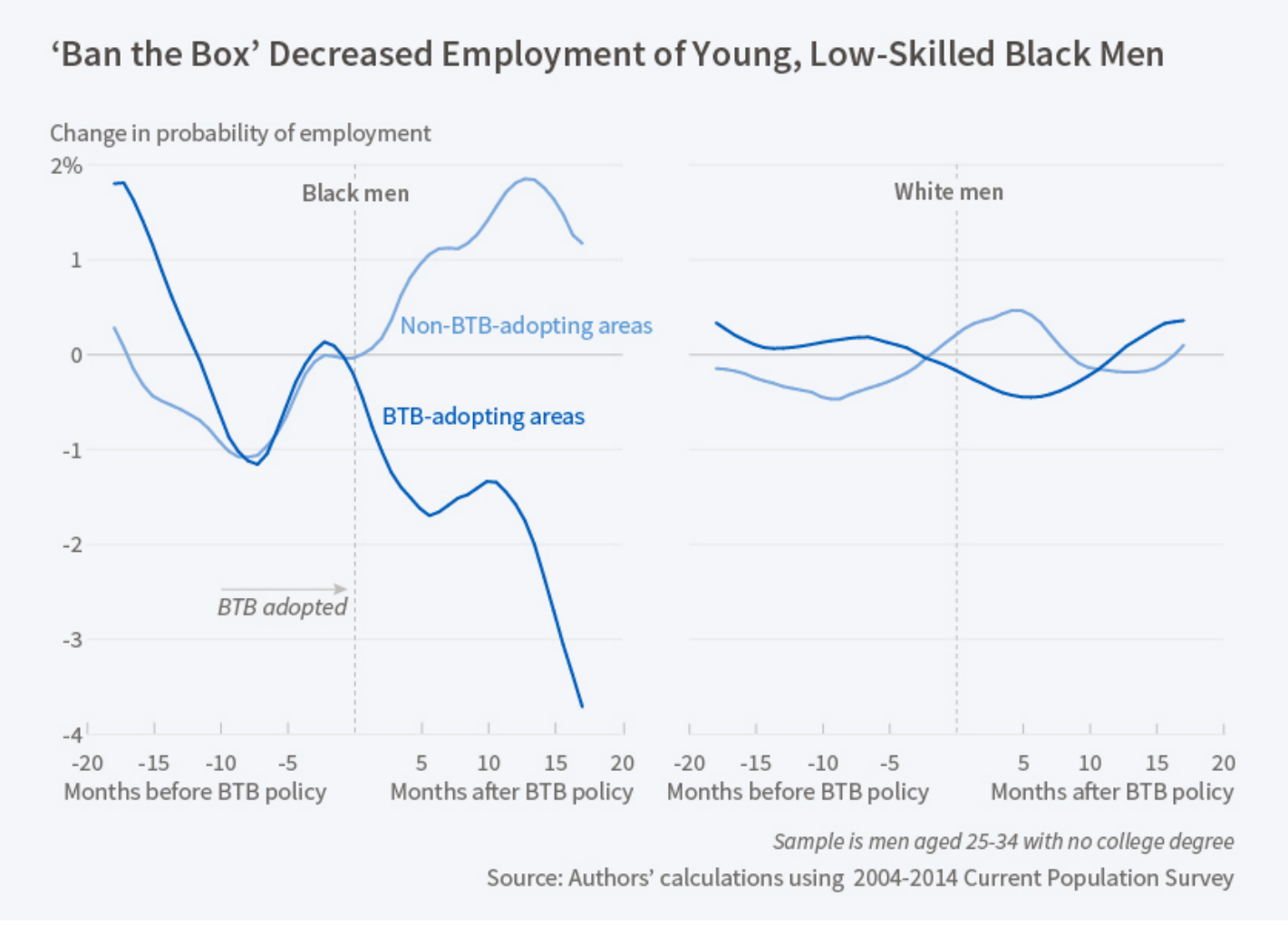
The problem: Ex-convicts often have difficulties finding a job as employers ask about applicants' criminal history.

The policy: A number of US states and districts passed "ban the box" (BTB) legislations that forbid employers from including a criminal record check box on job applications.

The concern: It could lead employers who don't want to hire ex-offenders to try to guess who the ex-offenders are, and avoid interviewing them. Especially a concern for young, low-skilled, black and Hispanic men.

Jennifer L. Doleac, Benjamin Hansen. 2016. Does "Ban the Box" Help or Hurt Low-Skilled Workers? Statistical Discrimination and Employment Outcomes When Criminal Histories are Hidden. NBER Working Paper No. 22469

Parallel trends in "Ban the Box" (BTB) legislation



Public Credit and Firm Growth in Brazil

Does public credit help firms grow and/or export?

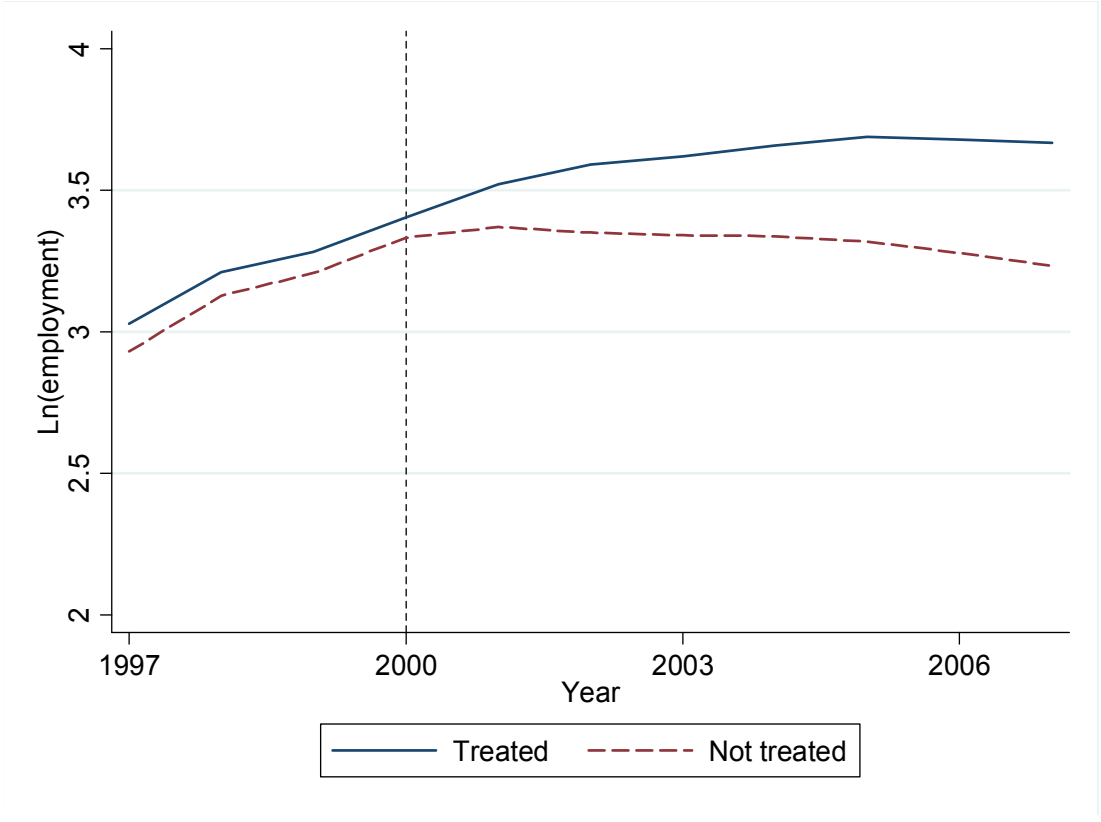
Researchers compared firms that received public loans (BNDES or FINEP) to firms that did not

- Compared treatment and control firms with similar trends in performance prior to loan receipt

De Negri *et al* (2011), "The Impact of Public Credit Programs on Brazilian Firms", IDB

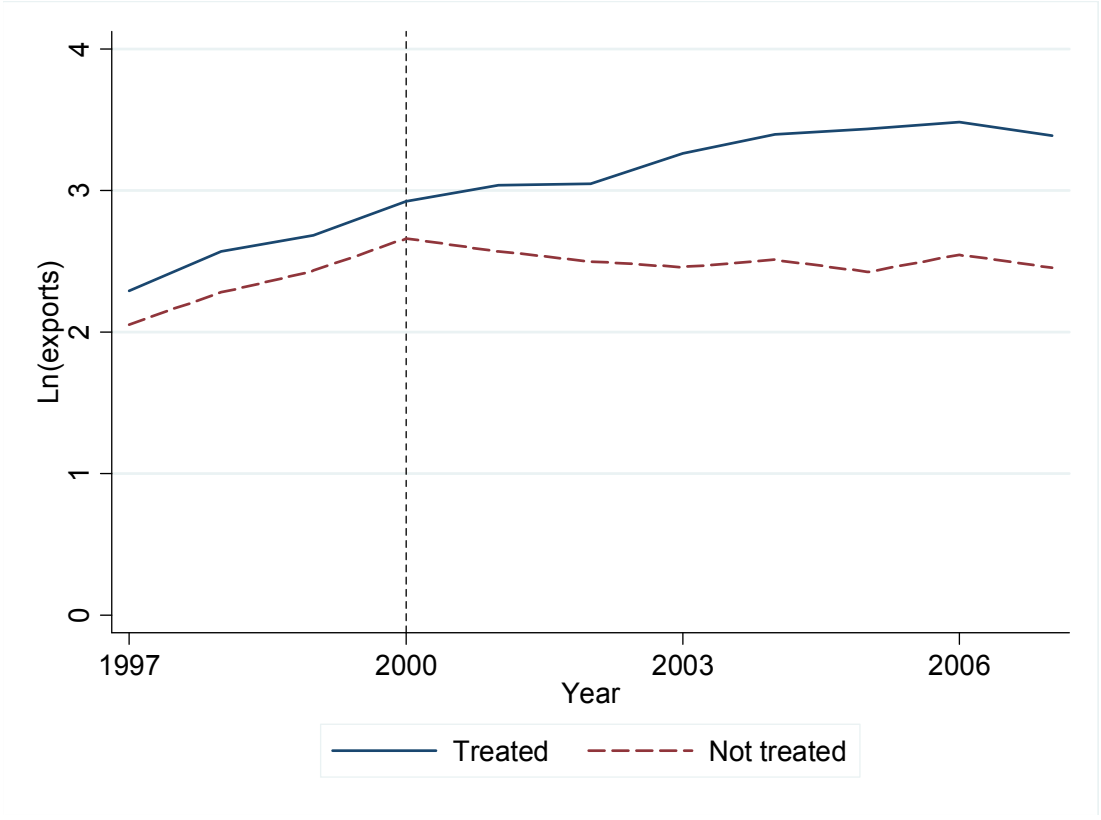
Impacts on Employment

Figure 4. Employment (Matched Sample)



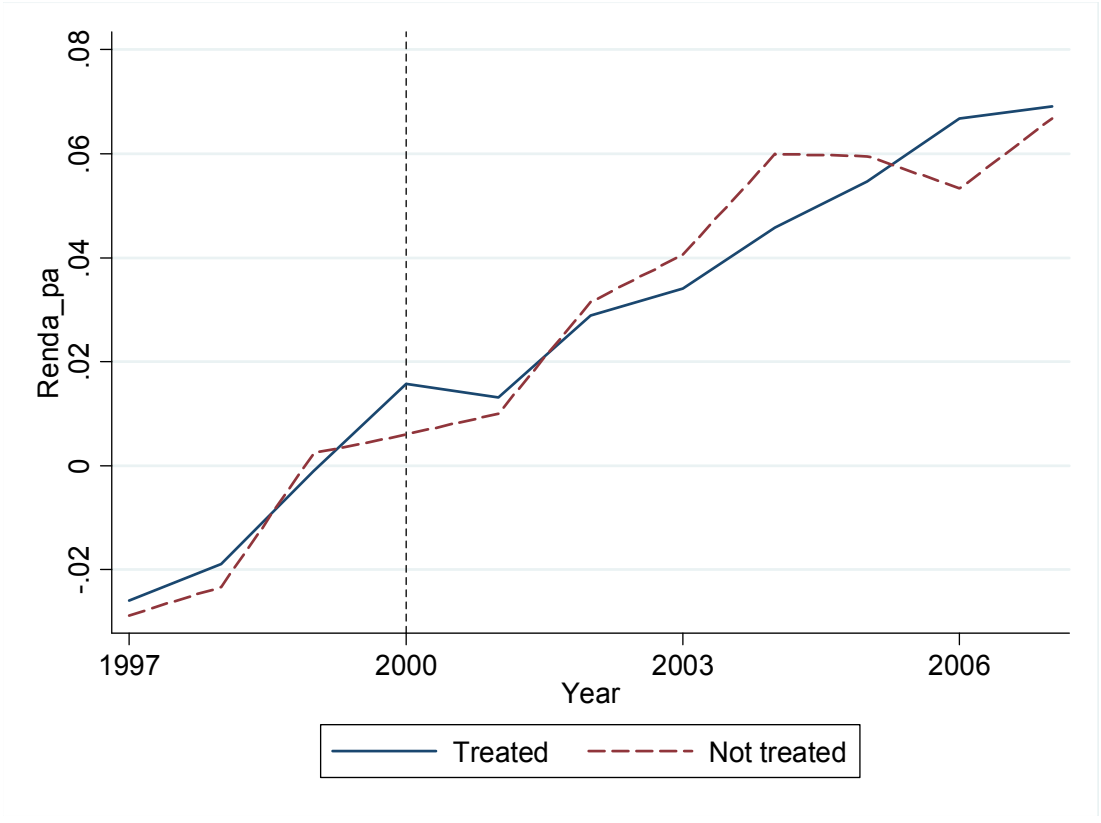
Impacts on Exports

Figure 5. Exports (Matched Sample)



Impacts on Labor Productivity

Figure 6. Labor Productivity (Matched Sample)



Can you think of other examples in Brazil?

What are some policies which could be evaluated by difference-in-difference?

- Policies that have been implemented in some states/municipalities but not others...
- Or policies that have been gradually rolled out over time

Recap: Difference-in-Difference

Difference-in-difference can allow us to identify causal impacts without randomization

- Compare change in treatment group to change in control group

Same potential problems as RCTs:

- Spillovers, non-compliance, small sample sizes

PLUS treatment and control groups must have parallel trends before policy was implemented

Questions?

Small group projects

Summary

10 policies that have been tried and evaluated outside of Brazil

- And 2 from Brazil

Choose one of these 12 based on your interest and form a group

Each group will:

- Read and critique the evaluation report
- Seek out additional evidence about the policy
- Map out the theory of change for the policy
- Consider how the policy would work in the Brazilian context
- Propose adaptations to make the policy more effective in Brazil
- Present this analysis to the whole group on Friday

The Policies (Part 1/2)

1. Health worker recruitment in Zambia
2. Public financial management reform in India
3. Longer school day in Chile
4. Improving management in manufacturing firms in India
5. Housing vouchers to move away from poor areas in the US
6. Road paving in Mexico

The Policies (Part 2/2)

7. E-procurement in India and Indonesia

8. Biometric ID cards in India

9. Anti-corruption in public works in Indonesia

10. Road safety in Kenya

11. Uma avaliação do impacto da qualidade da creche no desenvolvimento infantail (Rio de Janeiro)

12. Projeto Inverno Gaúcho da Secretaria de Saúde do Estado do Rio Grande do Sul

Or: choose your own policy (in small groups of people working on the same issue/sector)

Please go to your group's table and introduce yourselves

Describing Your Study

One way to understand a study is by using the PICO framework:

- Population (target group)
- Intervention (treatment)
- Comparison (counterfactual)
- Outcomes (measurement and results)

What was the overall finding? Did the policy work?

Take a few minutes to start doing this

Statistical uncertainty and significance

The basics of statistical uncertainty

Statistical uncertainty can make an evaluation misleading

- Even with a valid counterfactual

If we observe a positive treatment effect, is that due to a real causal impact or just chance?

- e.g. in Australian study, perhaps girls who were more likely to get pregnant just happened to be assigned to the treatment group

We can never know for sure, but we can use statistics to determine how confident we should be that an observed difference is real, not just pure chance

p-values

Suppose we want to compare two averages (e.g. outcomes in treatment and control groups) to see if the difference is a real one

- In Australia baby trial, treatment group girls had a 36% higher pregnancy rate than control group girls

Using statistics, we can calculate a p-value that tells us how likely we would be to see that difference purely due to chance

- In Australia, p-value was 0.0031
- This means that there is only a 0.3% chance of a 36% difference in pregnancy rates arising due to pure chance

We usually say an effect is statistically significant if its p-value is $< 5\%$

Brinkman, Sally *et al* 2016, "Efficacy of infant simulator programmes to prevent teenage pregnancy: a school-based cluster randomised controlled trial in Western Australia." *The Lancet*

Confidence intervals

We can then use the p-value to calculate a confidence interval: a range outside of which we think there is only a small chance that the true effect lies

We usually choose a 5% cut-off for p-values, which will create a 95% confidence interval

- In Australia, 95% confidence interval is 10% - 67% difference in pregnancy rates

Papers often report the confidence interval alongside the treatment effect (difference in means)

- Australia: treatment effect is 1.36 [95% CI: 1.10-1.67]

Uncertainty and sample size

Smaller samples $\rightarrow$ more statistical uncertainty (higher p-values)

- Smaller samples mean we should be more skeptical of evaluation findings
- As discussed yesterday, no exact rule on what is "too small"

Larger samples mean we can be more confident

- But even large studies with small p-values can be misleading

An evaluation with a valid counterfactual and statistically significant findings is usually a good indication of the true treatment effect

- But there is always some uncertainty

Statistical uncertainty in your studies

Take 5 minutes to look for evidence on statistical significance in your group's study

Look especially in these sections:

- Abstract
- Introduction
- Results

How statistically significant are your study's findings? Is there a chance the true effect could be zero?

Systematic reviews and meta-analysis

Aggregating studies to decrease statistical uncertainty

Studies with larger samples have more power

- But it is not always possible to conduct a large study

Another approach is to compare the results of many studies with:

- The same (or similar) treatment
- The same outcome measures

We can thus decrease statistical uncertainty by *aggregating* the results of these studies

- Gives us a broader picture than one study alone

Systematic review vs meta-analysis

Comparing the results of many studies lets us do two things:

1. Systematic review: Describe the results of these studies and what they tell us about the policy's effectiveness
    - Qualitative aggregation of results (*narrative synthesis*)

2. Meta-analysis: Combine information on effect sizes and confidence intervals to calculate an overall weighted effect
    - Quantitative aggregation of results

It is common for one paper to do both a systematic review and a meta-analysis

Example: nutritional advice and supplementation for young children

Food supplement programs for young children

- Similar to TINP and BINP

Researchers found 32 separate studies of these programs (26 RCTs)

- Described results and counted how many found positive/negative/zero impacts (systematic review)
- Combined data from all studies and analyzed jointly (meta-analysis)

Kristjansson *et al* 2015, "Food Supplementation for Improving the Physical and Psychosocial Health of Socio-economically Disadvantaged Children Aged Three Months to Five Years: A Systematic Review", *Campbell Systematic Reviews* 11.

Campbell Systematic Reviews
2015:11
First published: 04 May, 2015
Search executed: 20 March, 2014

Food Supplementation for Improving the Physical and Psychosocial Health of Socio-economically Disadvantaged Children Aged Three Months to Five Years: A Systematic Review

Elizabeth Kristjansson, Damian K Francis, Selma Liberato, Maria Benkhalti Jandu, Vivian Welch, Malek Batal, Trish Greenhalgh, Tamara Rader, Eamonn Noonan, Beverley Shea, Laura Janzen, George A Wells, Mark Petticrew

What did they find?

Supplementary feeding in low/middle-income countries usually had positive effects on child growth

- No clear effect on cognitive development

Note: this is the *average* effect of the interventions across all the contexts in which they've been evaluated

Results of systematic review and meta-analysis

We included 32 studies (21 RCTs and 11 CBAs); 26 of these (16 RCTs and 10 CBAs) were in meta-analyses. More than 50% of the RCTs were judged to have low risk of bias for random selection and incomplete outcome assessment. We judged most RCTS to be unclear for allocation concealment, blinding of outcome assessment, and selective outcome reporting. Because children and parents knew that they were given food, we judged blinding of participants and personnel to be at high risk for all studies.

Growth. Supplementary feeding had positive effects on growth in low- and middle-income countries. Meta-analysis of the RCTs showed that supplemented children gained an average of 0.12 kg more than controls over six months (95% confidence interval (CI) 0.05 to 0.18, 9 trials, 1057 participants, moderate quality evidence). In the CBAs, the effect was similar; 0.24 kg over a year (95% CI 0.09 to 0.39, 1784 participants, very low quality evidence). In high-income countries, one RCT found no difference in weight, but in a CBA with 116 Aboriginal children in Australia, the effect on weight was 0.95 kg (95% CI 0.58 to 1.33). For height, meta-analysis of nine

Forest plots (output of meta-analysis)

Analysis 1.3. Comparison 1 Low- and middle-income countries: feeding vs control - growth. RCT, Outcome 3 Weight-for-age z-scores (WAZ).

Review: Food supplementation for improving the physical and psychosocial health of socio-economically disadvantaged children aged three months to five years

Comparison: 1 Low- and middle-income countries: feeding vs control - growth. RCT

Outcome: 3 Weight-for-age z-scores (WAZ)

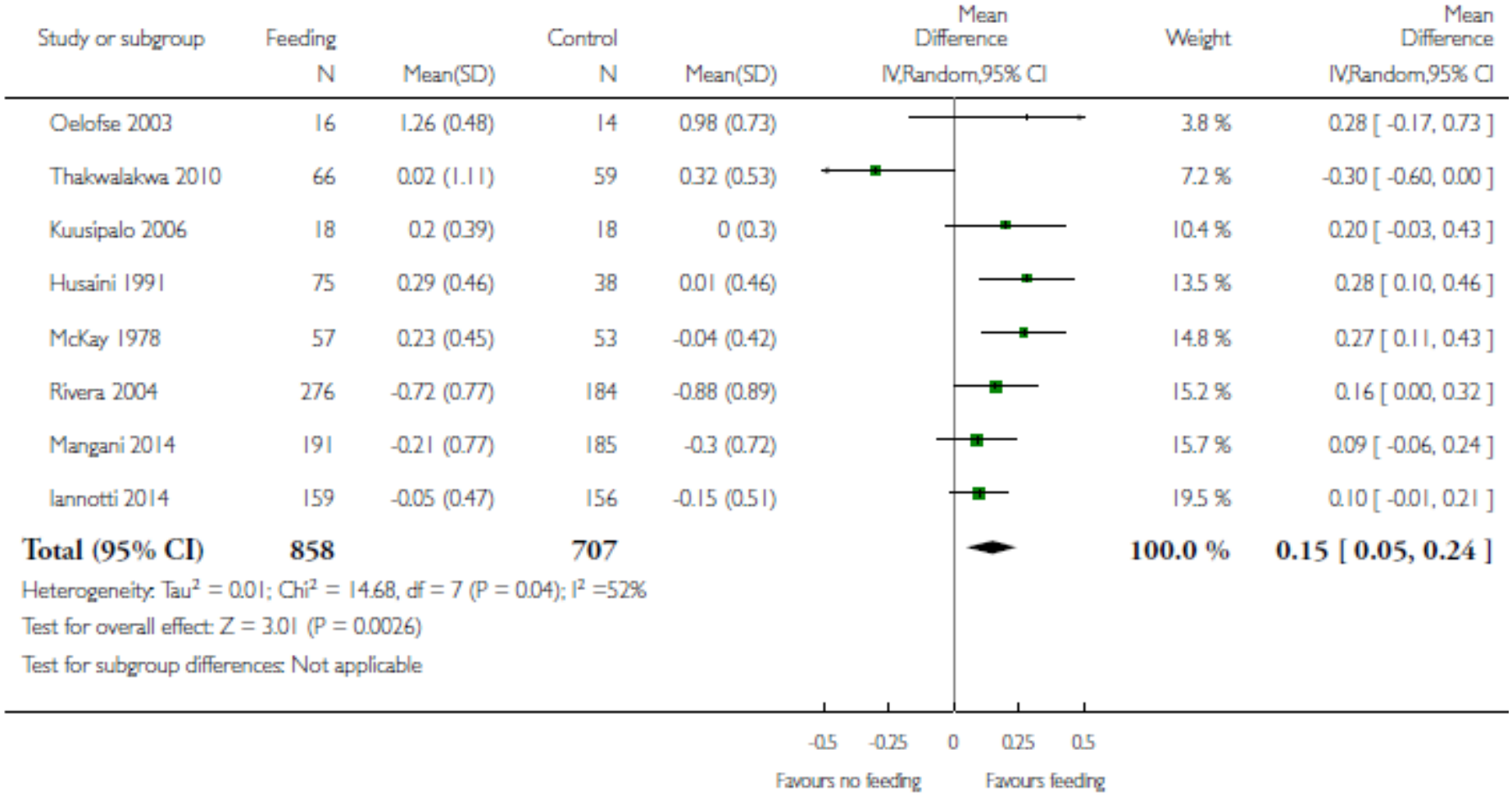
Study or subgroup	Feeding N	Mean(SD)	Control N	Mean(SD)	Mean Difference IV,Random,95% CI	Weight	Mean Difference IV,Random,95% CI
Oelofse 2003	16	1.26 (0.48)	14	0.98 (0.73)		3.8 %	0.28 [-0.17, 0.73]
Thakwalakwa 2010	66	0.02 (1.11)	59	0.32 (0.53)		7.2 %	-0.30 [-0.60, 0.00]
Kuusipalo 2006	18	0.2 (0.39)	18	0 (0.3)		10.4 %	0.20 [-0.03, 0.43]
Husaini 1991	75	0.29 (0.46)	38	0.01 (0.46)		13.5 %	0.28 [0.10, 0.46]
McKay 1978	57	0.23 (0.45)	53	-0.04 (0.42)		14.8 %	0.27 [0.11, 0.43]
Rivera 2004	276	-0.72 (0.77)	184	-0.88 (0.89)		15.2 %	0.16 [0.00, 0.32]
Mangani 2014	191	-0.21 (0.77)	185	-0.3 (0.72)		15.7 %	0.09 [-0.06, 0.24]
Iannotti 2014	159	-0.05 (0.47)	156	-0.15 (0.51)		19.5 %	0.10 [-0.01, 0.21]
Total (95% CI)	**858**		**707**			**100.0 %**	**0.15 [0.05, 0.24]**

Heterogeneity: $Tau^2 = 0.01$; $Chi^2 = 14.68$, df = 7 (P = 0.04); $I^2 = 52\%$

Test for overall effect: Z = 3.01 (P = 0.0026)

Test for subgroup differences: Not applicable

Forest plots (output of meta-analysis)



Strengths and limitations of systematic reviews and meta-analysis

Helps us reduce statistical uncertainty by increasing our sample size

Gives us a better estimate of the average treatment effect than a single study can give us

What is the major limitation of aggregating studies from various countries like this?

- The effect in a specific context may be different from the average effect worldwide
- We'll discuss this more tomorrow

Sources of evidence

Where can I find evidence?

There is not a single unified source for everything

- So you have to search multiple sources and be persistent

We'll review some sources for:

- Impact evaluations
- Systematic reviews and meta-analyses
- Other evidence (policy briefs, process evaluations, etc.)

International repositories of evidence (1/2)

Most repositories have impact evaluations and systematic reviews

Cochrane Collaboration - health-focused

- www.cochrane.org

Campbell Collaboration - evidence-based policy more generally

- www.campbellcollaboration.org

3ie - development-focused

- http://www.3ieimpact.org

International repositories of evidence (2/2)

UK government has seven "What Works Centres":

National Institute for Health and Care Excellence (NICE)	Health and social care
Sutton Trust/Educational Endowment Foundation	Educational achievement
College of Policing What Works Centre for Crime Reduction	Crime reduction
Early Intervention Foundation	Early intervention
What Works Centre for Local Economic Growth (hosted by LSE, Arup, Centre for Cities)	Local economic growth
Centre for Ageing Better	Improved quality of life for older people
What Works Centre for Wellbeing	Wellbeing

https://www.gov.uk/guidance/what-works-network

Example: finding evidence with 3ie

3ie International Initiative for Impact Evaluation

find Evidence | use Evidence | get Funding | improve Evaluation | engage with 3ie

New blog: Any chance to use impact evaluations with no impact? The case of Mexico

If a country has an institutional and relatively credible monitoring and evaluation system, the chances of using impact evaluations showing lack of programme impact increase, says CONEVAL Mexico's executive secretary and 3ie board member, Gonzalo Hernández Licona. In this blog, he gives examples of cases in Mexico where impact evaluations had found null or negative results but policymakers used the findings for informing decisions.

read more

Announcements

Funding

- Coming soon: Call for process evaluation proposal for Vegetable Oil Development Programme in Uganda
- Request for proposals: Birth registration impact evaluation, UNICEF, Nigeria

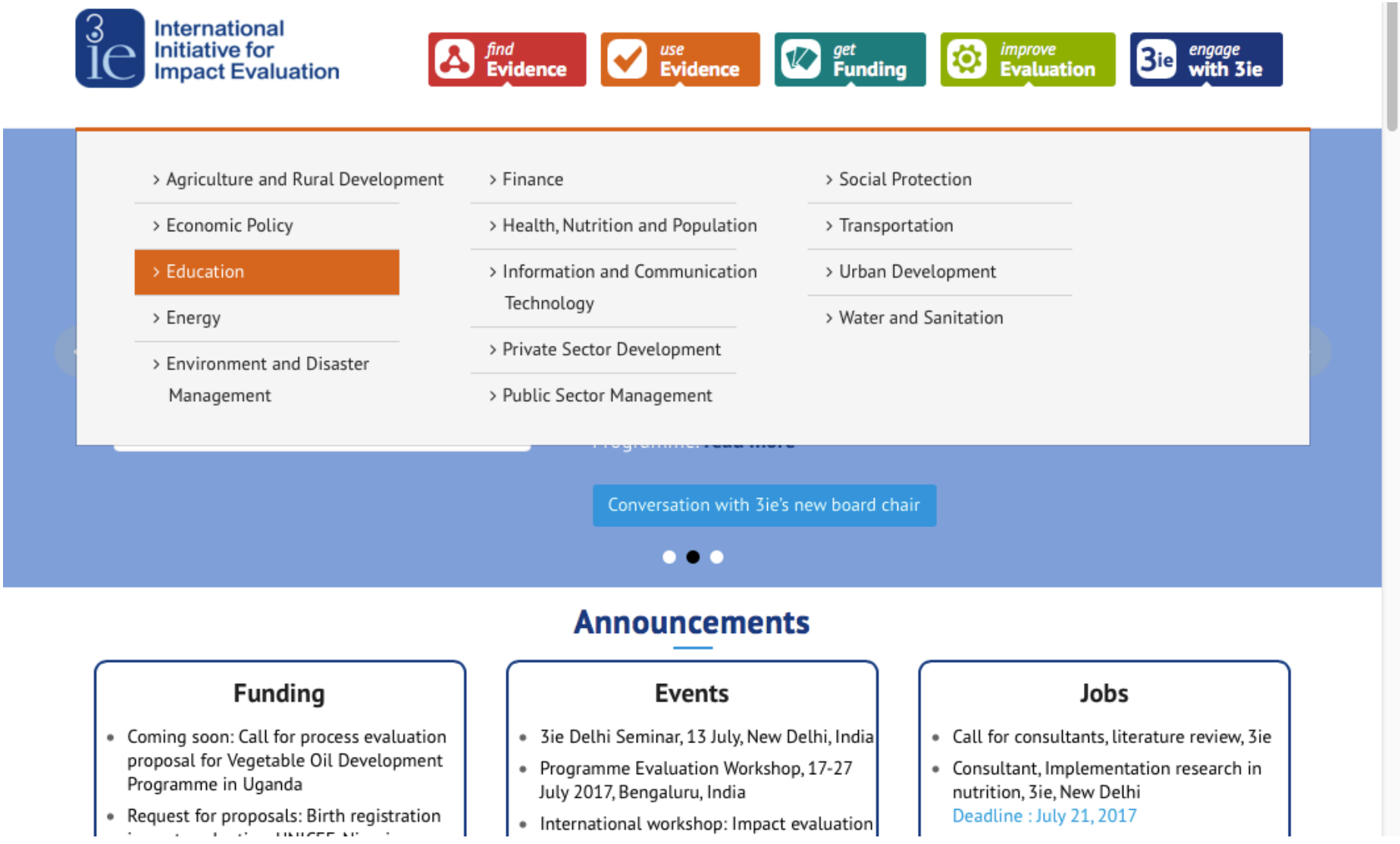
Events

- 3ie Delhi Seminar, 13 July, New Delhi, India
- Programme Evaluation Workshop, 17-27 July 2017, Bengaluru, India
- International workshop: Impact evaluation of population, health and nutrition

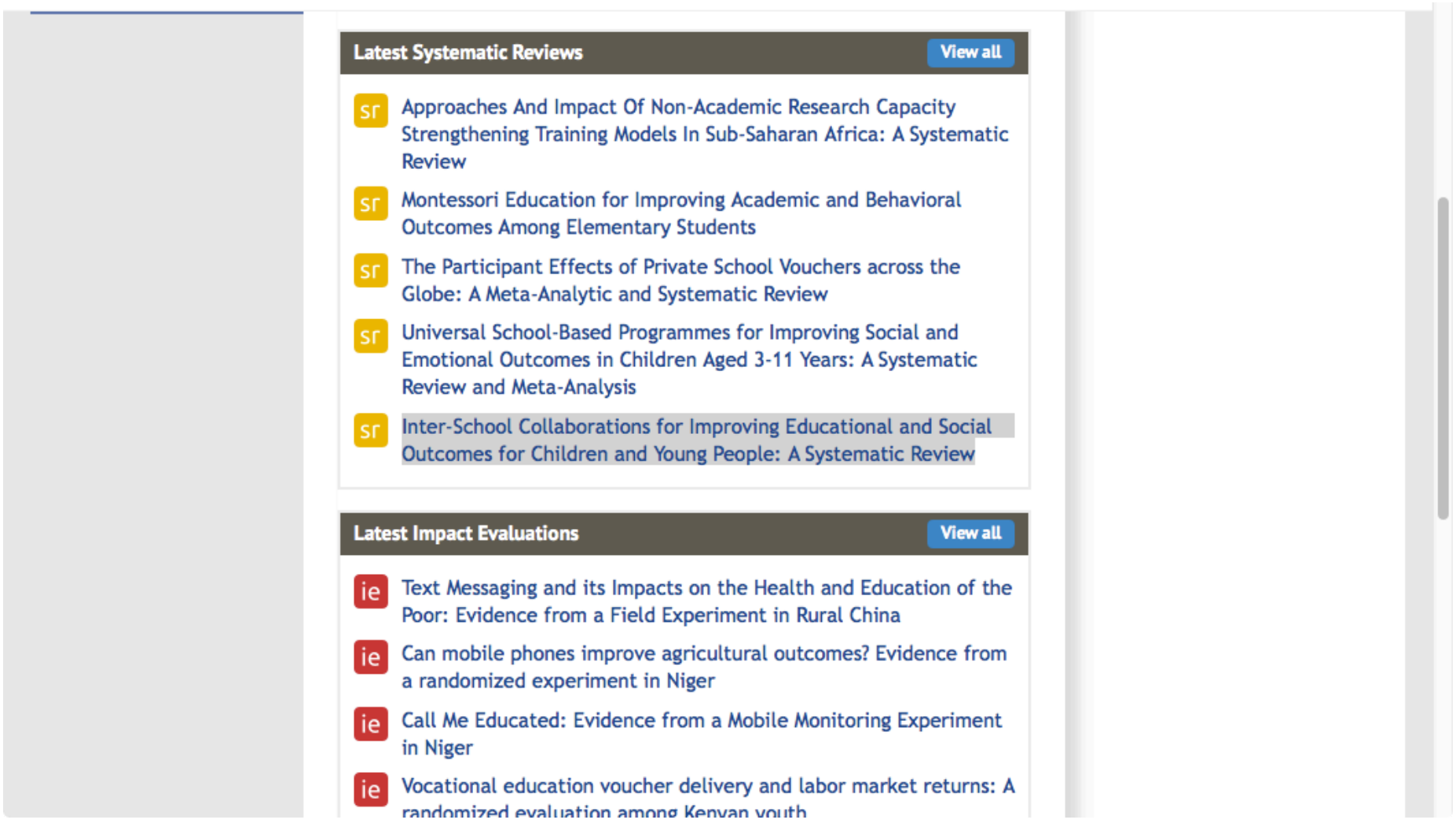
Jobs

- Call for consultants, literature review, 3ie
- Consultant, Implementation research in nutrition, 3ie, New Delhi
  Deadline : July 21, 2017
- Monitoring, evaluation and learning officer

Example: finding evidence with 3ie



Example: finding evidence with 3ie



Example: finding evidence with 3ie

Example: finding evidence with 3ie

International Initiative for Impact Evaluation

Impact evaluation brief
Education

Evaluating the impact of computer-assisted instruction and learning in China

The use of information and communication technology (ICT) for teaching is a promising approach for improving learning outcomes, particularly for disadvantaged children in developing countries. In many countries, large investments have been made to integrate ICT into education systems. For instance, as part of its 12th Five-Year Plan, China's central government earmarked funds to provide a computer room in every rural school.

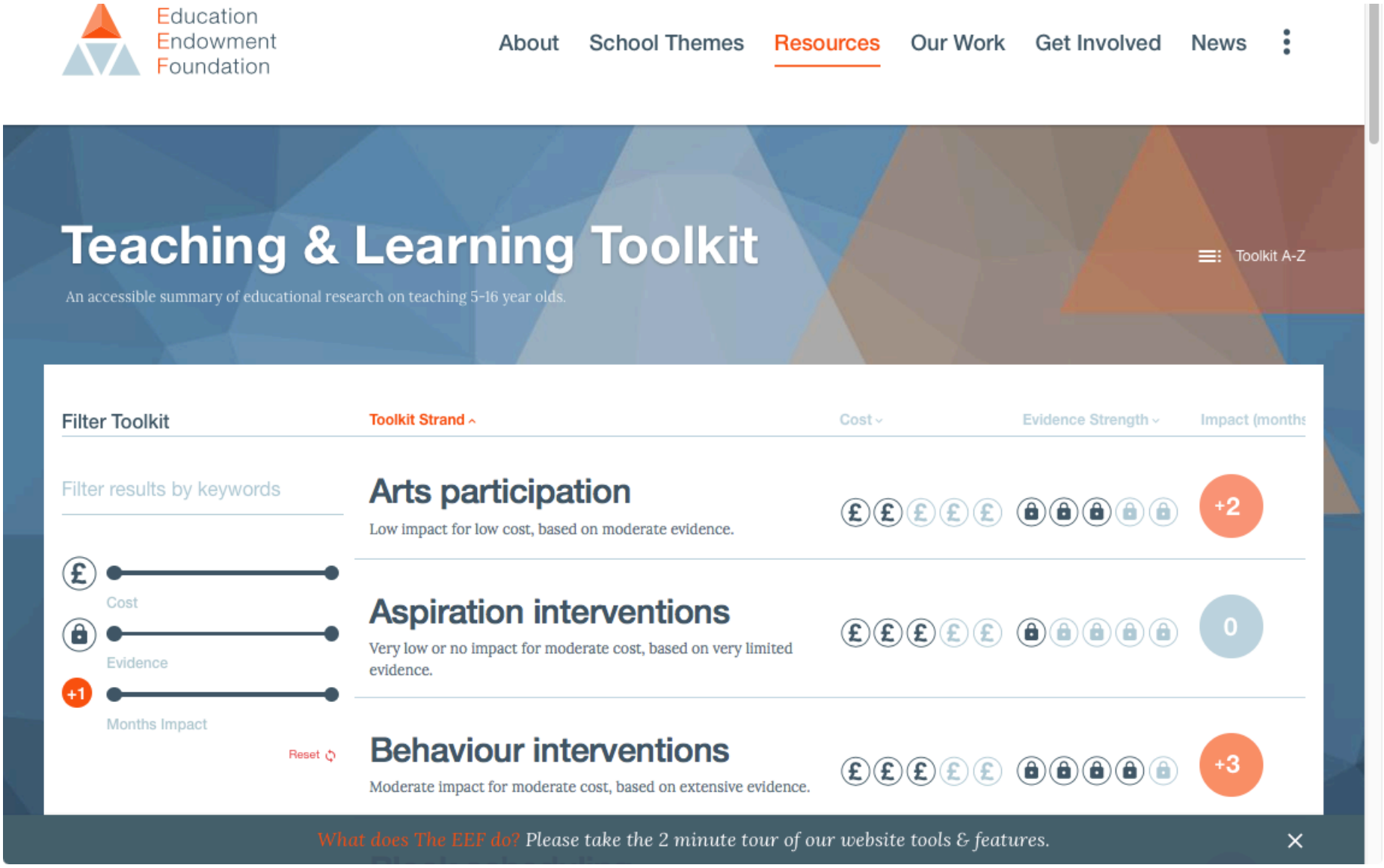
Despite the popularity of ICT use in education, researchers have so far found significant variability in the impact of ICT programmes on students' academic achievement. A recent 3ie systematic review[1] on education effectiveness found that computer-assisted learning (CAL) programmes have had mixed effects on learning outcomes, and in some contexts the effects have been negative.

3ie supported a research team to examine some of

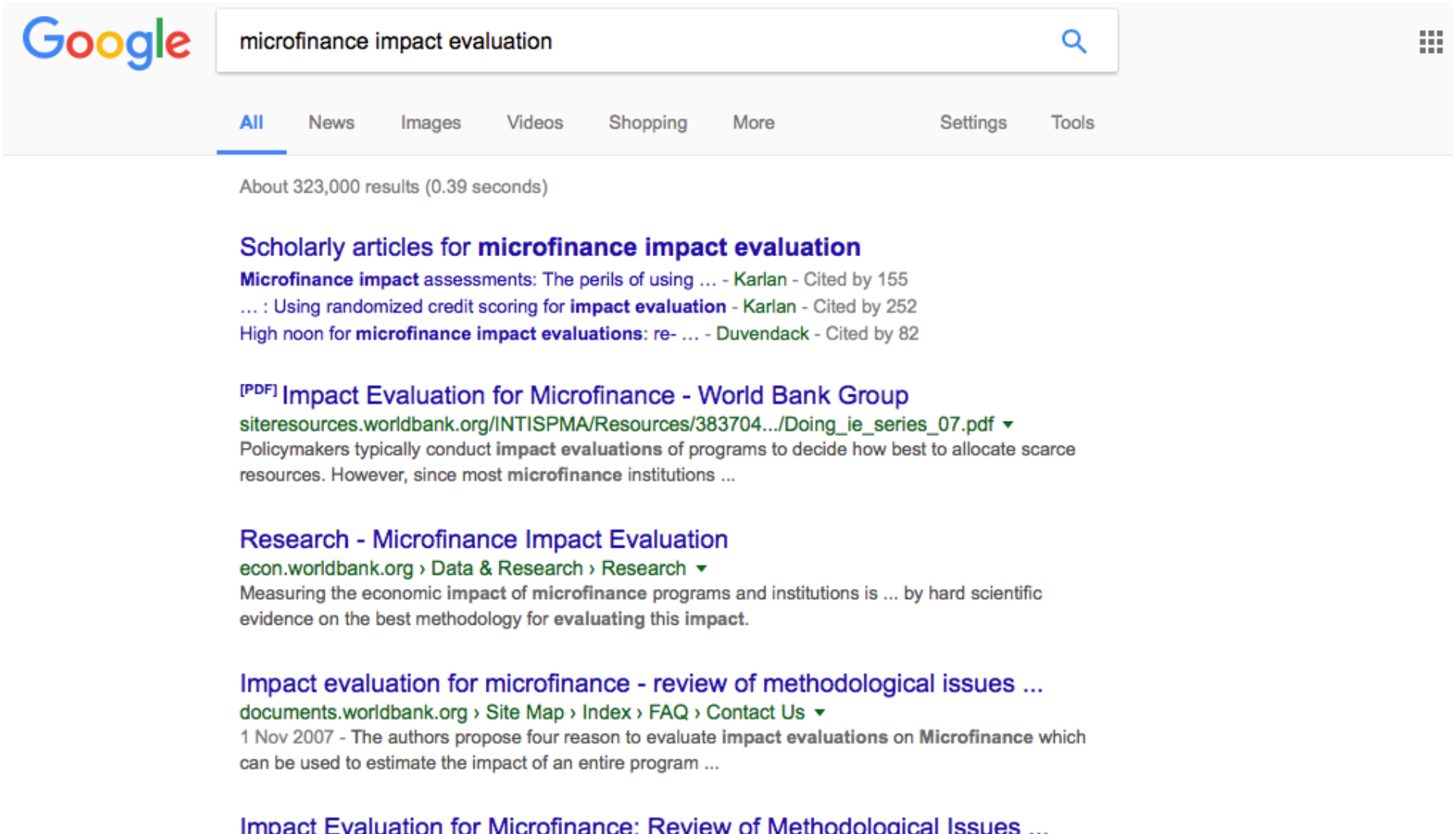
Main findings

- The researcher-implemented CAL programme was effective in improving test scores in English. However, the same programme was ineffective when implemented by the government.
- Schools in the government-implemented programme were more likely to use existing English teachers to supervise the CAL programme and replace English classes with CAL sessions. This substitution could have been one of the reasons that the programme did not have an impact.
- The computer-assisted instruction (CAI) programme was more effective than the CAL programme in improving students' English language test scores.
- Both the better-performing and worse-

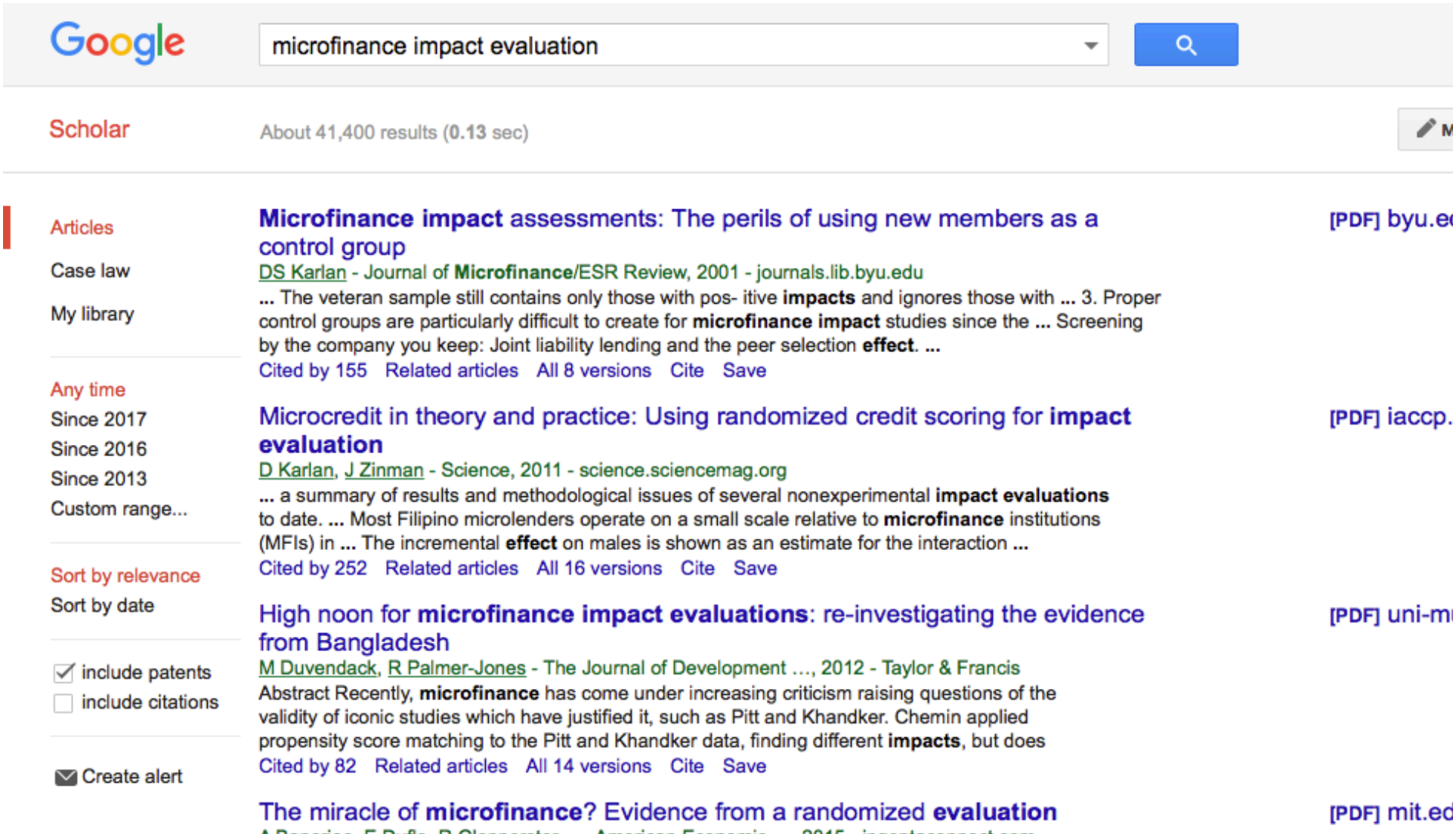
Example: impact, reliability, and cost of education policies



When in doubt, Google



Google Scholar



Evaluation organizations

Independent or pseudo-governmental

- e.g. MDRC

Government organizations

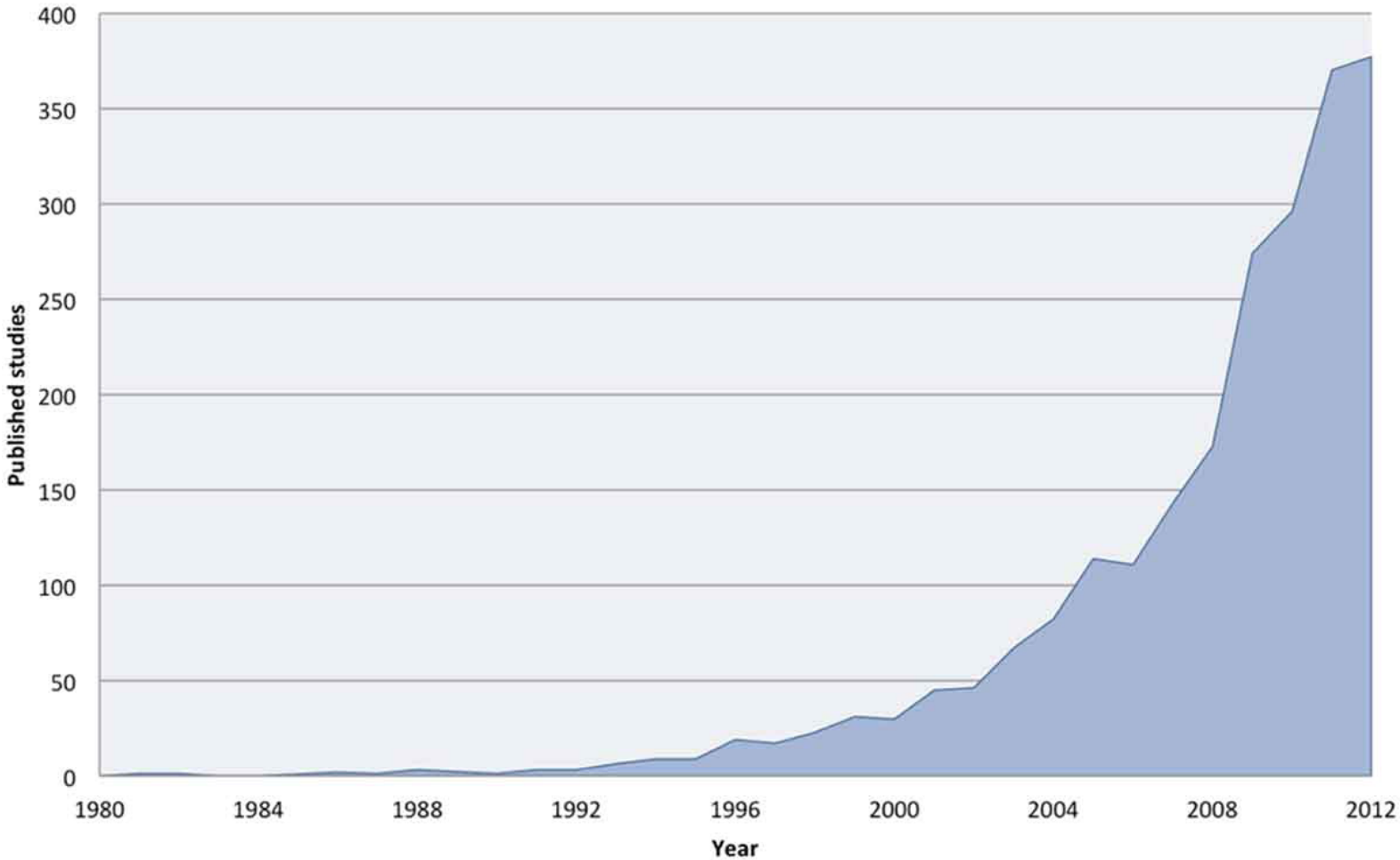
- e.g. CONEVAL

MDRC - USA


mdrc
BUILDING KNOWLEDGE TO IMPROVE SOCIAL POLICY
Solutions We're Working On
Questions We're Asking
Issues We Focus On
Populations We Work With
Publications
Projects
implementation research incubator
idea
MDRC FEATURE
Latest from @MDRC_News
Follow mdrc:
RT @RichardvReeves: No, research does not say that you produce more when working 40 hours per week https://t.co/LeUMJk1GM1 via @luispedrocoelho
RT @EdTrust: Thank you Sandy McMakin of @uiwcardinals, Alexander Mayer of @MDRC_News, Thomas Kane of @HarvardCEPR & @JohnBKing for a great panel on RCTs
REPORT
Aligning Aid with Enrollment
Interim Findings on Aid Like A Paycheck
ISSUE FOCUS
A Bipartisan Way to Make Work Pay for Low-Wage Workers
BRIEF
Every Step Counts
Building a School Choice Architecture

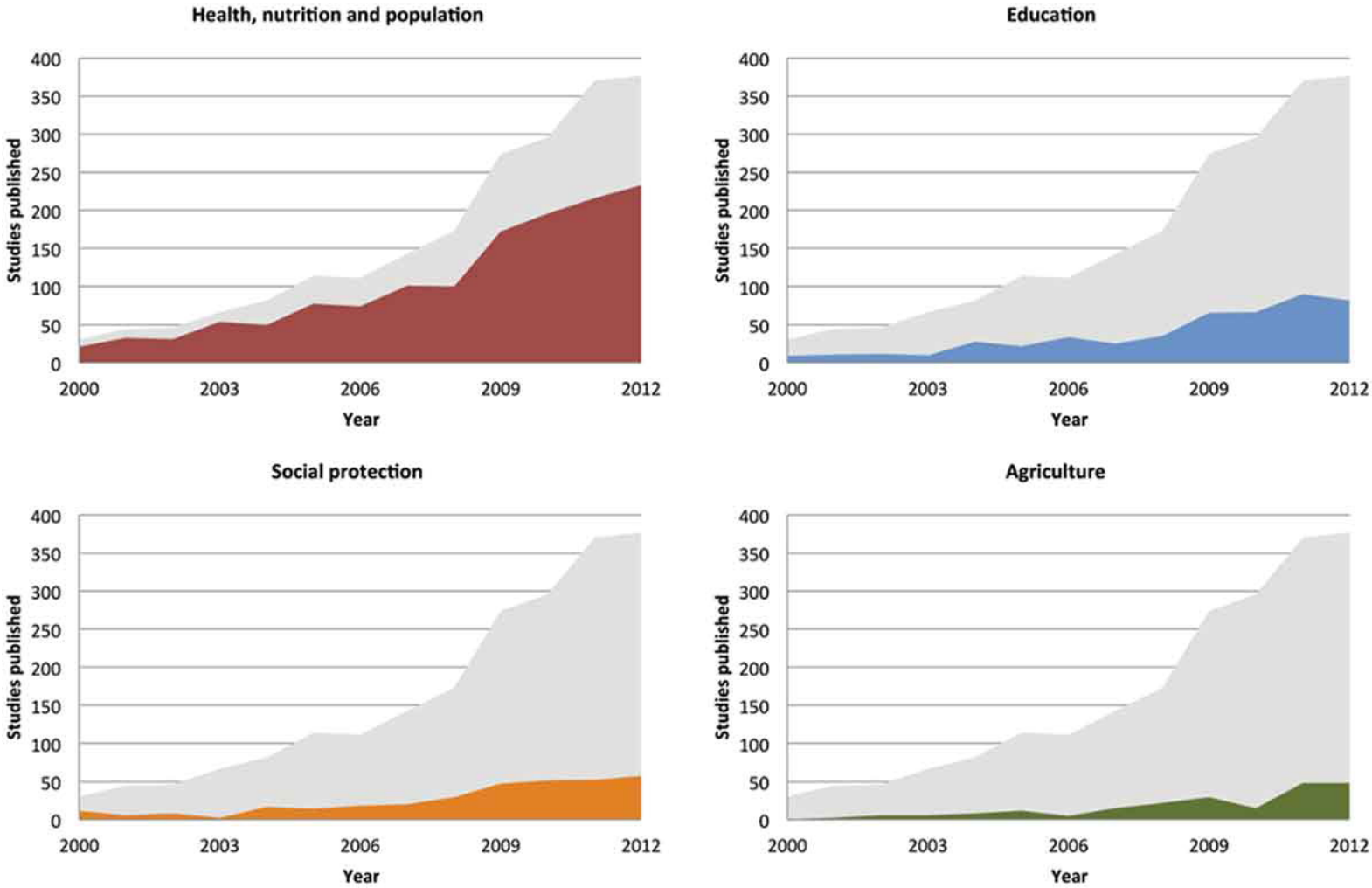
CONEVAL - Mexico

Increasing availability of evidence



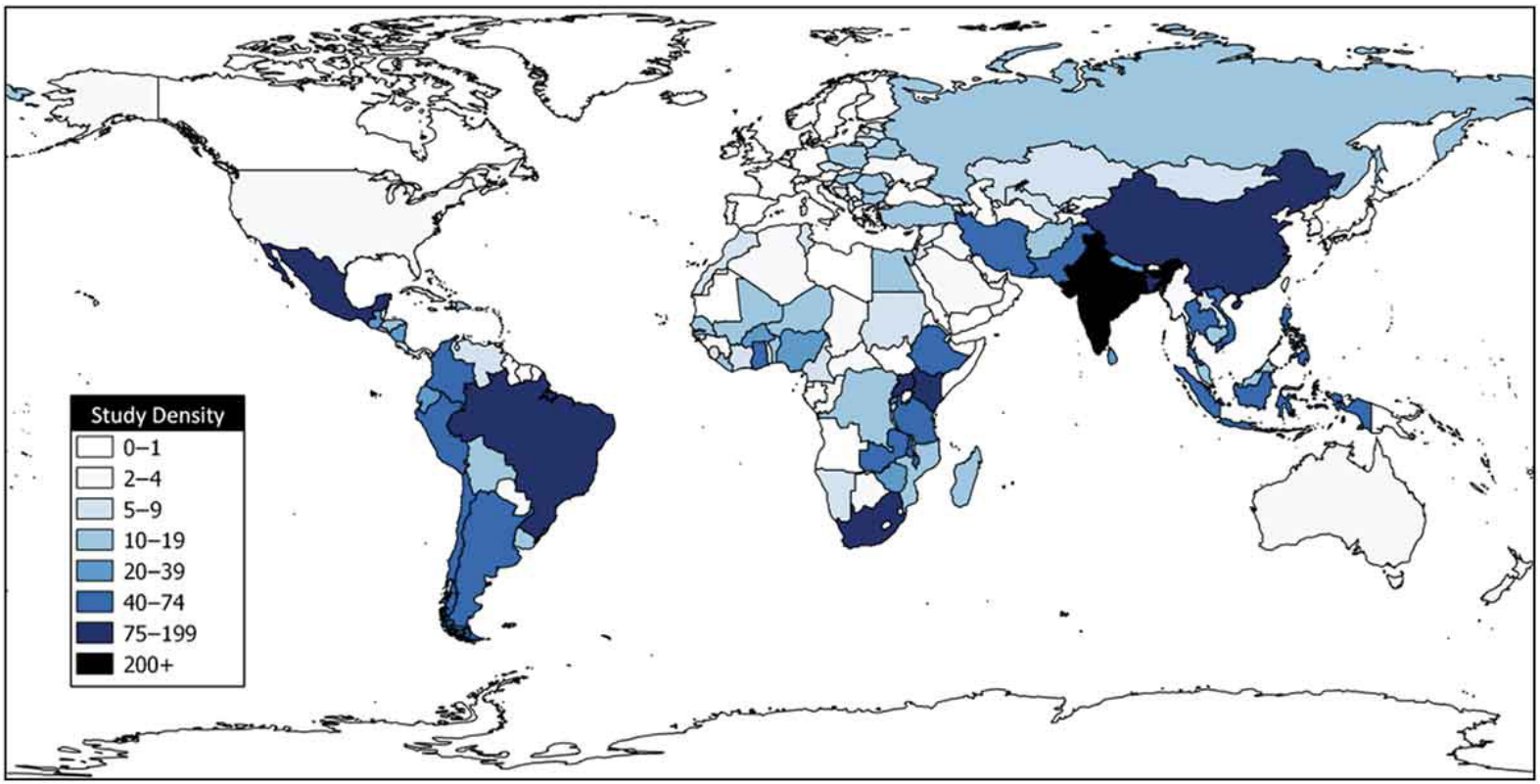
Source: Drew B. Cameron, Anjini Mishra, and Annette N. Brown. The growth of impact evaluation for international development: how much have we learned? Journal Of Development Effectiveness Vol. 8 , Iss. 1,2016

But in some sectors more than others



Source: Drew B. Cameron, Anjini Mishra, and Annette N. Brown. The growth of impact evaluation for international development: how much have we learned? Journal Of Development Effectiveness Vol. 8 , Iss. 1,2016

Geographical coverage (developing countries only)



Source: Drew B. Cameron, Anjini Mishra, and Annette N. Brown. The growth of impact evaluation for international development: how much have we learned? Journal Of Development Effectiveness Vol. 8 , Iss. 1,2016

Be critical of evaluation quality

There are many low-quality evaluations

- Invalid counterfactual, poor measurement, small sample. . .

Read published impact evaluations with a critical lens

- Check the issues we discussed yesterday and this morning
- Critical Appraisal Skills Programme (CASP) RCT checklist

  http://www.casp-uk.net/casp-tools-checklists
- Is it from a reputable source? (even if yes, still be critical)

Recap: evidence sources

Institutional repositories

- e.g. Campbell, Cochrane, 3ie, What Works

Google, Google Scholar

Evaluation organizations

Recursos em português?

Recap of the day

What one idea or concept will you remember from today?

Impact evaluation is important

- But the same policy can have different effects in different contexts

Review of impact evaluation methods - counterfactuals!

- RCTs and difference-in-difference

The plan for Thursday

Thursday: external validity

- Finish reading your study by Thursday
- Please bring a laptop on Thursday!

See you Thursday!

External Validity and Mechanism Mapping

Martin J. Williams

Blavatnik School of Government, University of Oxford

ENAP, 4-8 June, 2018

Some review from Monday

Fundamental problem of evaluation

- Counterfactual: what would have happened without the policy?

Many ways people often think of policy effectiveness do not have a valid counterfactual

- Before-after comparisons
- Treated-untreated comparisons

Two ways to estimate valid counterfactuals

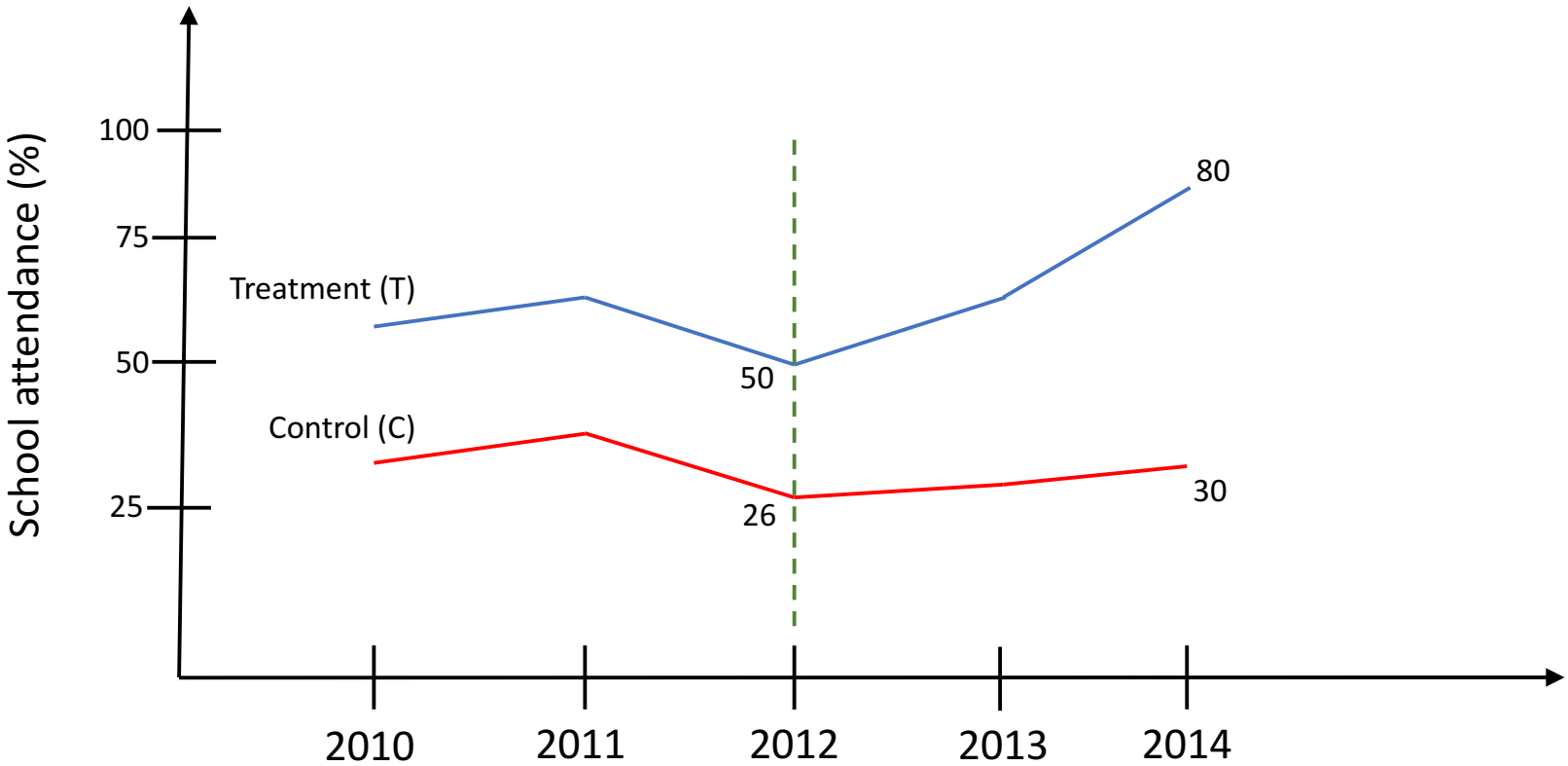
Randomized controlled trials (RCTs)

- Randomly select which people/units get policy
- For large enough sample, control group can then be assumed to be a valid counterfactual
- Can then just compare average outcomes of treatment vs control groups

Difference-in-difference (DiD)

- Compare treatment and control before AND after implementation
- . . . and look at *change* in treatment group outcome compared to *change* in control group outcome

Difference-in-difference



Uncertainty

Statistical uncertainty can make an evaluation misleading

- Even with a valid counterfactual

Ways to assess how statistically significant a finding is

- p-value: how likely we would be to see a difference between treatment and control purely due to chance, if the true difference is 0
- We usually choose a 5% cut-off for p-values, which we can use to create a 95% confidence interval around the observed difference in means

Smaller samples and/or more variable outcomes $\rightarrow$ more statistical

External validity

A tale of evidence and context

In 2015, Zimbabwe rolled out a new HIV drug, efavirenz

- Recommendation from World Health Organization (WHO)
- Based on evidence from numerous RCTs worldwide

But soon, many people began quitting the drug

- Disastrous for treatment and for transmission prevention

Efavirenz causes hallucinations, depression, etc., in individuals with a certain gene

- Gene is rare worldwide, but common in Zimbabwe

A tale of evidence and context

But Zimbabwean scientists had identified this interaction years before

- "It's not a bad drug. We just know it can be improved in Africa"
  (Masimirembwa in Nordling 2017)

Sole reliance on evidence from elsewhere can be seriously misleading

This is the problem of external validity of evaluation results

Sources: Nordling 2017, Masimirembwa *et al* 2016, Nyakutira *et al* 2008

External validity

On Monday, we learned to evaluate whether a policy worked in other contexts

But "it worked there" $\neq$ "it will work here"

- There vs here
- That group vs this group
- Then vs now
- Pilot vs full-scale
- Etc . . .

Today, we learn how to think about making this leap

Remember our example from Monday

Tamil Nadu Integrated Nutrition Project (TINP)

- Nutritional counselling for mothers + supplementary food for infants
- Improved nutritional knowledge in mothers, significant decline in child malnutrition

Bangladesh Integrated Nutrition Project (BINP)

- Program design copied from TINP; covered $1/8^{th}$ of Bangladesh
- Improved nutritional knowledge in mothers, but no impact on child malnutrition
- Because in Tamil Nadu mothers make food decisions, but in Bangladesh fathers and mothers-in-law make them

Example from Cartwright and Hardie, (2012), *Evidence-Based Policy: A Practical Guide to Doing it Better*, OUP.

Outline for today

1. External validity
2. Ways to address external validity
3. Mechanism mapping
4. In groups: finding and interpreting evidence
5. Panel: Evidence in Policymaking in Brazil

External validity - concept and definitions

Internal vs external validity

Internal validity

- "whether the results [of an evaluation] are a true reflection of the impact on the individuals being studied"
- Did it work (there)?

External validity

- "whether the impact estimated for those directly studied [in an evaluation] can be extrapolated / generalised to others"
- Will it work here?

UK Government Magenta Book, p.127

Two types of external validity

1. **Policy transportation**: moving a policy from one context to another
   - From one target population to a new target population
   - Example: Bangladesh Integrated Nutrition Programme (BINP)
   - What you are doing in your small groups

2. **Scaling up**: going from small-scale implementation to large-scale implementation
   - From a study sample (or small target population) to a larger target population
   - Example: Tools of the Mind program (USA)

Scale-up

Tools of the Mind in the USA

- Pre-K program targeting improved executive function
- Small RCT in New Jersey found it was effective

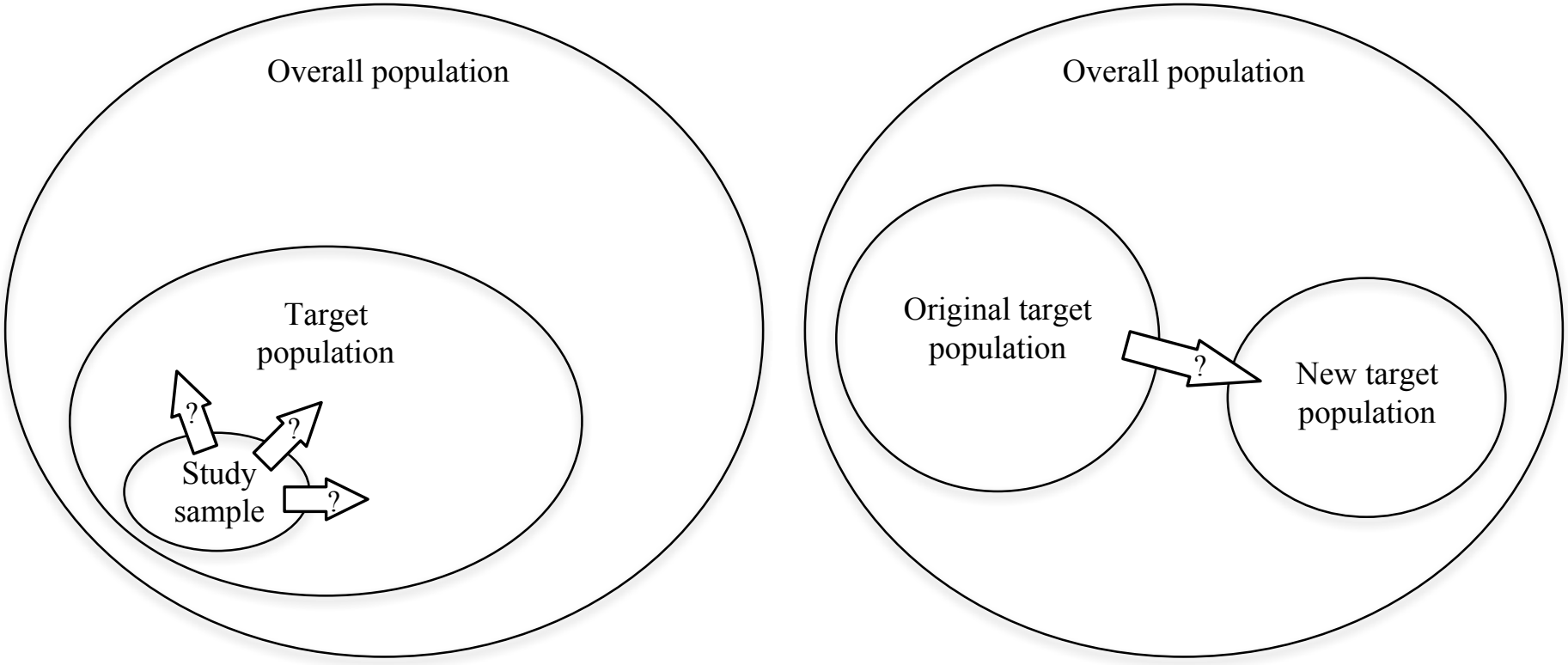
Large federally funded RCT in other states found negative impacts

- Why?
- Teachers had limited time to learn how to deliver it effectively
- Curriculum didn't fit well into the school day

Diamond *et al* 2007, "Preschool Program Improves Cognitive Control", *Science* 318(5855): 1387 - 1388.

Farran and Wilson 2014, "Achievement and Self-Regulation in Pre-Kindergarten Classrooms: Effects of the Tools of the Mind Curriculum"

Two types of external validity



External validity

Suppose a policy worked somewhere else (i.e., there was an internally valid evaluation that showed it worked)

Why might this not be a good predictor of whether the same policy will work in your country?

- Or on a larger scale in the same country?

Because differences in context can interact with a policy's theory of change

Theory of change

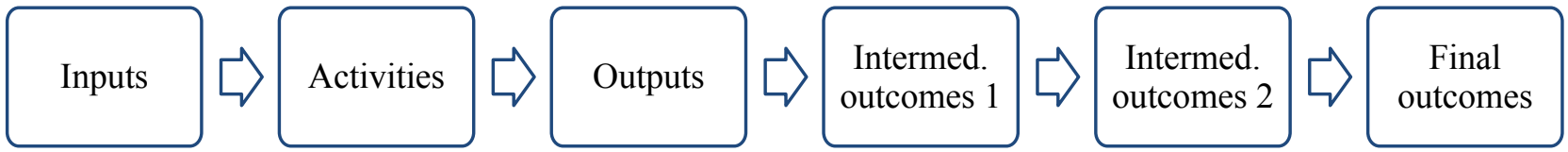
A policy's theory of change is the sequence of steps linking inputs to outcomes

- Theory of change is sometimes called mechanism, results chain, logic model

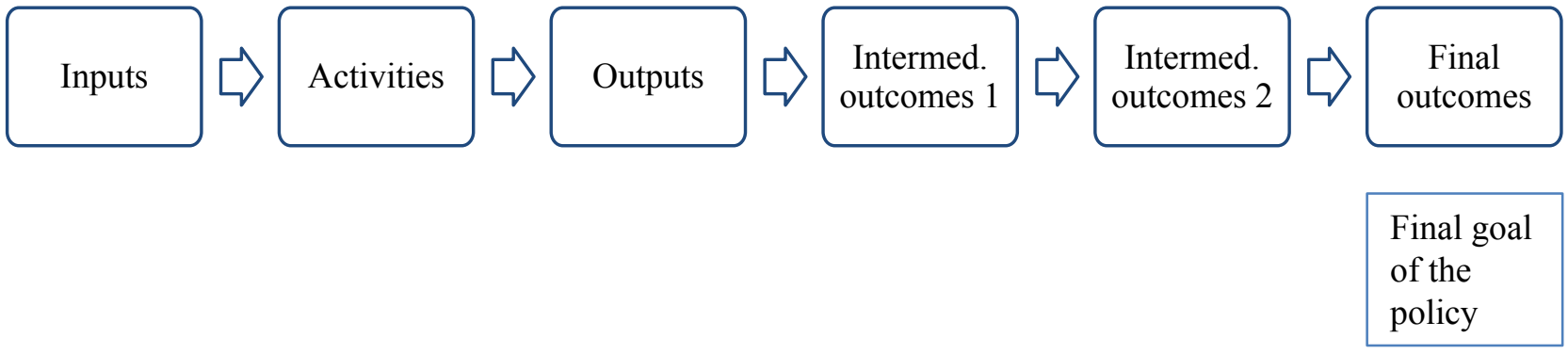
A theory of change explains *how* a policy is supposed to work

- e.g. Improved nutritional knowledge in mothers + additional food $\rightarrow$ mothers and children eat more food
- e.g. Taking a photo of teachers/nurses every day and rewarding attendance will reduce absenteeism and increase service quality

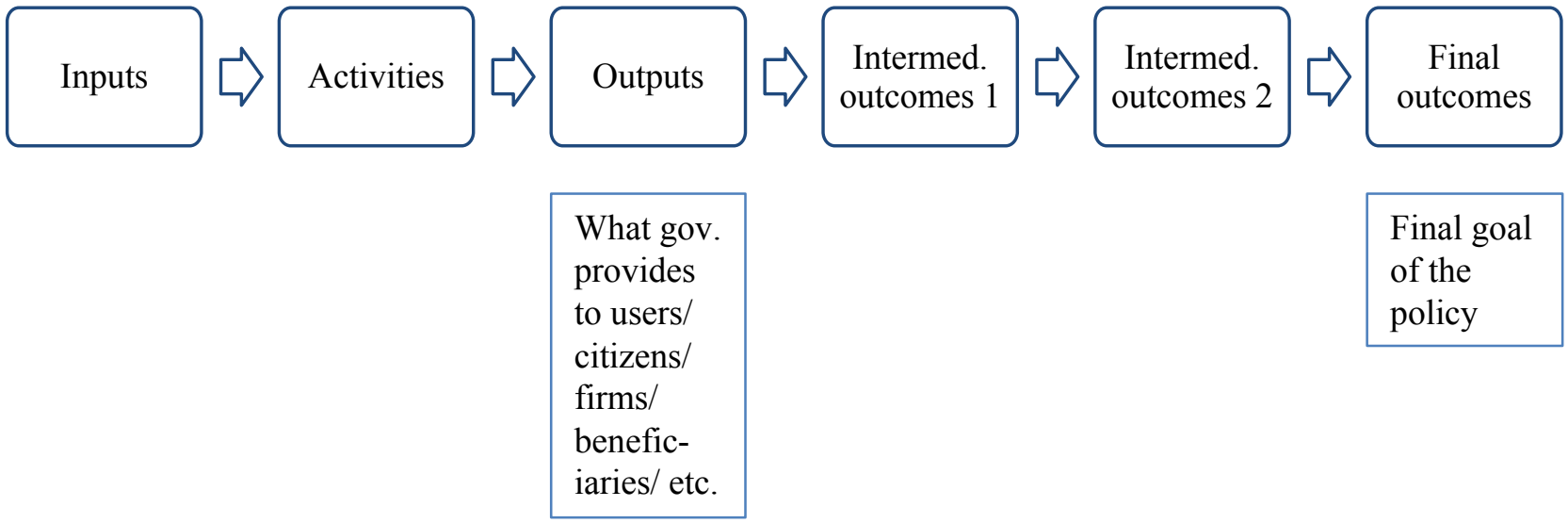
A theory of change



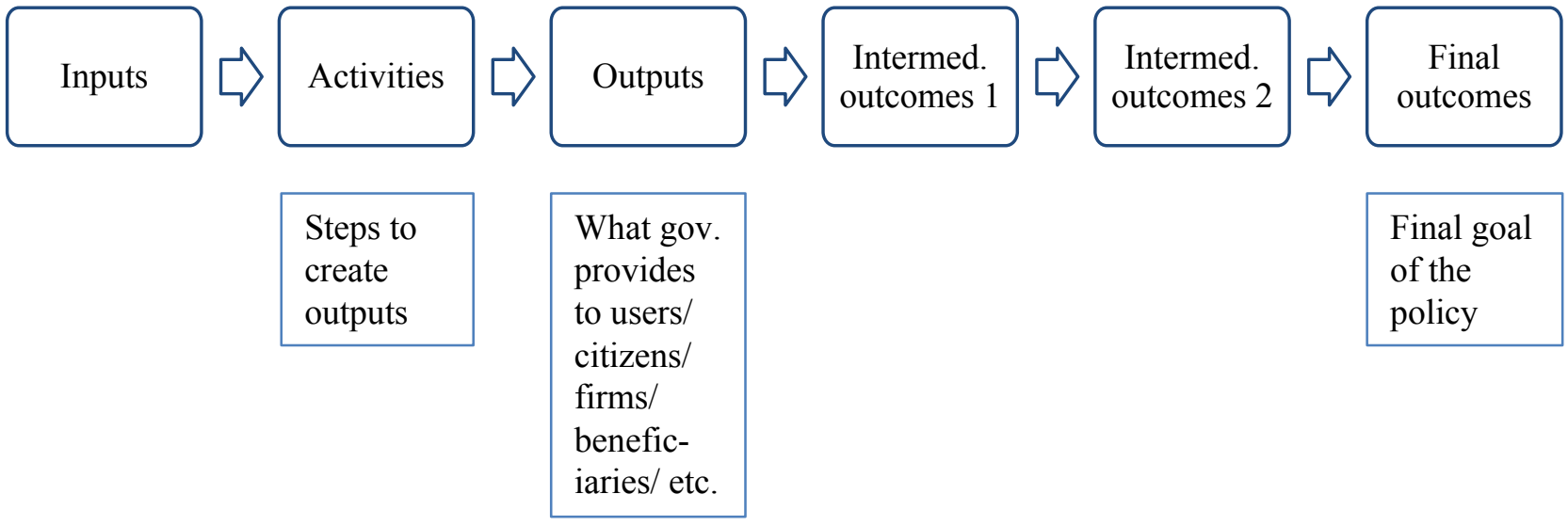
A theory of change



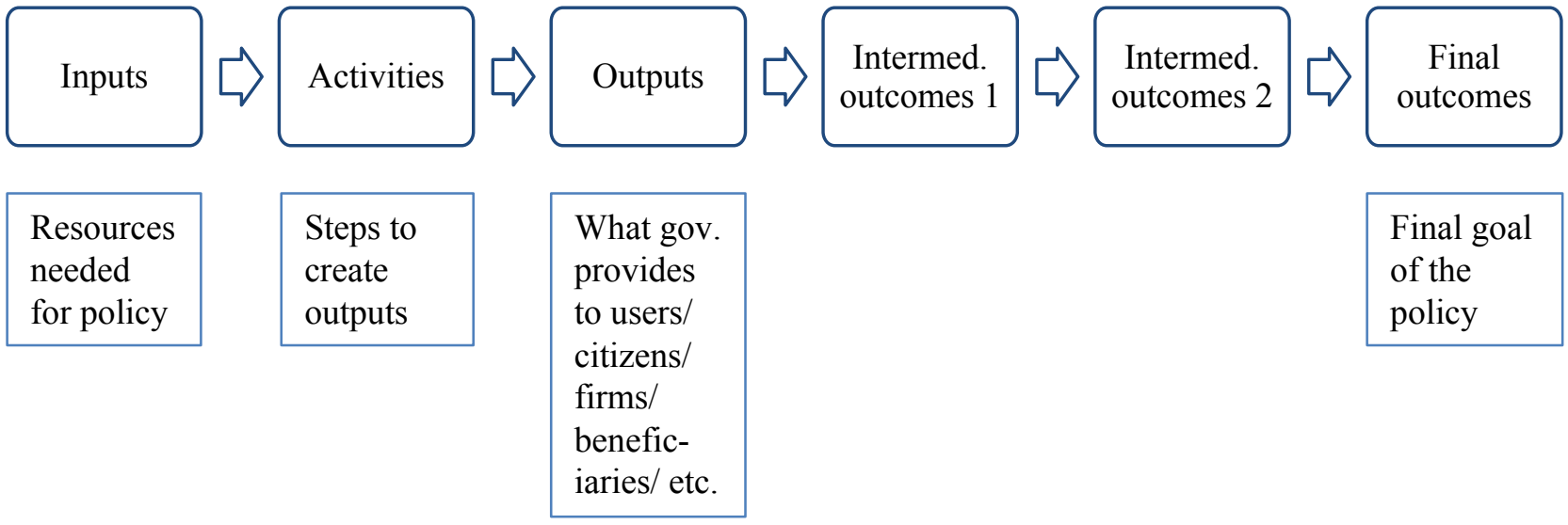
A theory of change



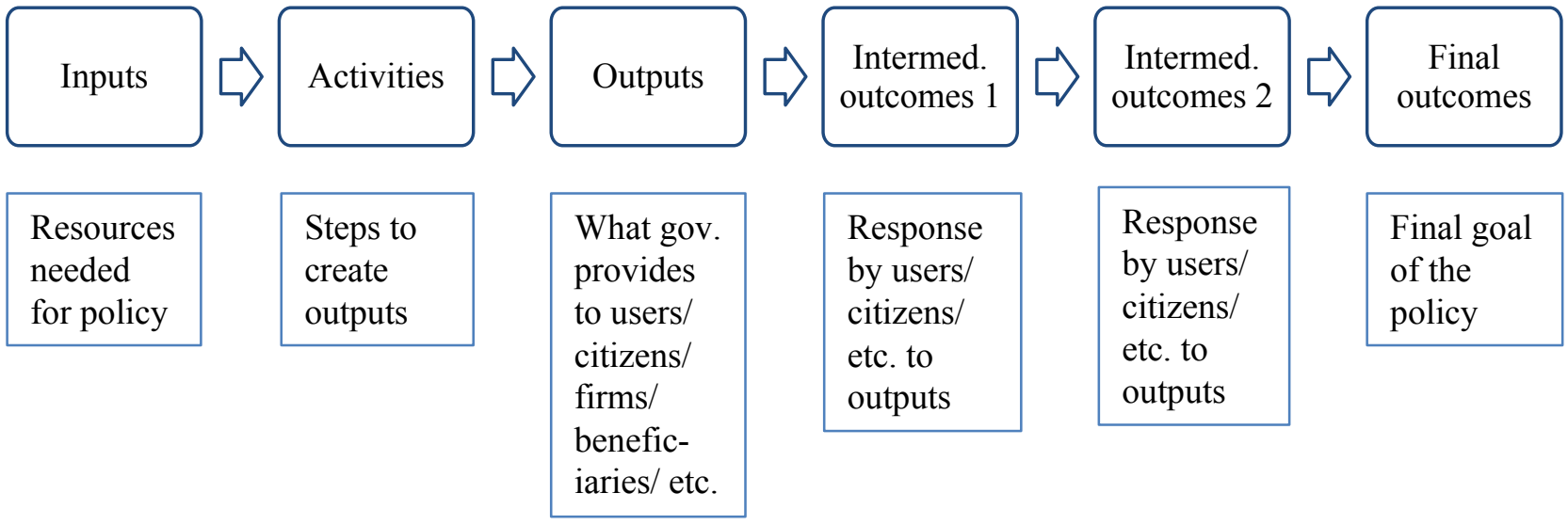
A theory of change



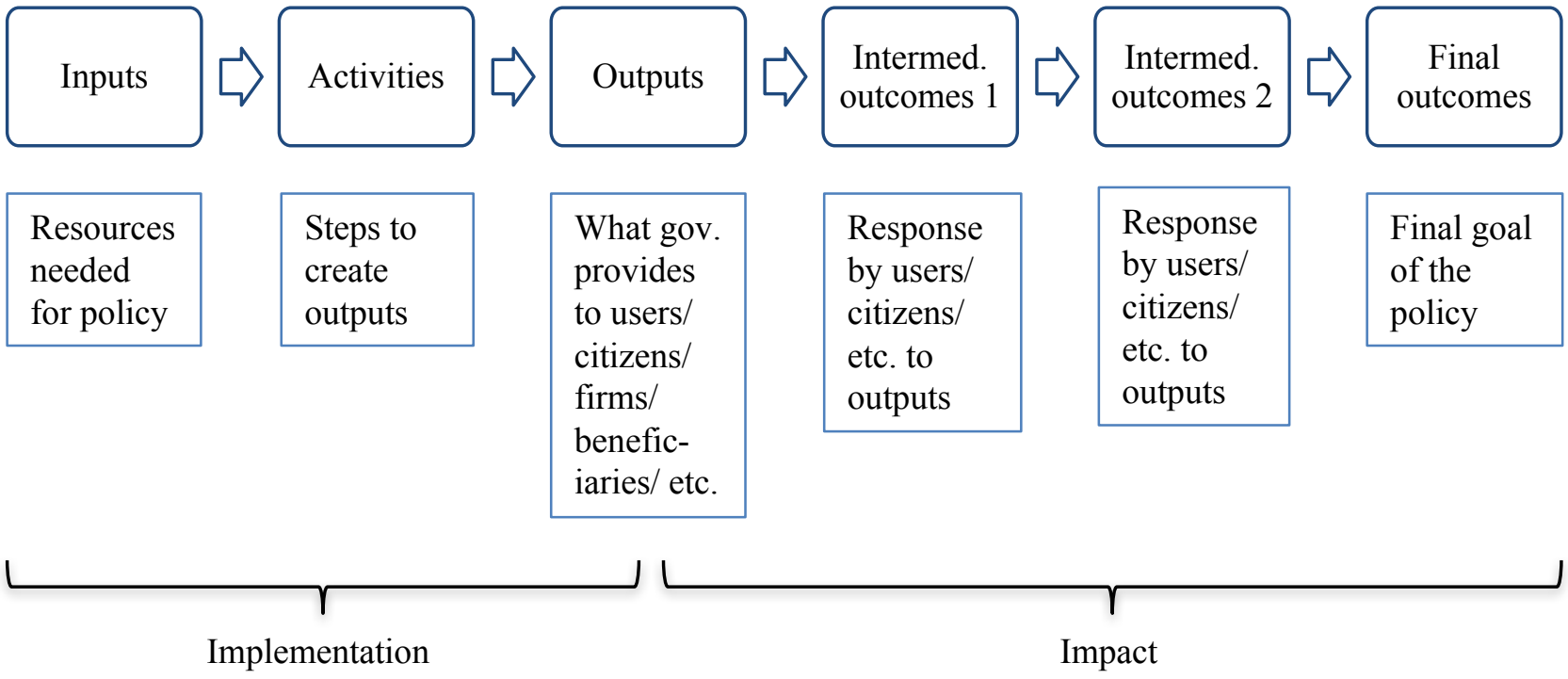
A theory of change



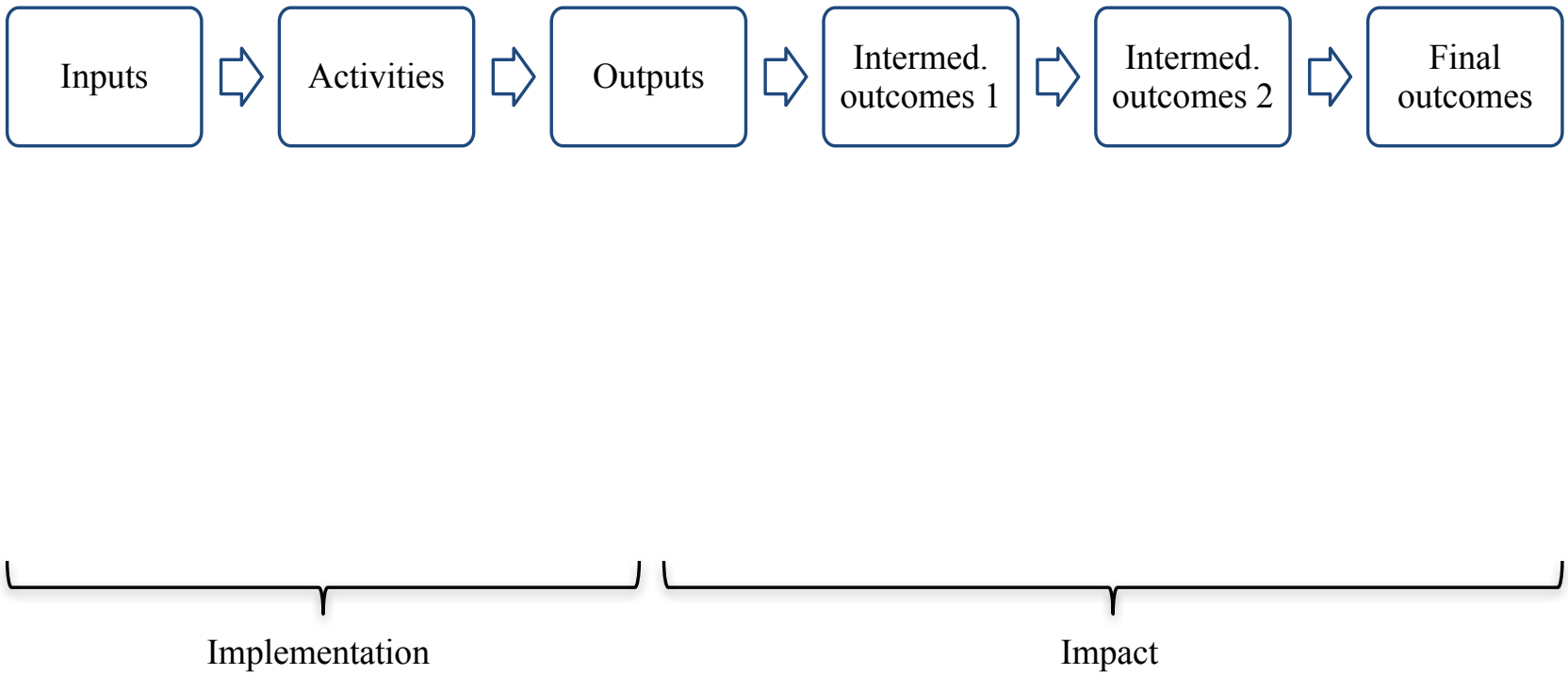
A theory of change



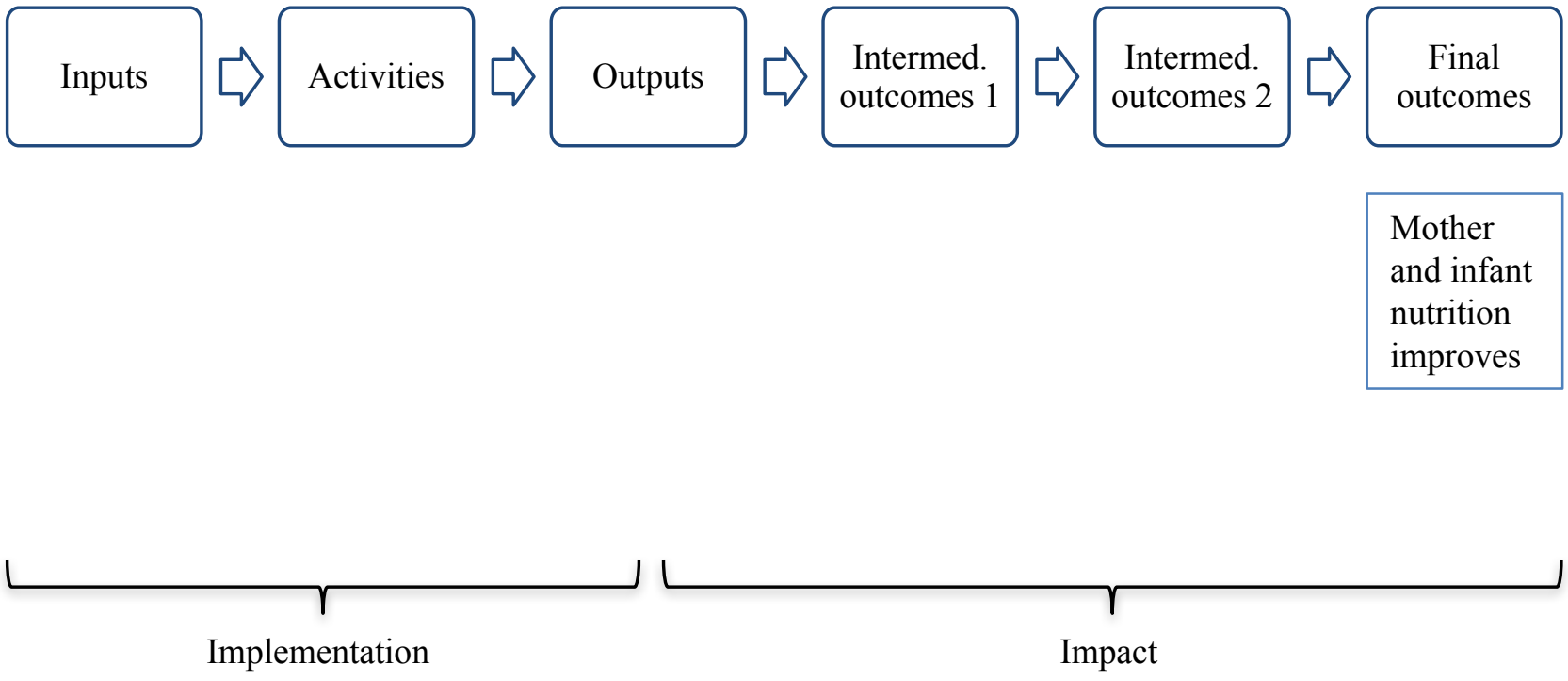
A theory of change



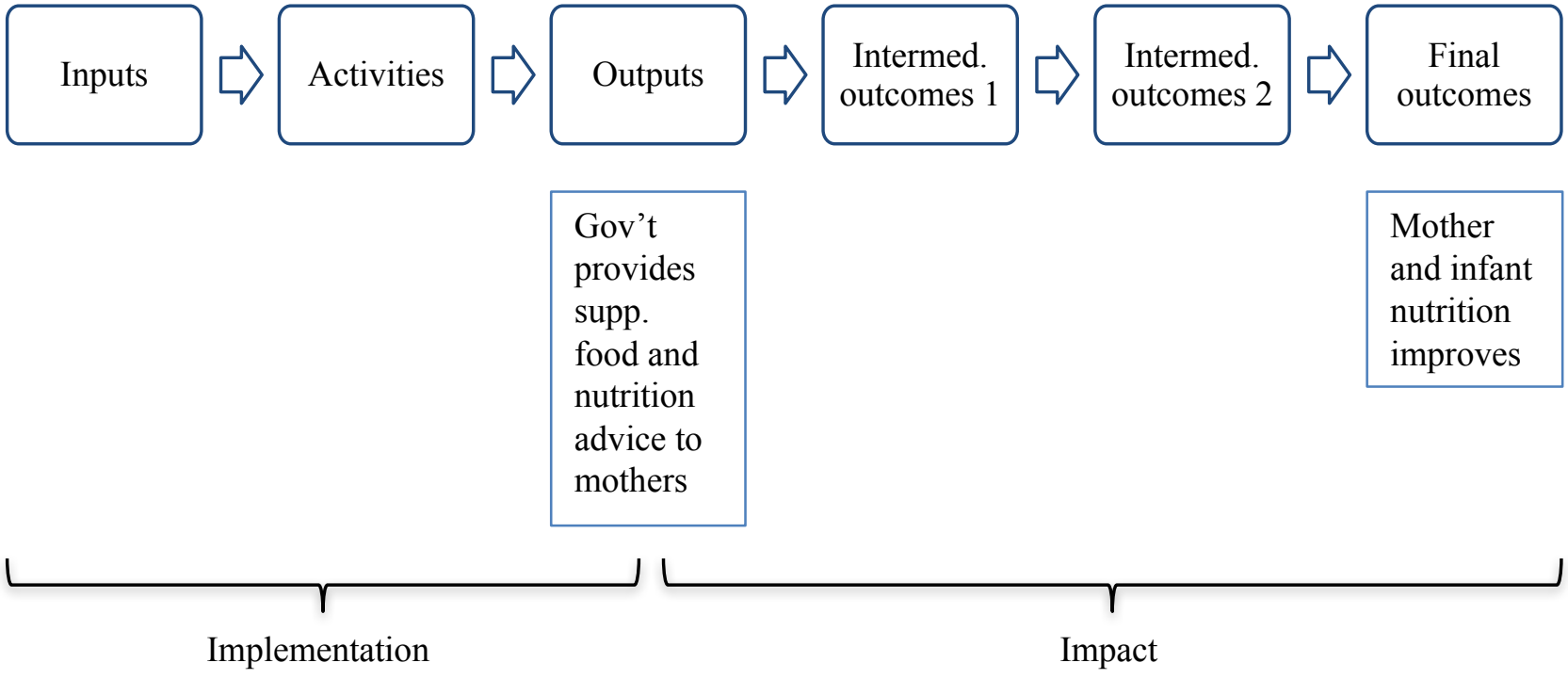
Example theory of change: BINP



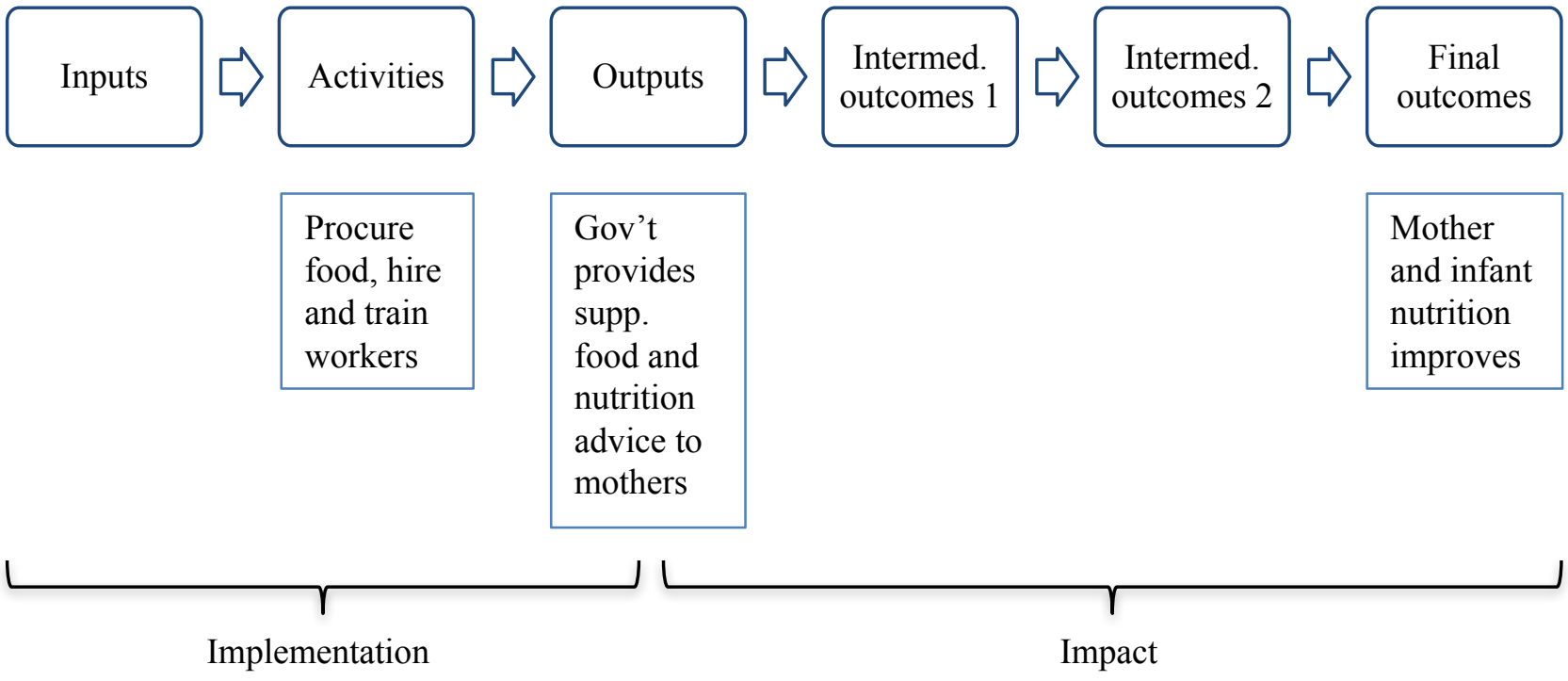
Example theory of change: BINP



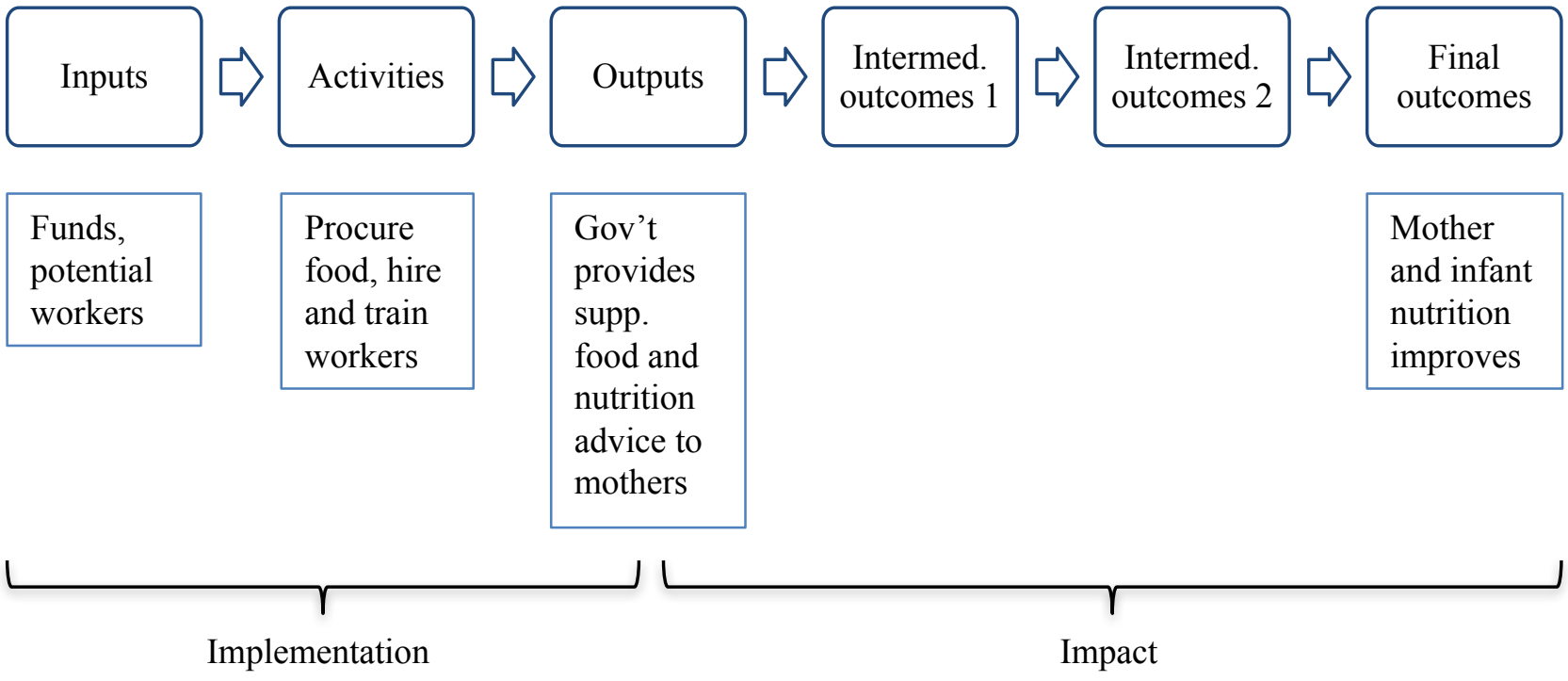
Example theory of change: BINP



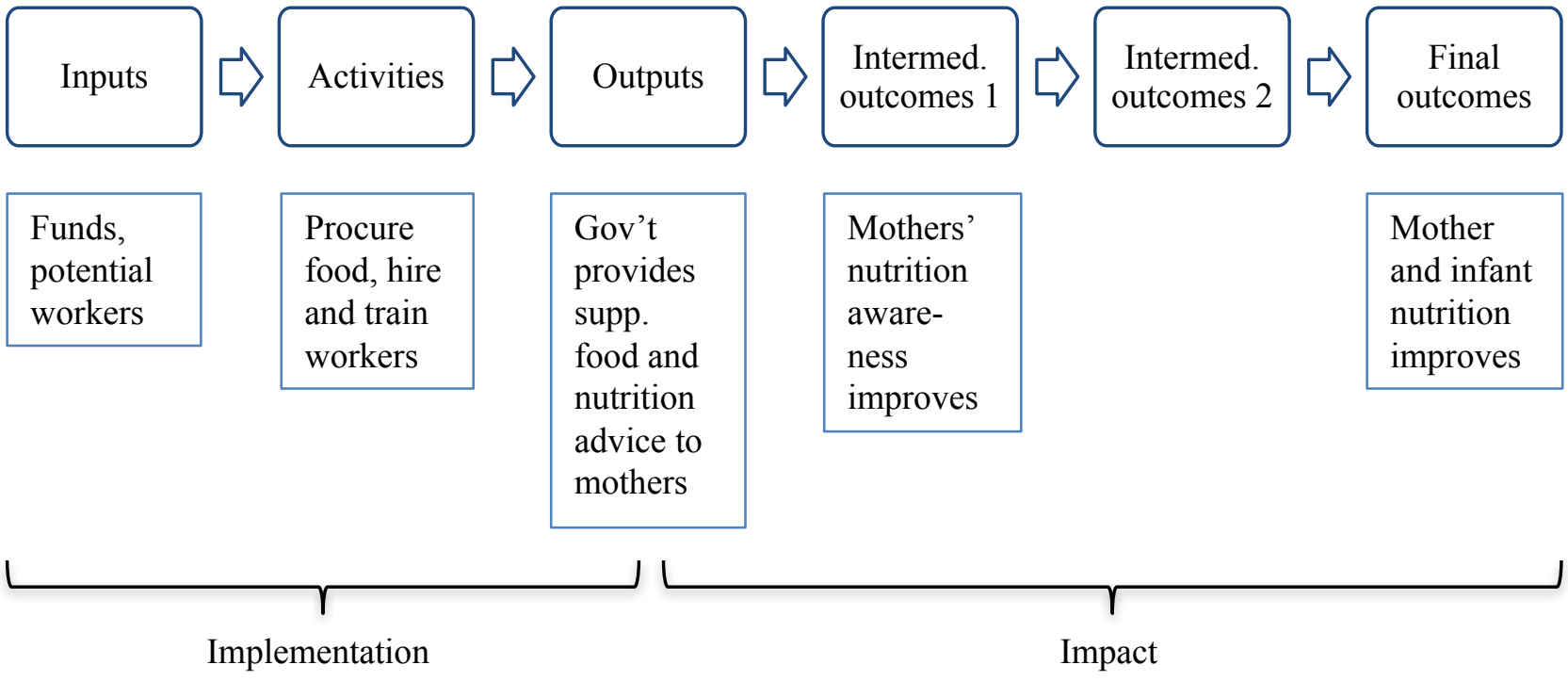
Example theory of change: BINP



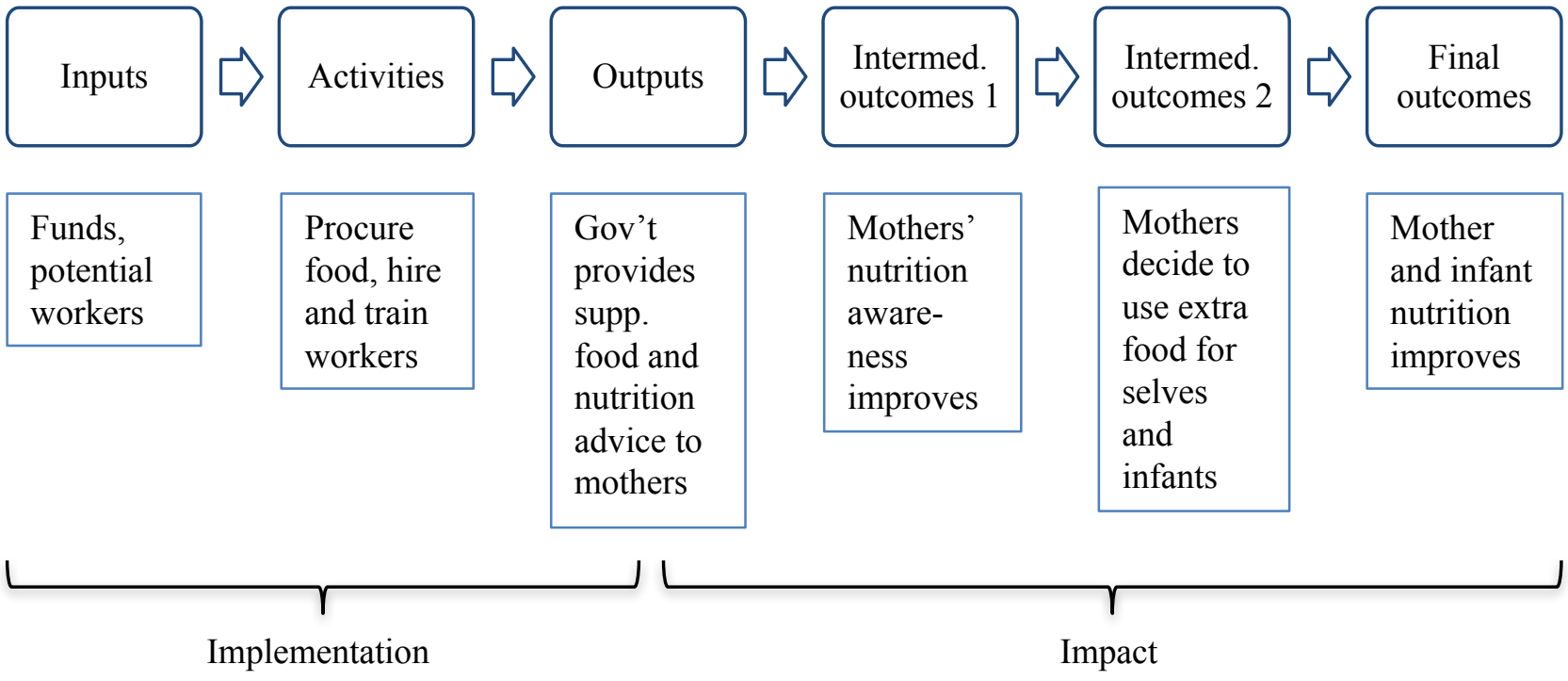
Example theory of change: BINP



Example theory of change: BINP



Example theory of change: BINP



Context

A policy's context is. . . the context in which it is being implemented

Context has many dimensions:

- Location (geographical, social, economic, cultural differences)
- Target group(s)
- Time
- Existence of related policy interventions (including spillovers)
- Who is implementing the policy (and how well)

Theory of change *x* Context

The same theory of change can operate in different contexts

A difference in context doesn't mean that the same policy can't be effective

The same policy can have different effects in different contexts **if** part of the theory of change **interacts** with a difference in context

What can we do about it?

Good news and bad news. . .

Bad news: no simple answer

Good news: 4 different approaches

- Not mutually exclusive

Approach 1: Systematic review and meta-analysis

See whether a policy has worked in many different contexts

- If yes, then there's a better chance it will work in this context

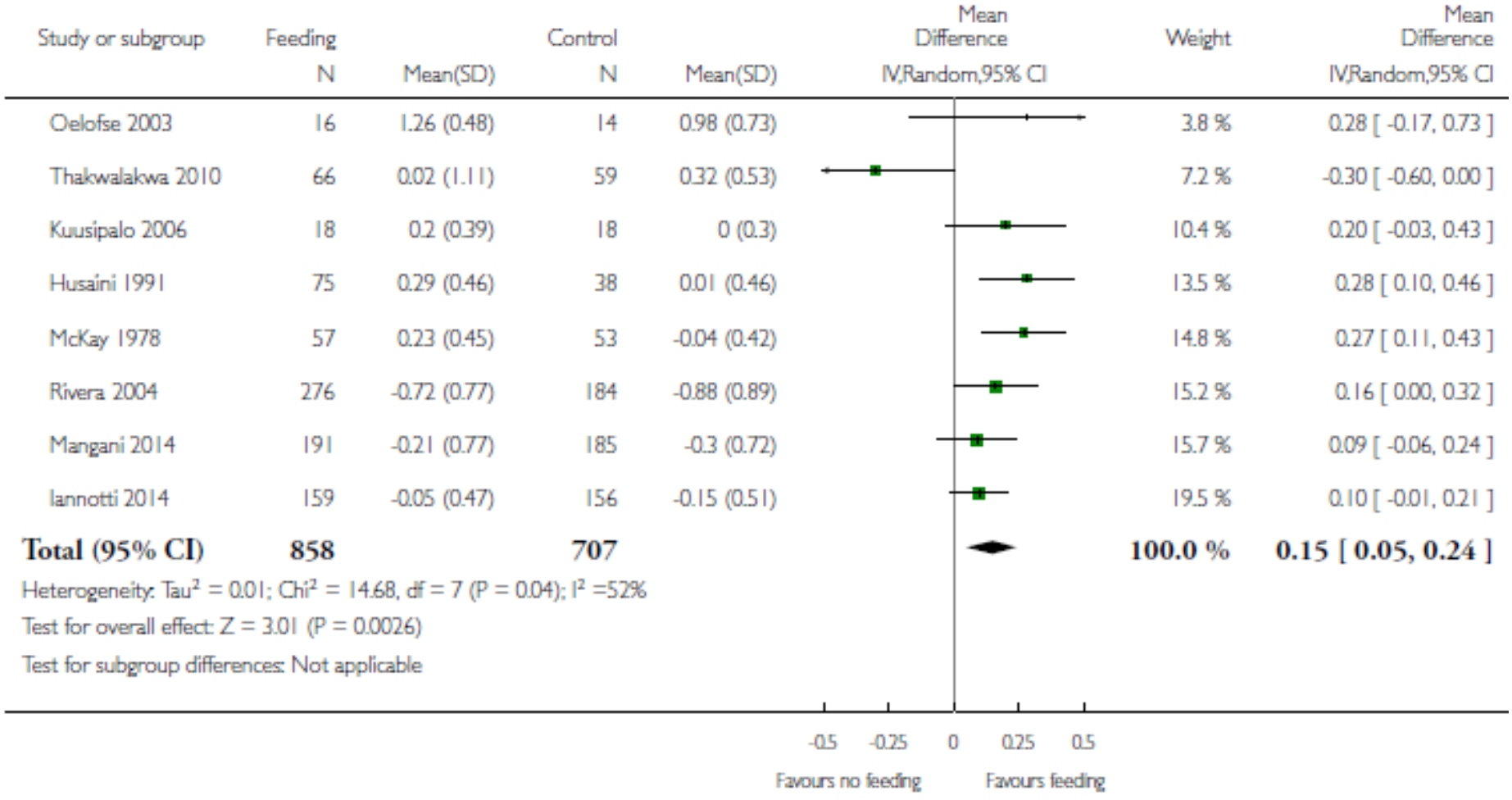
Food supplement programs for young children: researchers found 32 studies

- Supplementary feeding in low/middle-income countries usually had positive effects on child growth
- No clear effect on cognitive development

Note: this is the *average* effect of the interventions across all the contexts in which they've been evaluated

Kristjansson *et al* 2015, "Food Supplementation for Improving the Physical and Psychosocial Health of Socio-economically

Forest plots (output of meta-analysis)



Approach 2: Sub-group analysis

Sub-group analysis examines variations in policy impact across different groups within the population

- Works within a single evaluation, or with meta-analysis

Compare effect sizes across sub-groups, by:

- Age, sex, urban/rural, poverty level...
- Any other dimension of context you can measure

Use this information to assess:

- What dimensions of context are likely to matter
- How they interact with policy mechanisms

Sub-group analysis with nutrition programmes

Meta-analysis shows positive average impact

- But with a lot of heterogeneity in effects

More effective for younger and poorer children

- No effects in high-income countries
- Mixed results by sex

Variation in leakage by delivery method

- At home, children only received 36% of the supplement
- In day cares or feeding centres, rose to 85%

Based on this, where are nutrition programmes likely to be more effective?

Strengths and limitations of sub-group analysis

Gives some insight into heterogeneous effects

- For which groups / in which contexts is it likely to work better?
- Which context-mechanism interactions seem important?

But can't ever cover *all* dimensions of context

- And details of policy design/implementation might be important

Approach 3: Judge similarity of contexts

The more similar the contexts, the more likely a policy is to have a similar effect

What are some important aspects of context you could judge?

Strengths and limitations of judging context similarity

Before implementing, it's often unclear what the relevant dimensions of context are

Remember: contextual differences are only a problem if they interact with the policy mechanism

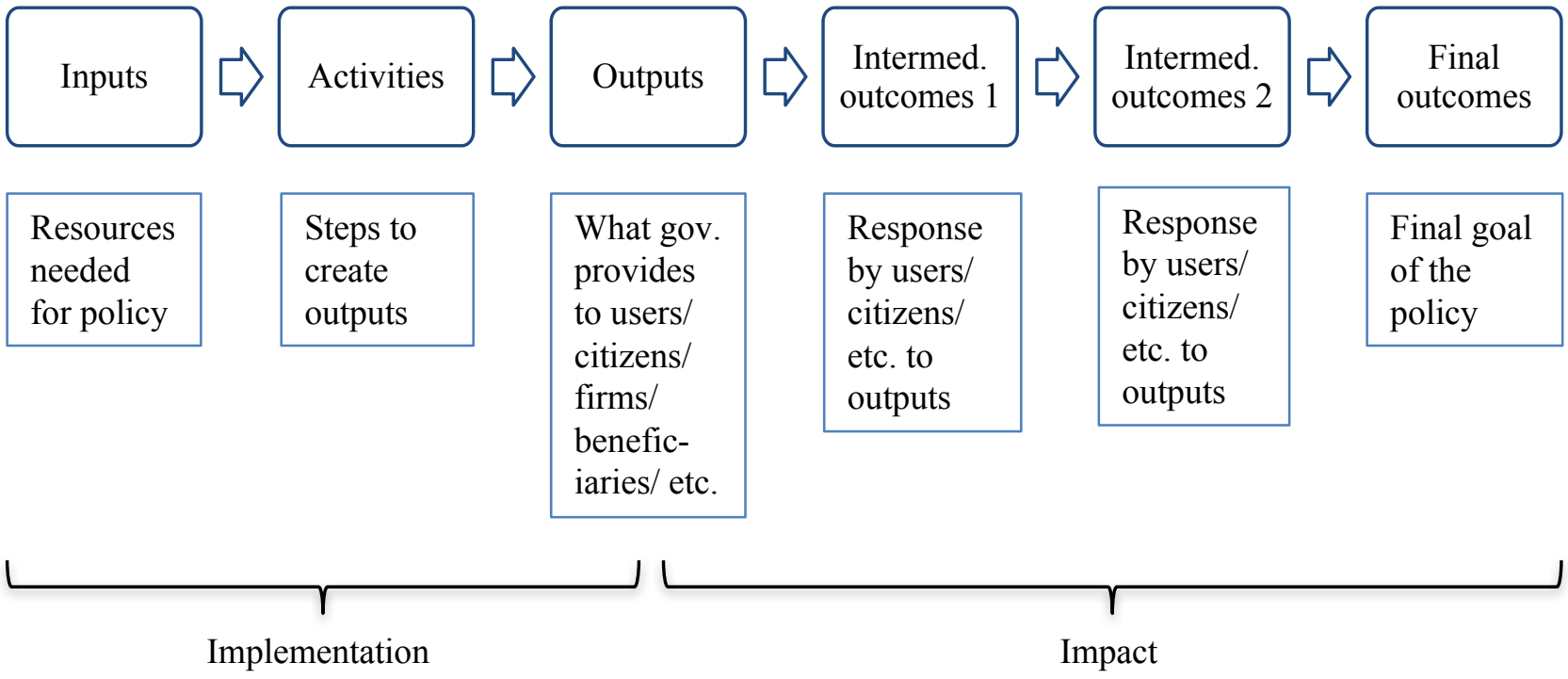
Approach 4: Mechanism mapping

Mechanism mapping is a way to pinpoint *which* aspects of a policy are likely to be problematic in a new context

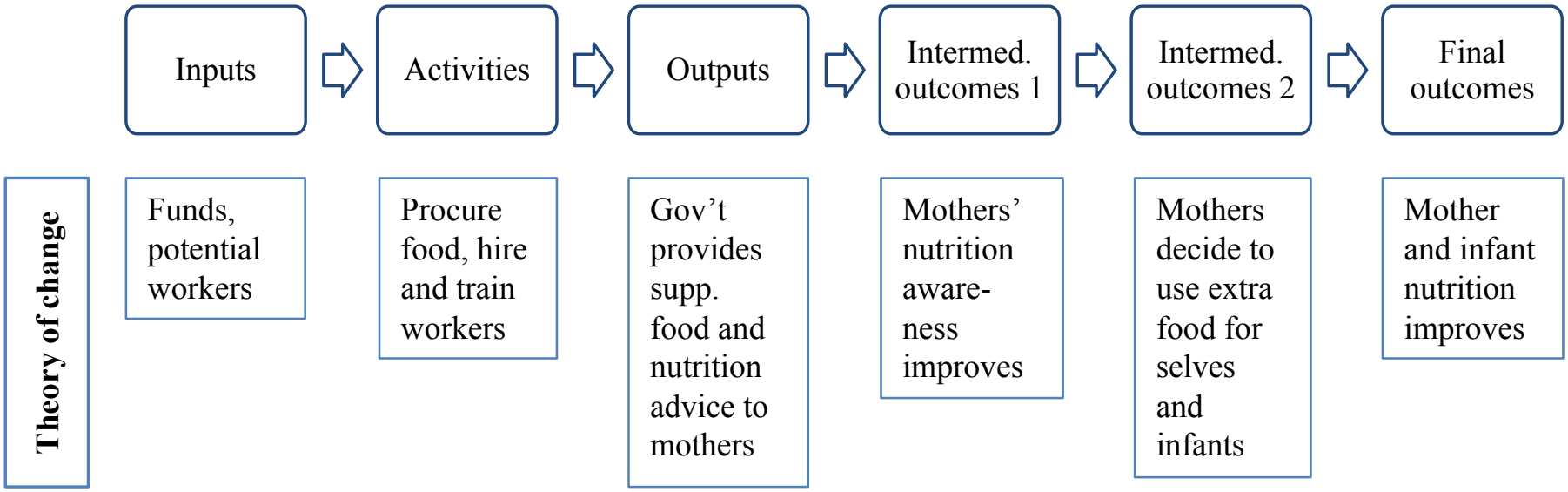
Five steps. Today we will look at the first three:

1. Map out a policy's *theory of change*
2. Identify the *contextual assumptions* necessary for each step of the theory of change
3. Compare this to the *actual characteristics* of your context

Step 1: A theory of change



Example: BINP & TINP



How to make a theory of change

Good starting point (where possible): ToC from an existing evaluation/review

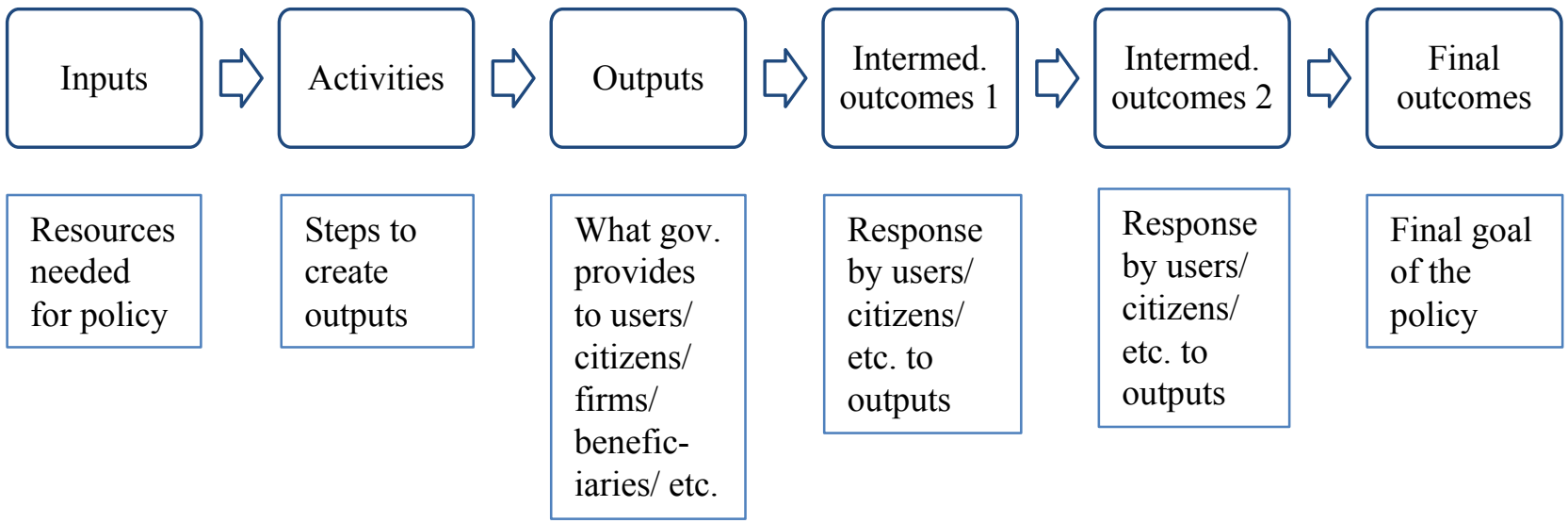
Most ToCs are simple and linear, but can add:

- Multiple activities, outputs, outcomes
- Multiple competing mechanisms
- Complementarities, feedback mechanisms
- Negative outcomes

Making a theory of change

With the person next to you:

- Pick a policy that exists somewhere outside Brazil
- Write a theory of change for it
- 5 minutes
- Don't choose your group project policy: we'll do that later



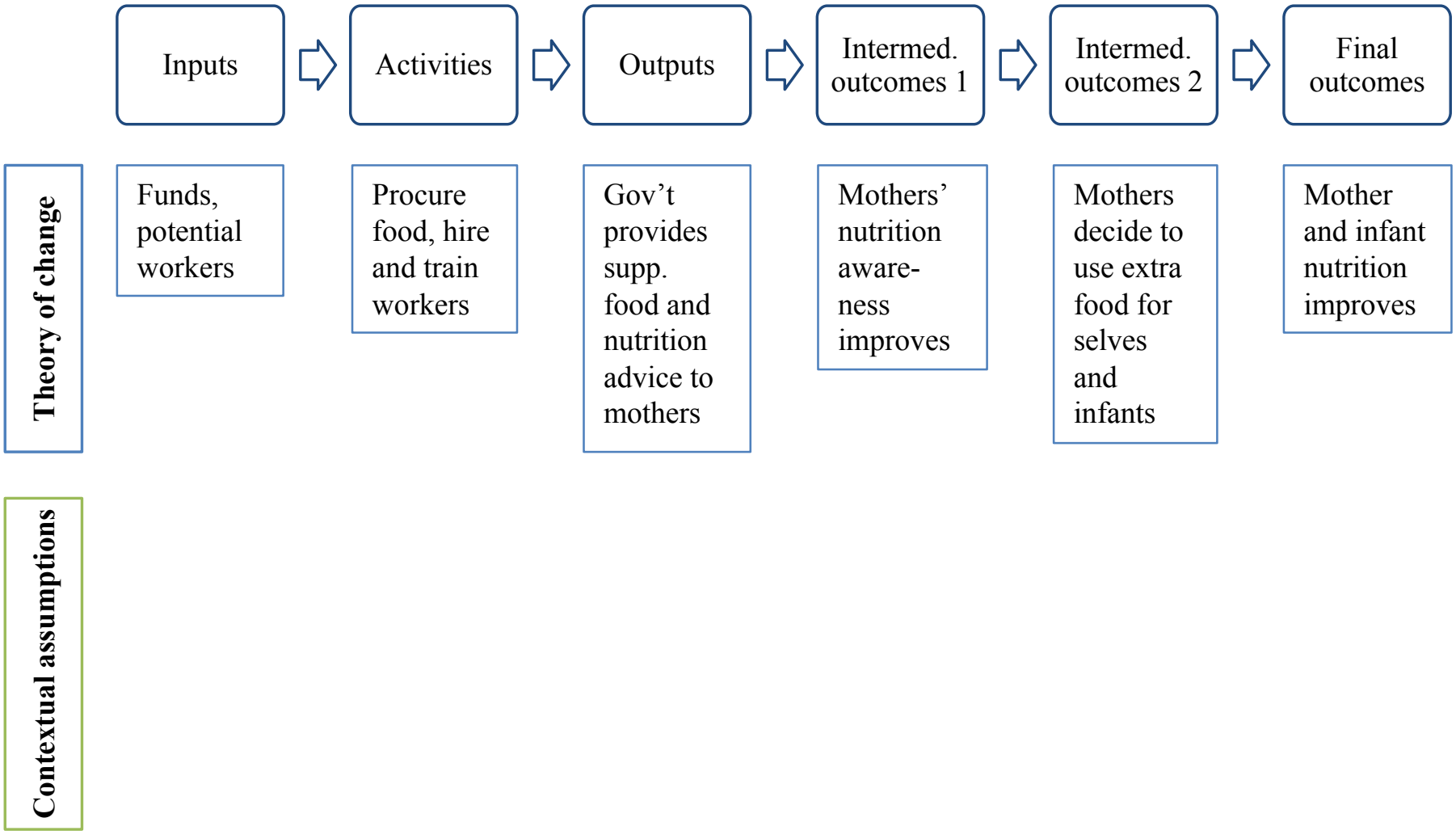
Step 2: Contextual assumptions

What does the theory of change assume about the context? For example:

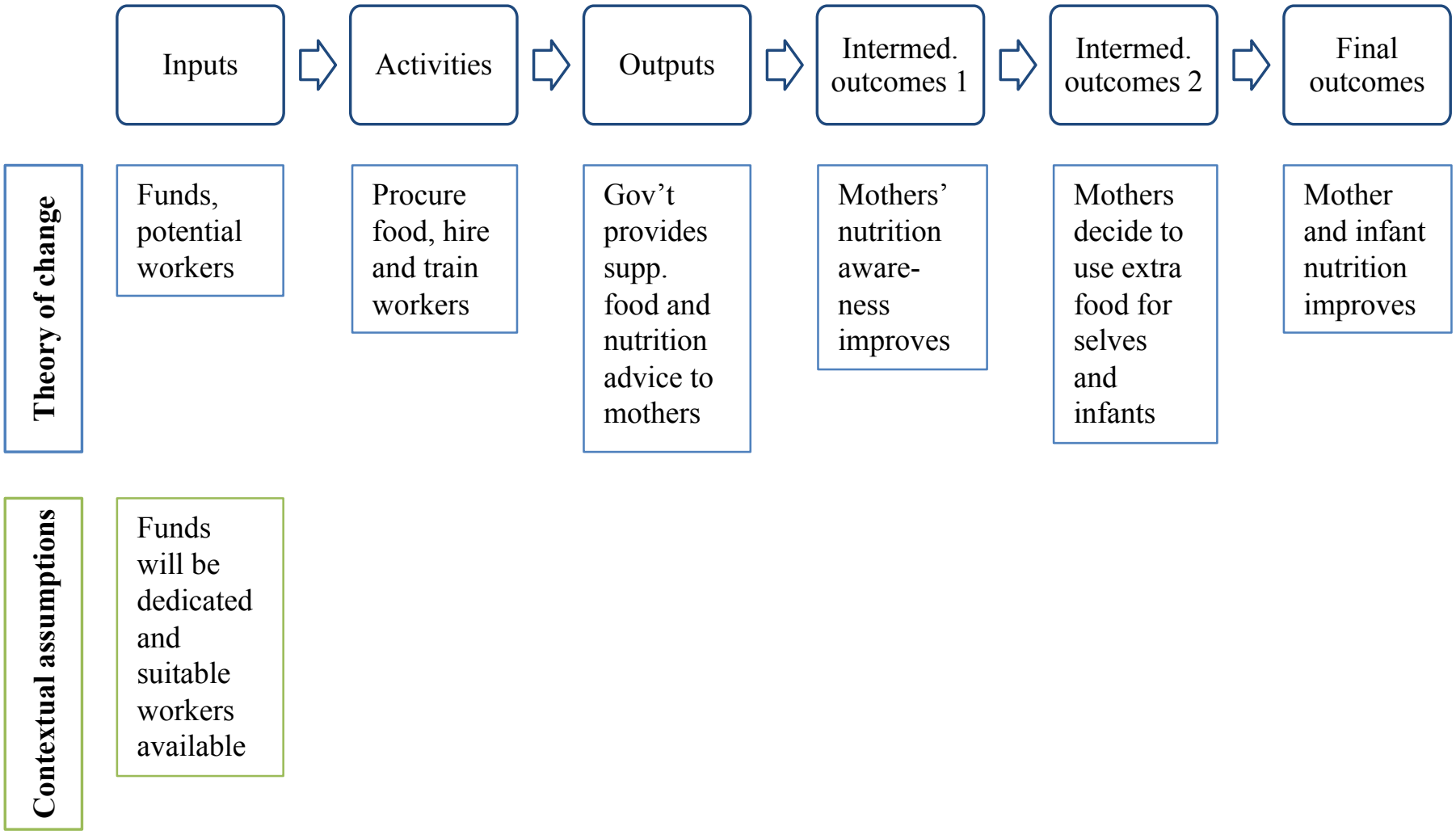
- Availability of resources
- Personnel with technical skills
- Coordination: all activities in a project will come together at the right time
- Service users will use the service the way you expect
- Having the service is enough to achieve the outcome
- Infinite number of possibilities. . .

Step two in mechanism mapping is to map the most salient contextual assumptions

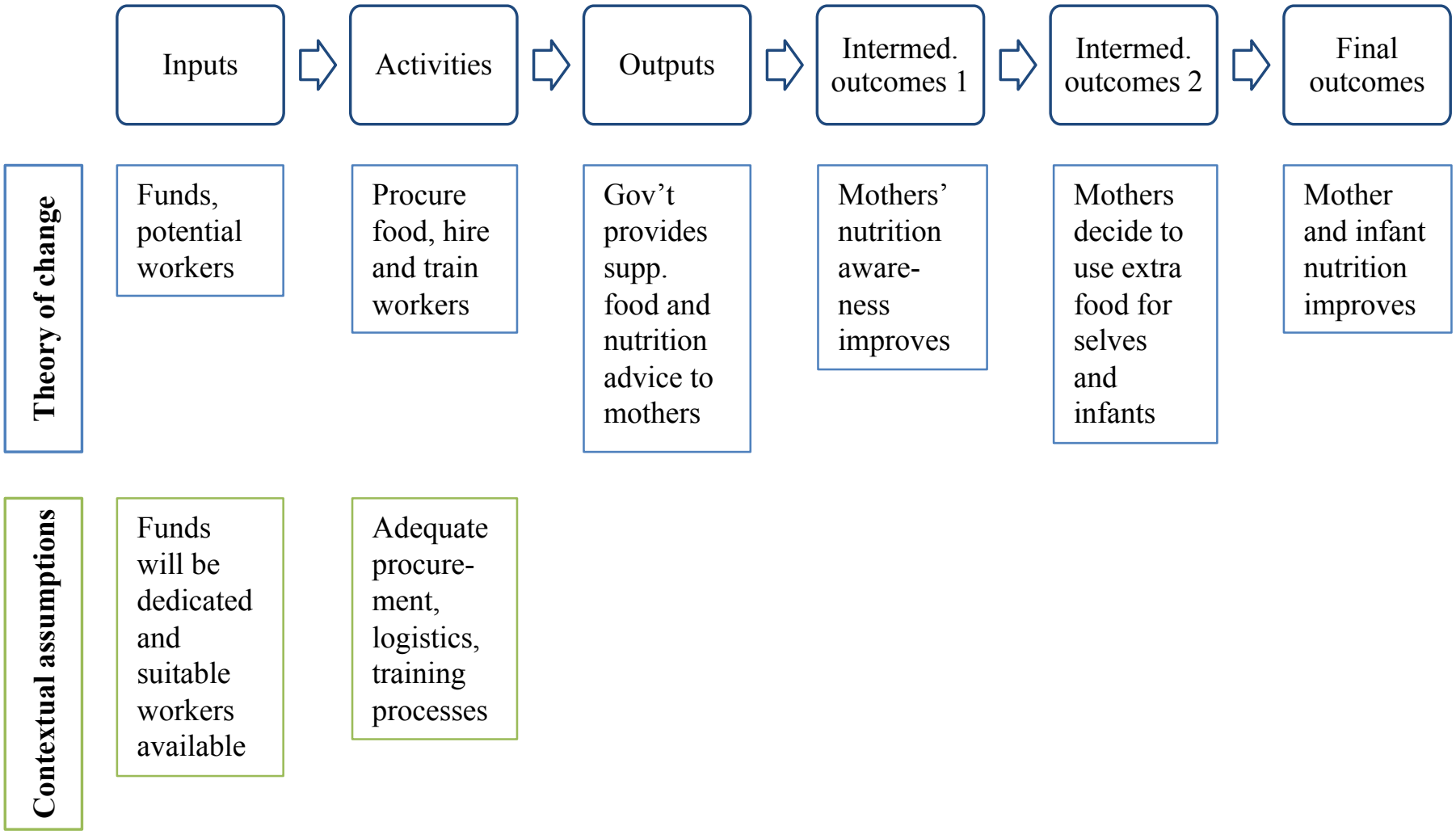
Contextual assumptions for BINP & TINP



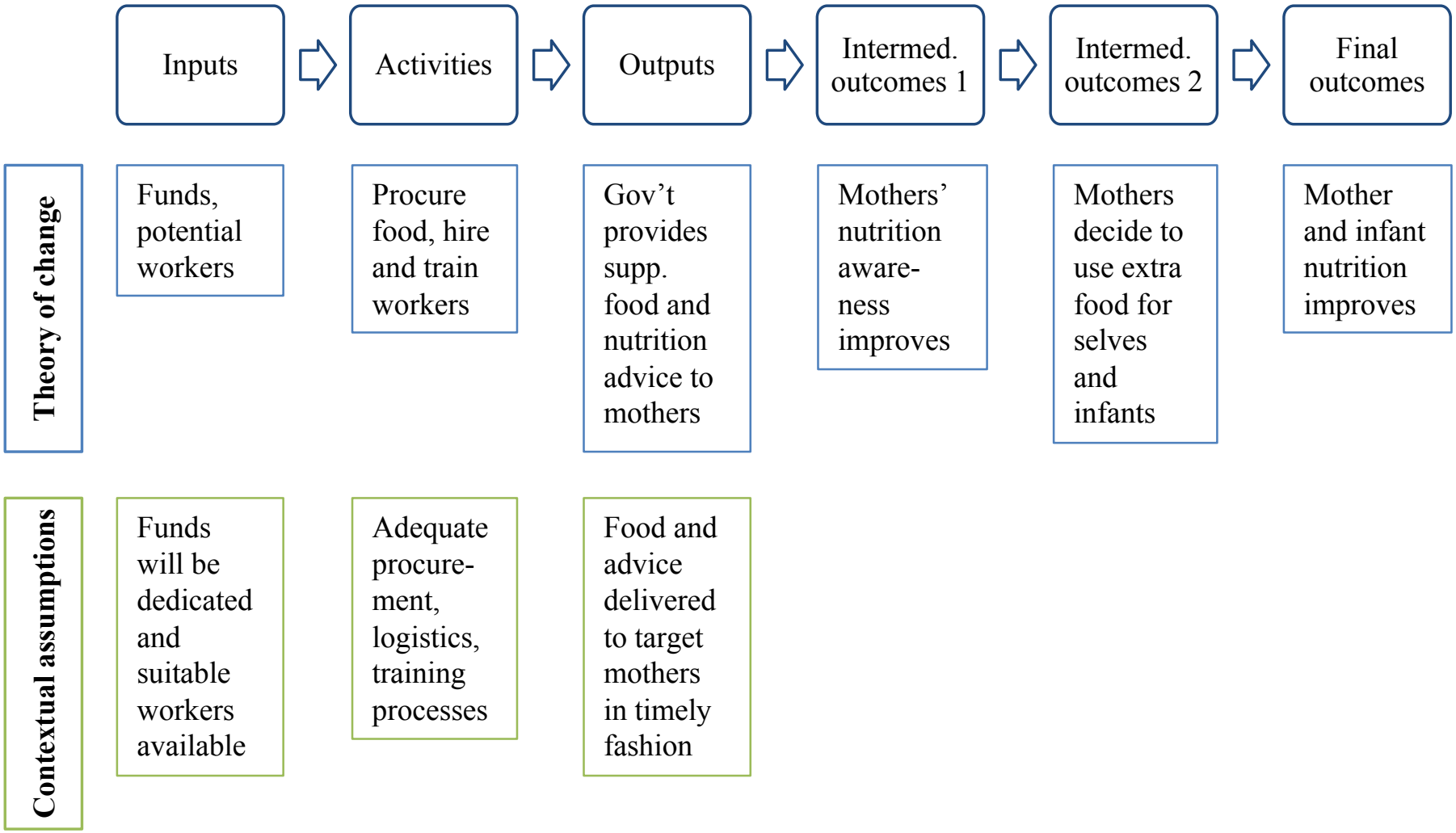
Contextual assumptions for BINP & TINP



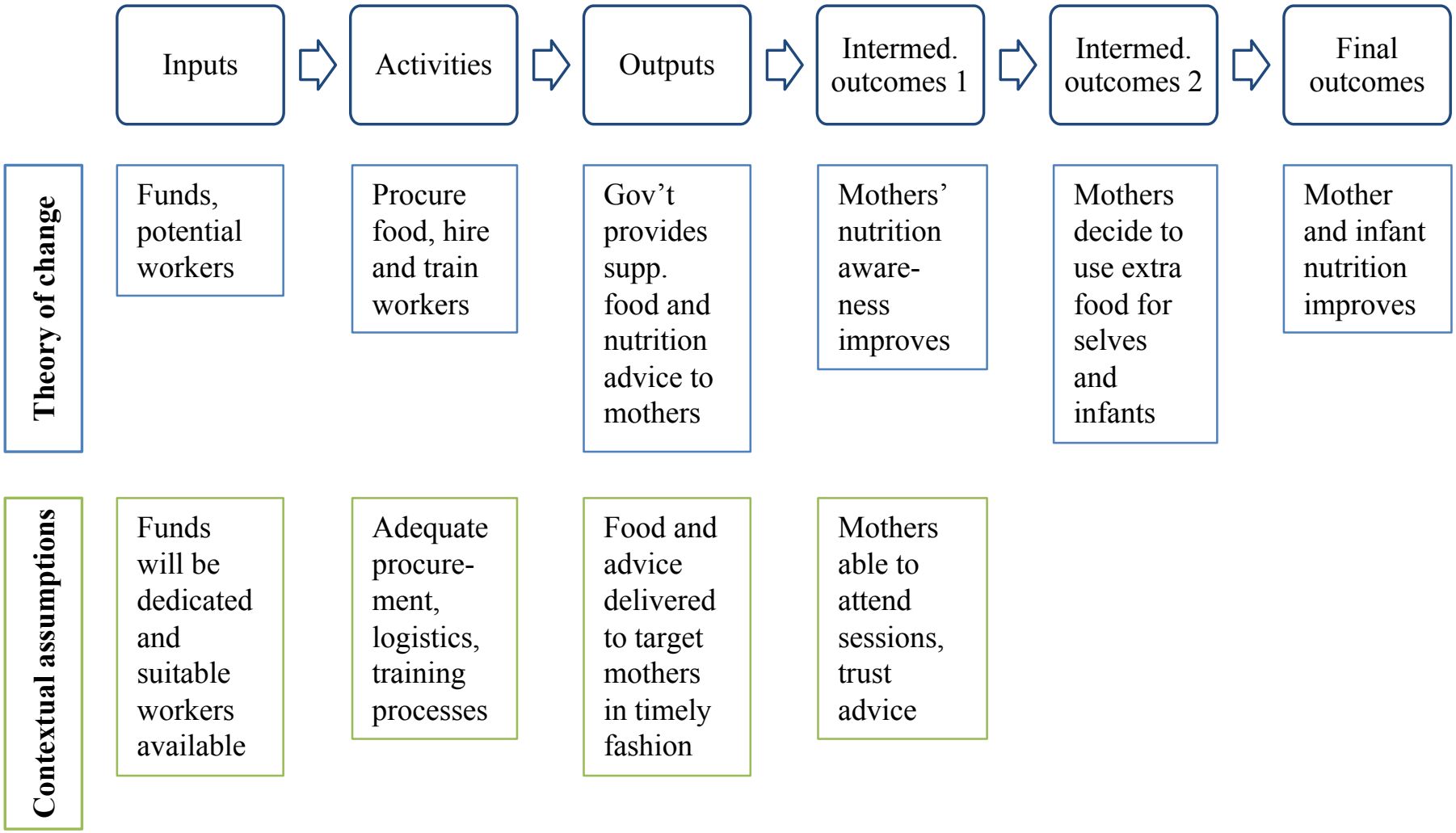
Contextual assumptions for BINP & TINP



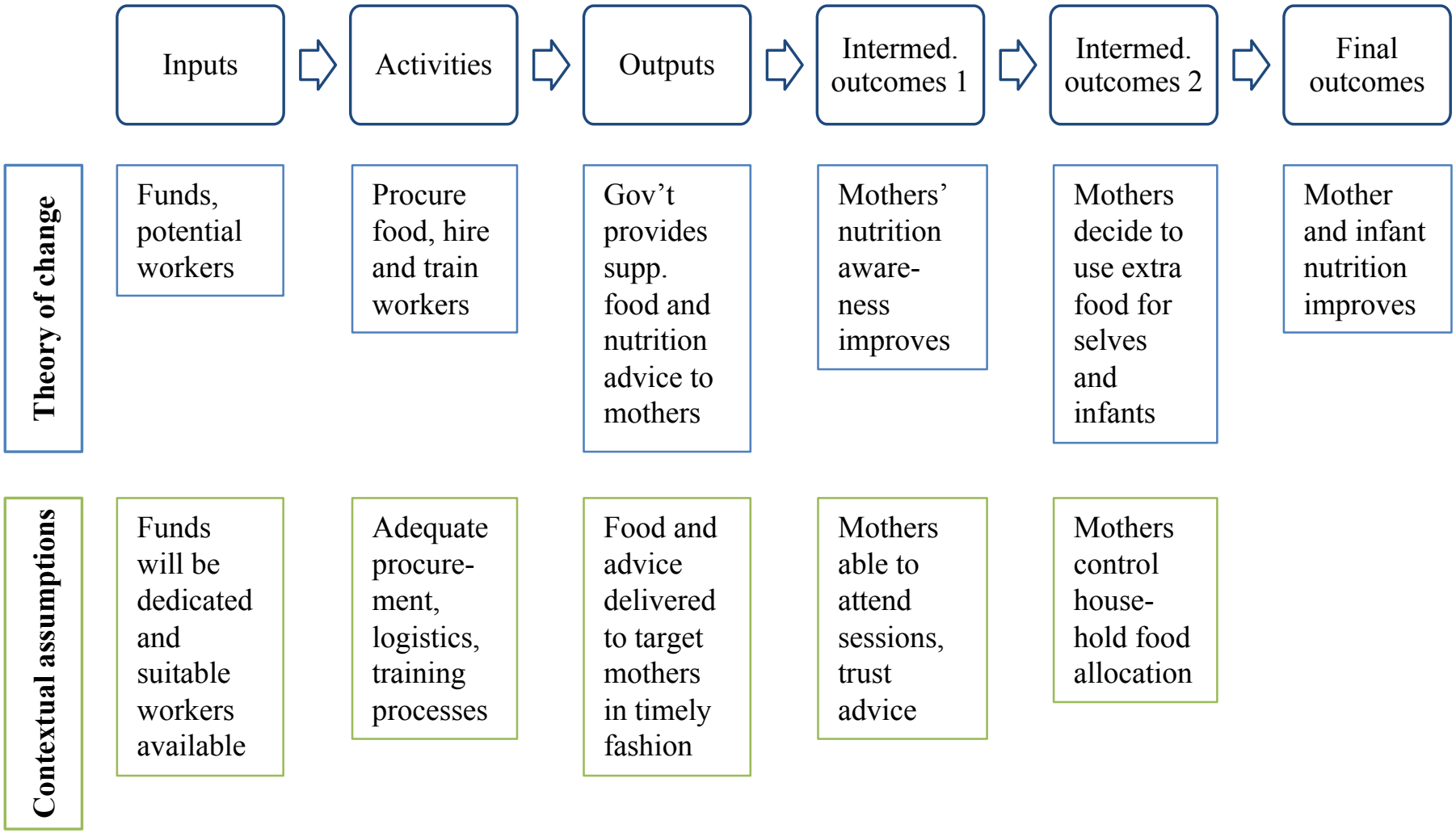
Contextual assumptions for BINP & TINP



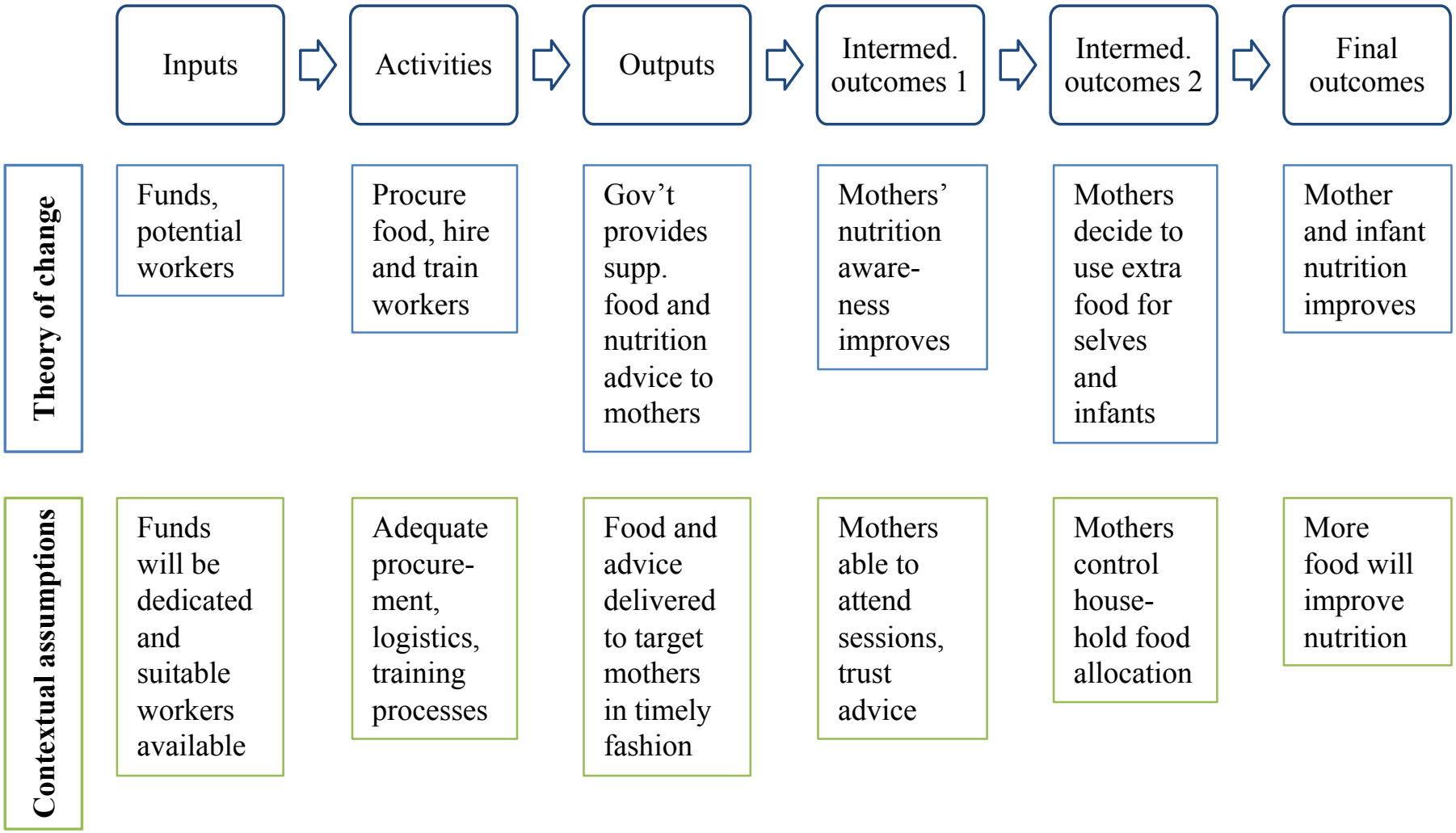
Contextual assumptions for BINP & TINP



Contextual assumptions for BINP & TINP



Contextual assumptions for BINP & TINP



How to identify contextual assumptions

1. Sub-group analysis from other studies/reviews
2. Some dimensions are commonly important
   - demographic and socioeconomic variables; resource availability; political support and resistance; social and cultural norms; implementing organizations' effectiveness; potential for corruption; etc.
3. Inspect theory of change for salient aspects of context
   - e.g. household decisionmaking over food in BINP's ToC
4. Participatory policy processes: implementers and beneficiaries

Identifying contextual assumptions

With your partner, write out the *most important* contextual assumptions of your policy

- Each step of a theory of change includes assumptions
- 5 minutes

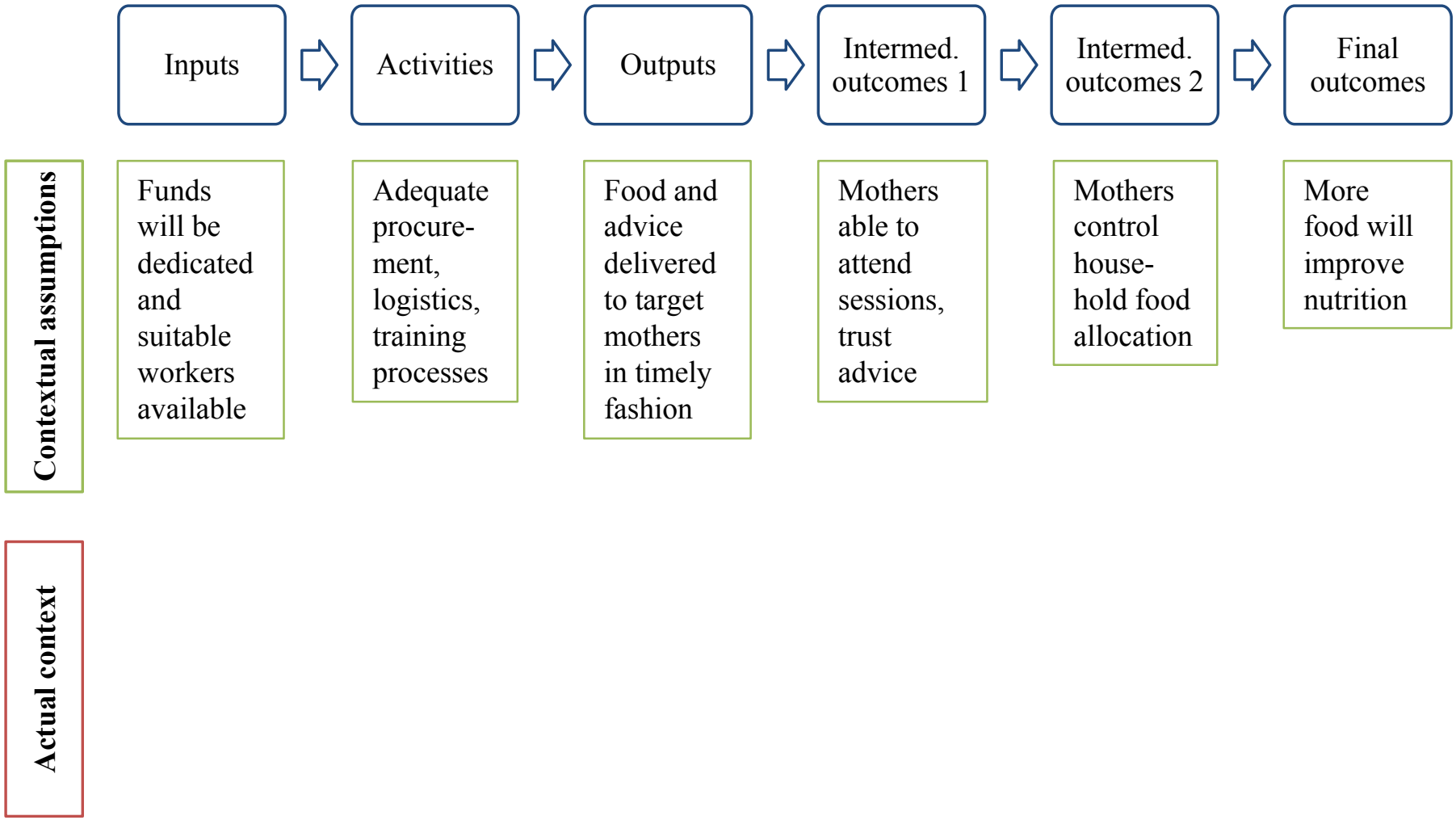
How did it go? What was easy about it? What was difficult?

Step 3: Contextual Characteristics

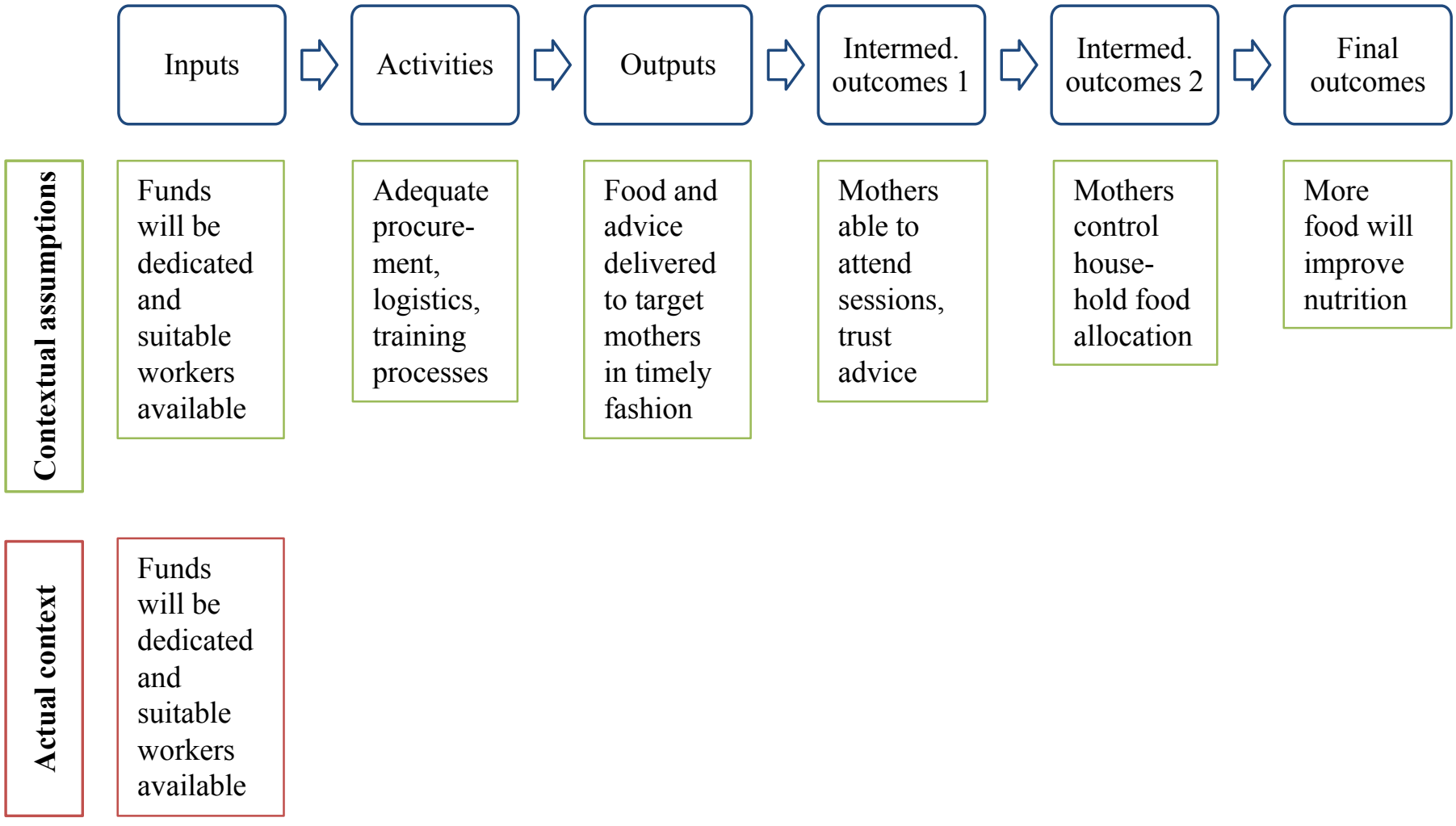
Mapping the theory of change against the contextual assumptions has shown us when/where we expect the policy to work

Now, let's compare those assumptions to the reality of our context

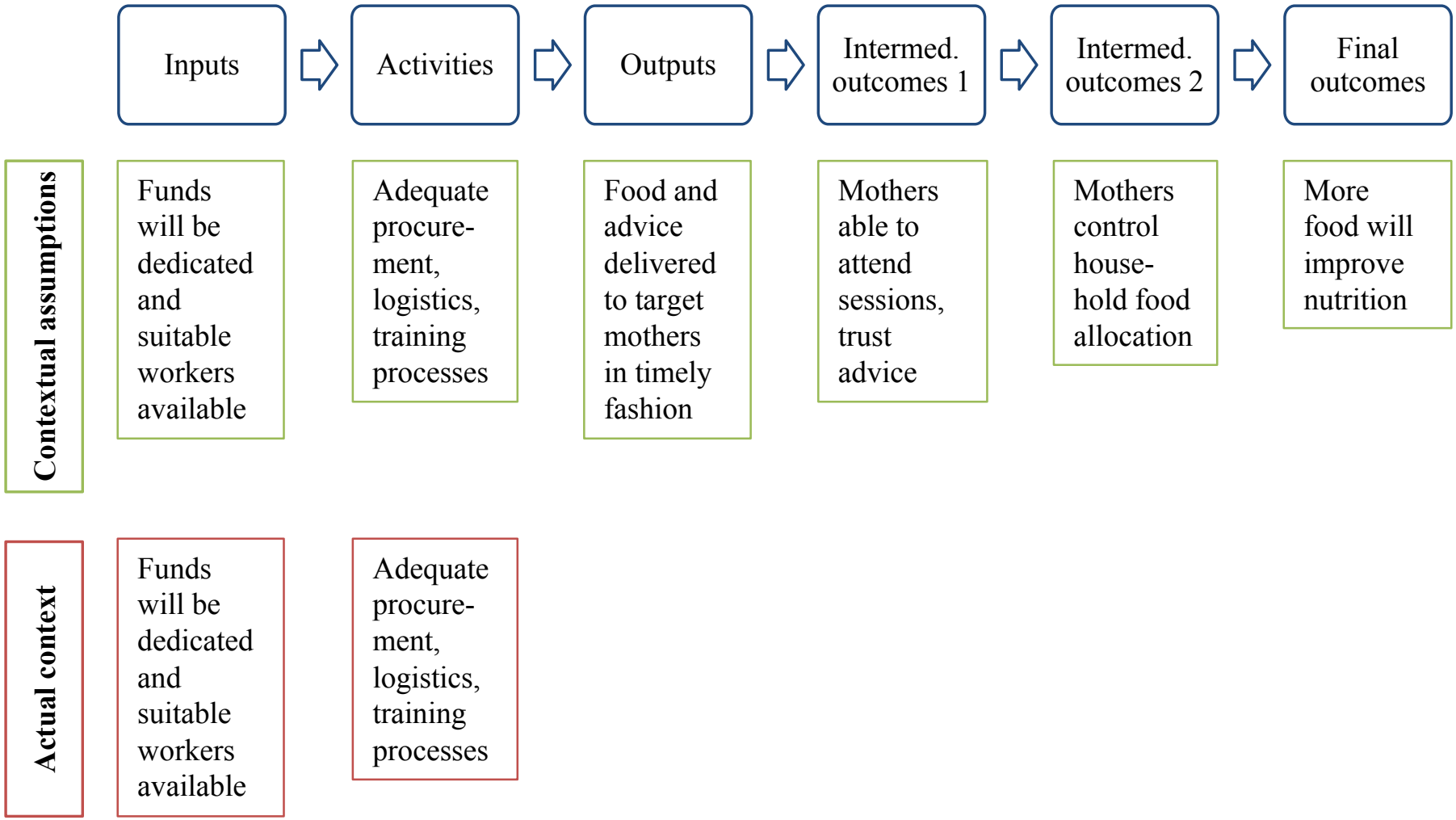
Contextual Characteristics: BINP



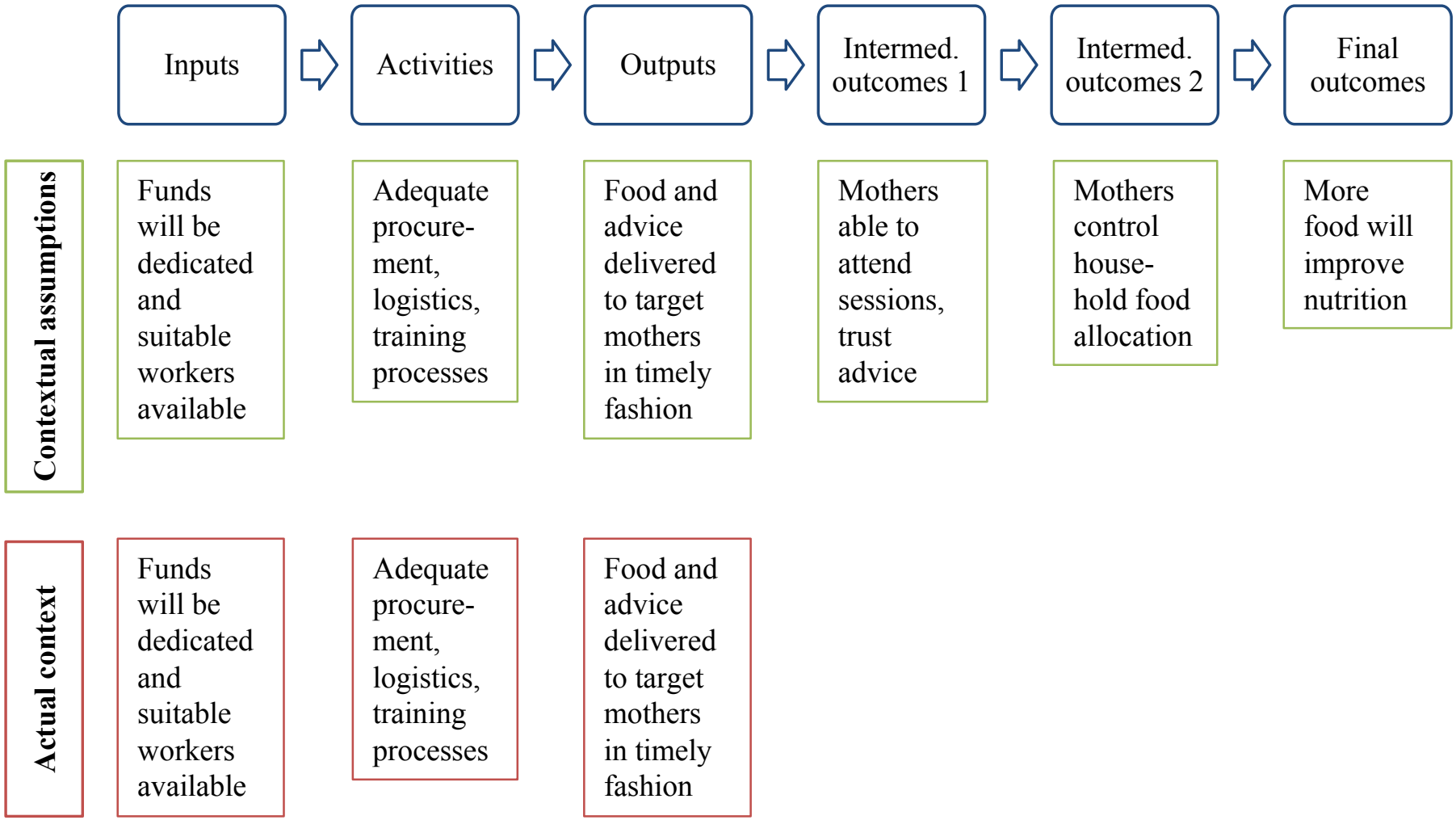
Contextual Characteristics: BINP



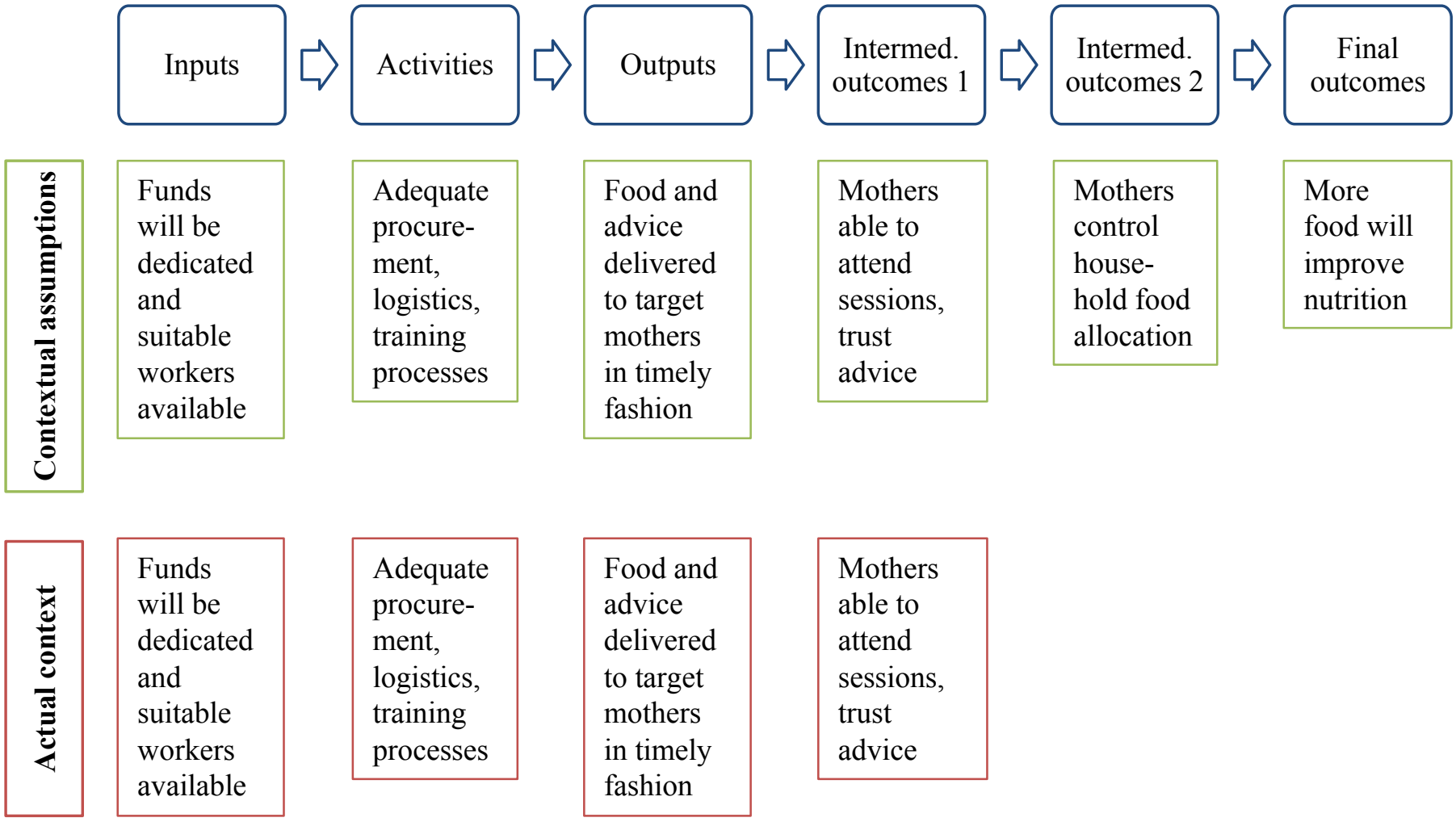
Contextual Characteristics: BINP



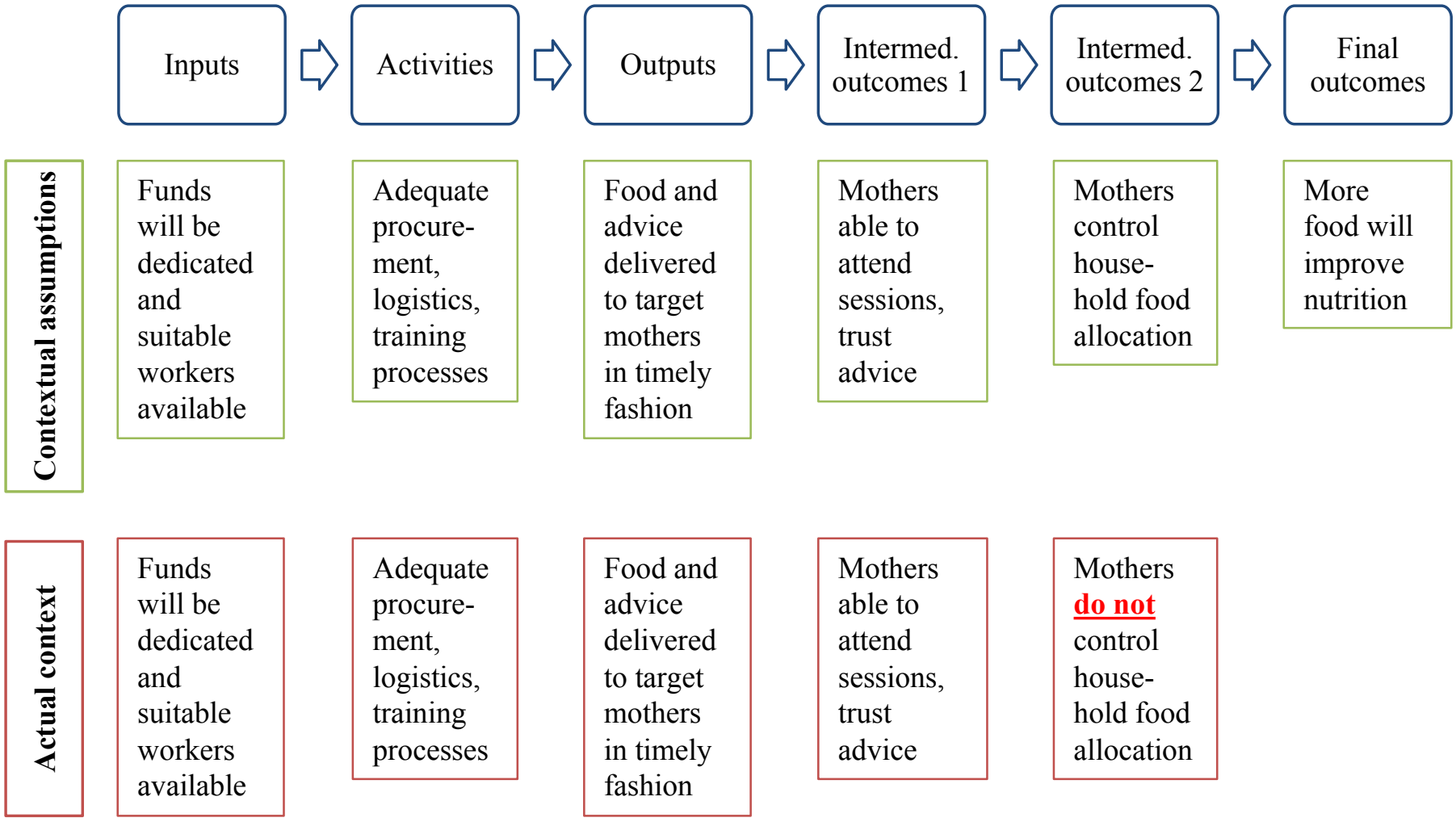
Contextual Characteristics: BINP



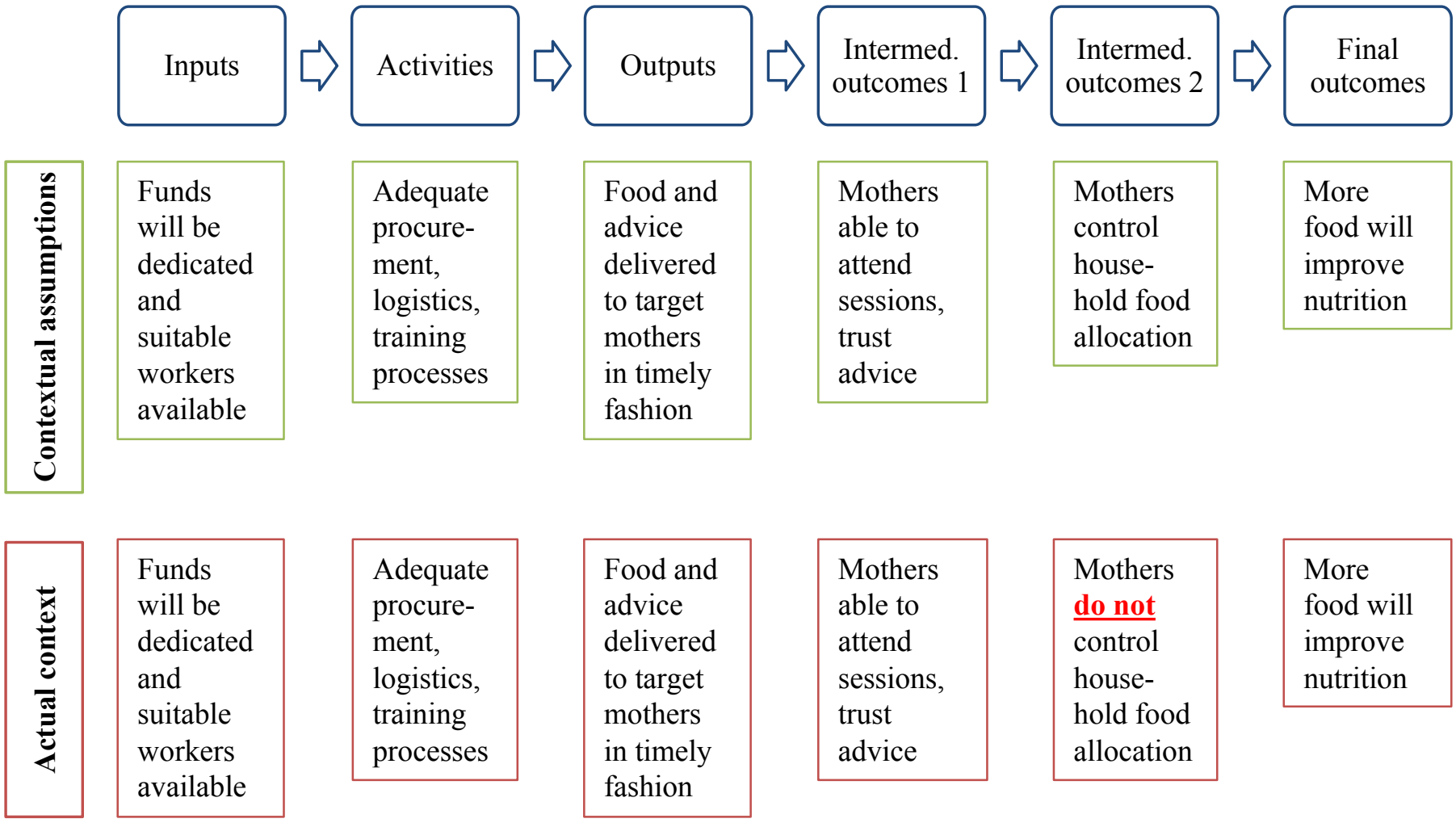
Contextual Characteristics: BINP



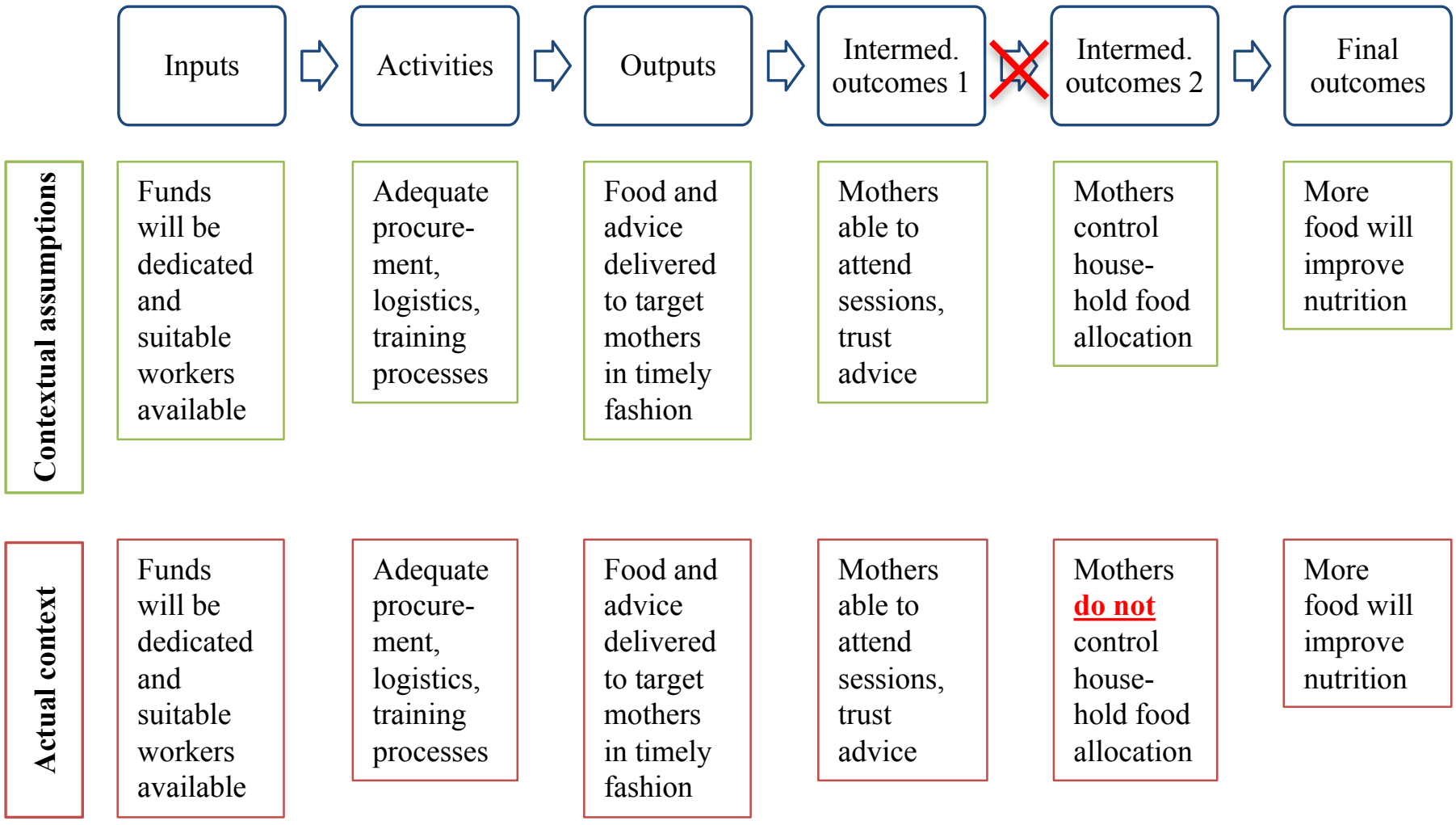
Contextual Characteristics: BINP



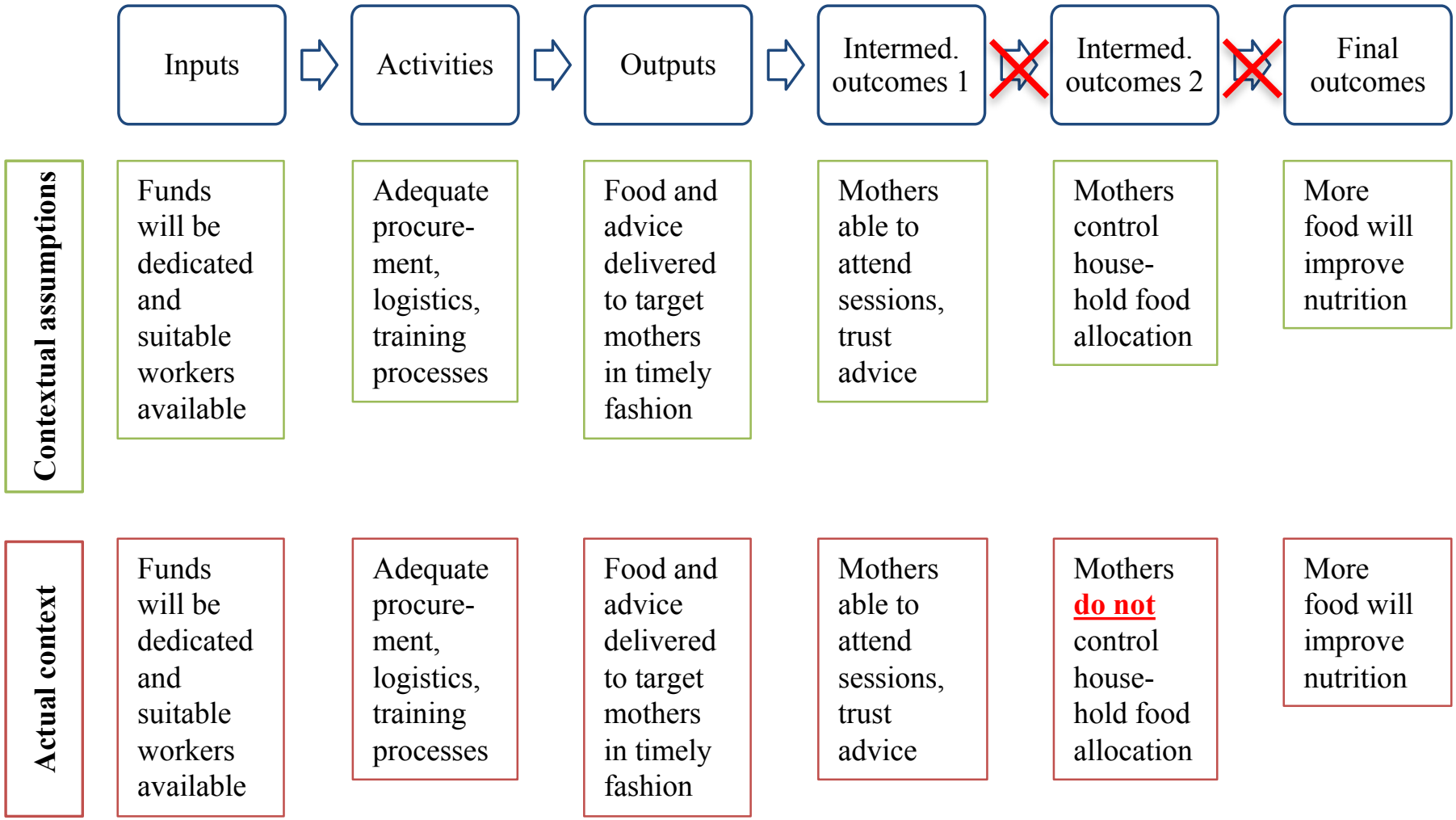
Contextual Characteristics: BINP



Contextual Characteristics: BINP



Contextual Characteristics: BINP



Contextual Characteristics

Assume that your ministry is considering implementing your group's policy at national scale in Brazil

How would the contextual characteristics of Brazil compare to the policy's contextual assumptions?

First, think about Brazil in general - i.e. the nationwide average

Then think about heterogeneity within Brazil

Finding and interpreting evidence on your group's policy

Reminder: sources for evidence

Cochrane Collaboration - health-focused

- www.cochrane.org

Campbell Collaboration - evidence-based policy more generally

- www.campbellcollaboration.org

3ie - development-focused

- http://www.3ieimpact.org

UK government “What Works Centres”

- https://www.gov.uk/guidance/what-works-network

Finding additional evidence on your policy

Use these evidence sources to find additional evidence on your group's policy

- A systematic review
- Similar impact evaluations from other countries
- A policy brief
- Media discussion (e.g. newspapers)
- Official material (government or NGO)

Recommend that each group divide up search responsibilities

Remember: finding evidence usually means combining multiple sources

How did it go?

Were you able to find resources?

What were the challenges?

Any tips for others?

How did different accounts of the policy differ?

- Did it have different impacts in other contexts?

Recap

External validity

- *Interaction* of *theory change* with *context*

Four ways to address external validity

1. Systematic review and meta-analysis
2. Sub-group analysis
3. Judge similarity of contexts
4. Mechanism mapping

Finding and interpreting evidence

Tomorrow

Tomorrow we will think about how to adapt policies from elsewhere to better fit the local context

Group presentations

Evidence in Policymaking in Brazil

Panel discussion:

- Fernando Sertã - General Coordinator of Evaluation and Public Policies / Finance and Budget Secretariat / Ministry of Development and Management
- Miguel Crisóstomo Leite - General Coordination for Policy and Program Evaluation / Planning Secretariat / Ministry of Development and Management
- Daienne Machado - ENAP
- João Sigora - Ministry of Social Development

From Impact Evaluation to Policy Design:

Evidence, External Validity, and Effective Policymaking

Martin J. Williams

Blavatnik School of Government, University of Oxford

ENAP, 4-8 June, 2018

Today

Monday: did a policy work elsewhere?

- Impact evaluation, systematic reviews

Thursday: from "it worked there" to "will it work here?"

- Systematic reviews, mechanism mapping

This morning: whether and how policies should be modified when transporting them to a new context or scaling them up

Plan for today

1. Fidelity vs adaptation
2. Adapting policies
3. Mechanism mapping of group policies
4. Presentating group work

The complexity of policy interventions

Even simple policies are complex (i.e. have many dimensions)

- Eligibility
- Duration
- Dosage
- Timing
- Implementing organization and personnel
- And many more

And these dimensions can interact with each other

Details, Details, Details...

Designing policies that work is HARD

- And many policies that sound sensible don't work

Even small details can have major impacts on outputs and outcomes

- "Identical" policies can be very different in practice

Another example of how "the same" policy can have different results:

- Contract teachers in Kenya (Bold *et al* 2016)

Details matter

Example: contract teachers in Kenya

- Teachers hired on short-term contracts with performance incentive (outside the normal teacher system)
- An NGO-led RCT showed it was effective in Western Kenya

Nationwide scale-up with randomized implementation; government in some districts, an NGO in others

- On paper, policy details (theory of change) was the same
- Large impacts in NGO schools; zero impact in government schools

Bold *et al* 2016, "Experimental Evidence on Scaling Up Education Reforms in Kenya"

Fidelity

Because small policy details matter so much for effectiveness, fidelity gets emphasized when transporting policies

- Fidelity = stick as close to the original policy design as possible

Piecing together a successful set of components is hard

- So if a policy has been shown to work, you should replicate even tiny details as closely as possible

Adaptation

But contexts are different

- External validity depends on interaction of context and mechanism
- If contexts differ, shouldn't you try to adapt the policy?
- TINP and BINP

Adaptations can be superficial or more substantive

- Superficial: translating language, cultural references, etc.
- Substantive: eligibility, dosage, timing, etc.

The fidelity vs adaptation dilemma

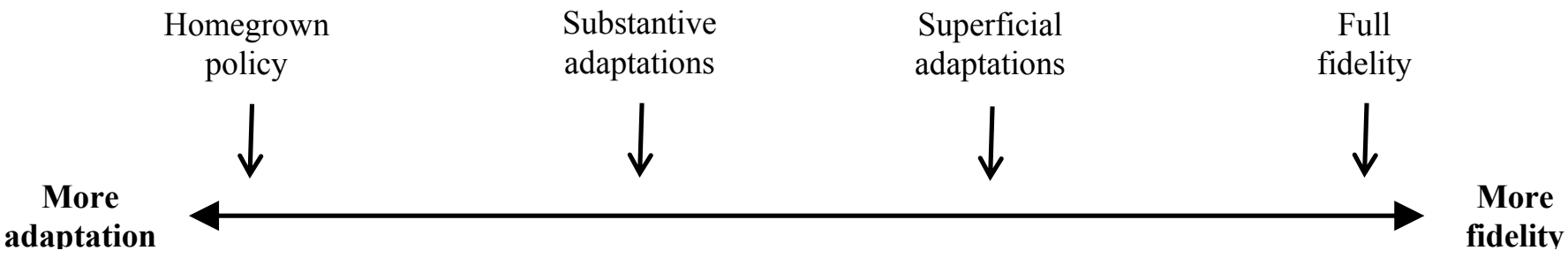
Fidelity

- Takes advantage of existing knowledge and past experimentation
- Keeps details as constant as possible
- But risks poor fit with local context

Adaptation

- Takes advantage of information on local context
- But risks ignoring valuable experience from other contexts

The fidelity - adaptation spectrum



External validity and adaptation

Two key questions for policy adaptation:

1. *Which aspects* of policy to adapt, which to leave alone?
2. *How much* to adapt a policy?

Difficult questions - judgment, not science

- Mechanism mapping can help structure thinking

Mechanism mapping in five steps

Diagnosis

1. Map out a policy's *theory of change*
2. Identify the *contextual assumptions* necessary for each step of the theory of change
3. Compare this to the *actual characteristics* of the context

Adaptation

4. *Adapt* the policy to eliminate mismatches between assumptions and characteristics

Iteration

5. Repeat steps 1-4 for the adapted policy

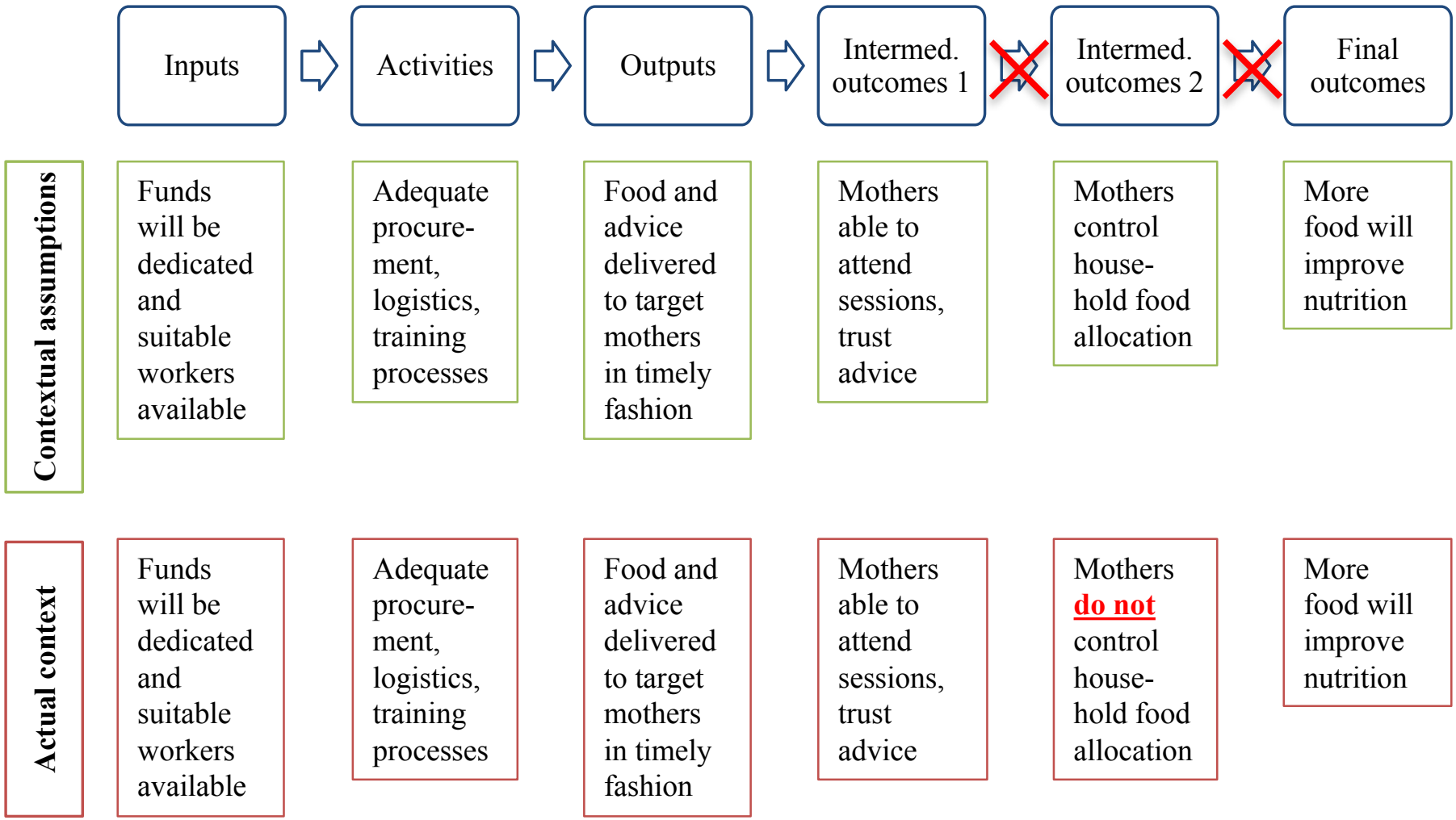
Which aspects to adapt?

Mechanism mapping doesn't determine *how* to adapt a policy. . .

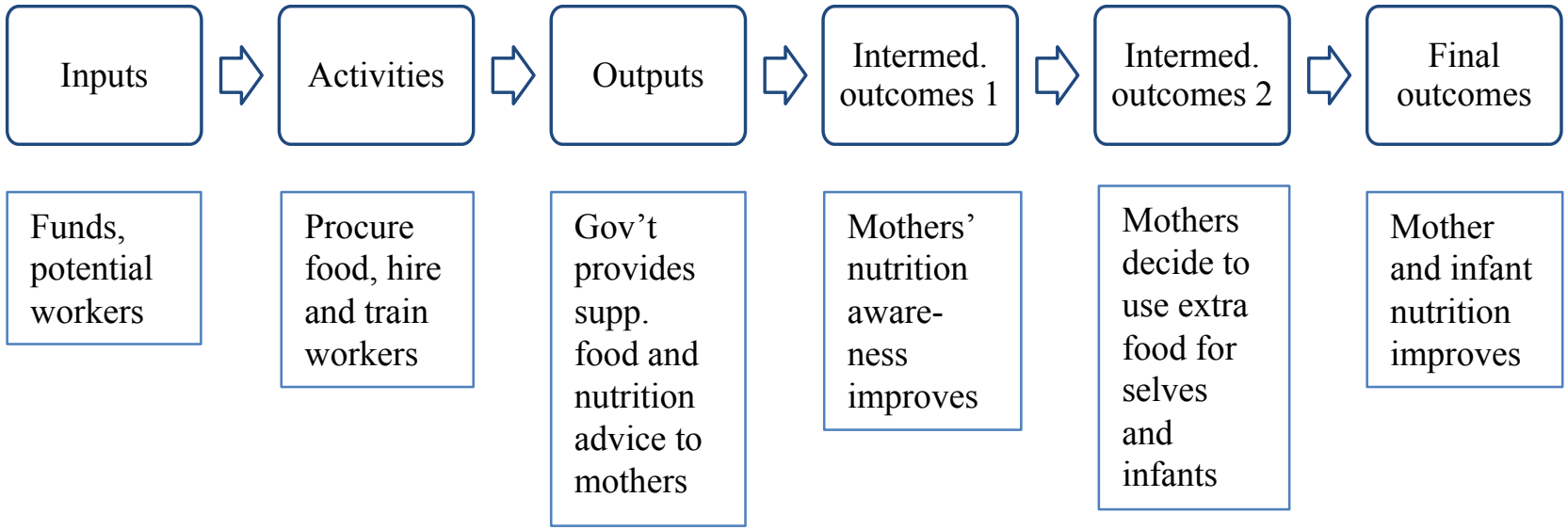
. . . but the diagnosis aspect of mechanism mapping focuses attention on *which aspects* need to be adapted

Let's use BINP as an example

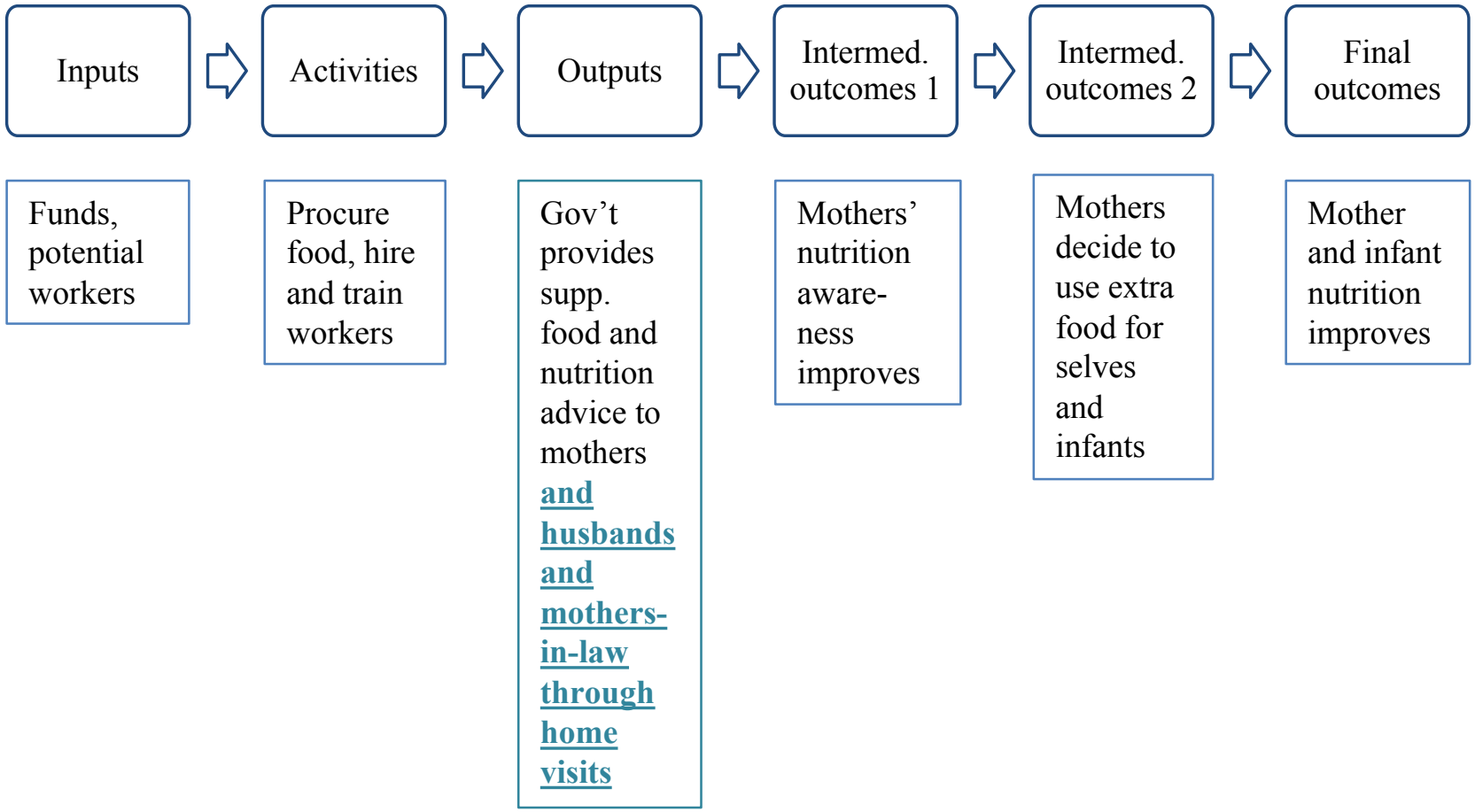
BINP Mechanism Map



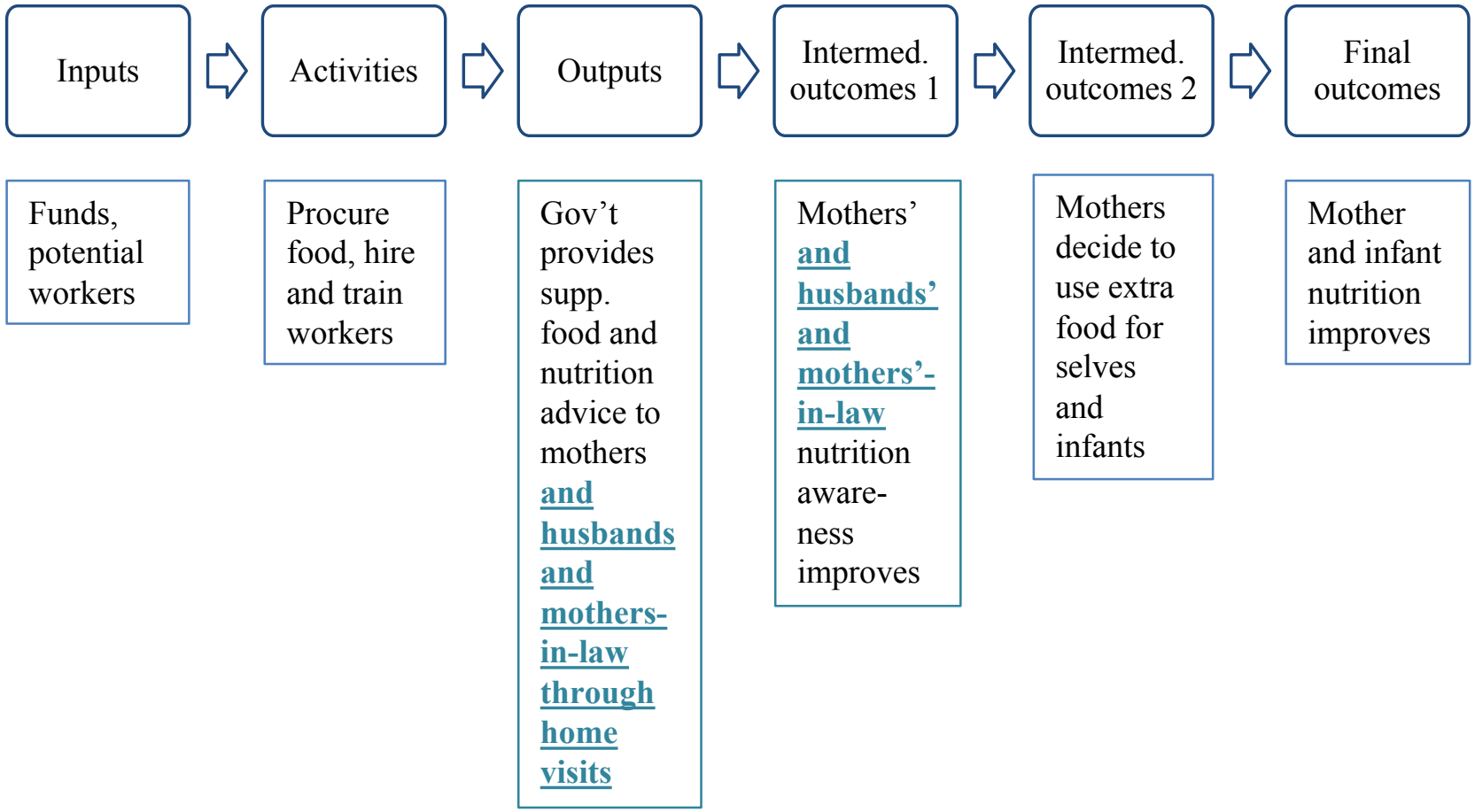
BINP Theory of Change



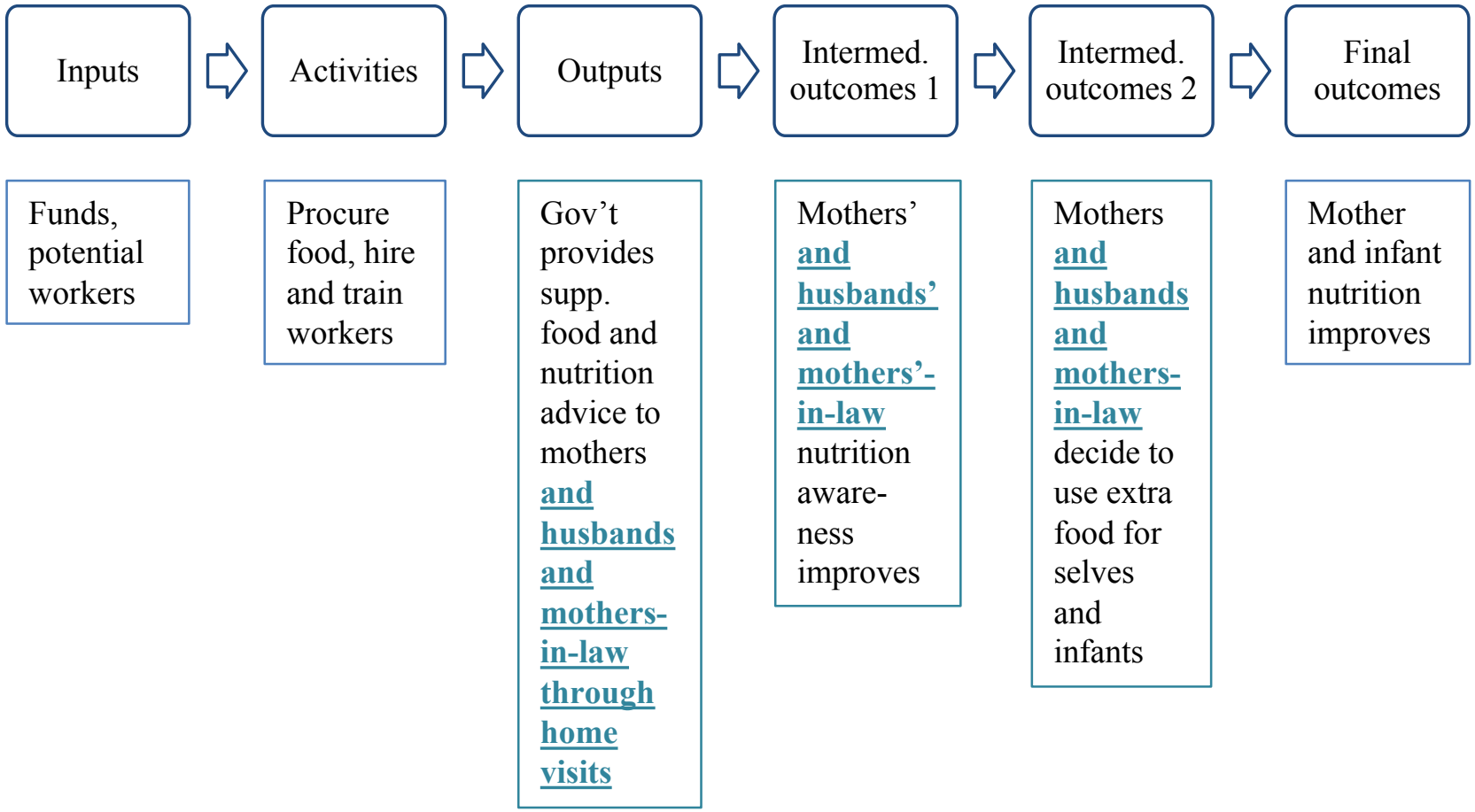
Adapted BINP Theory of Change



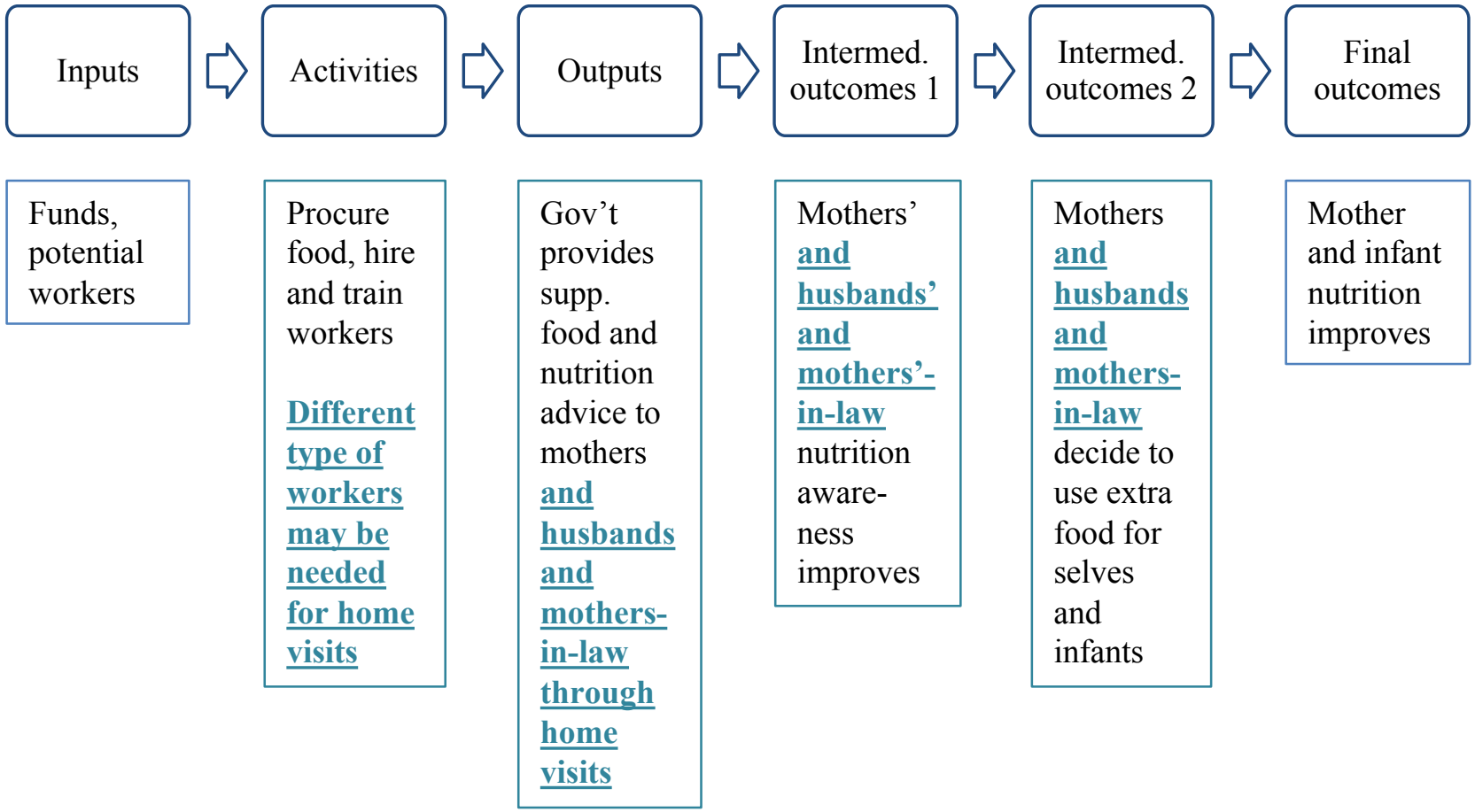
Adapted BINP Theory of Change



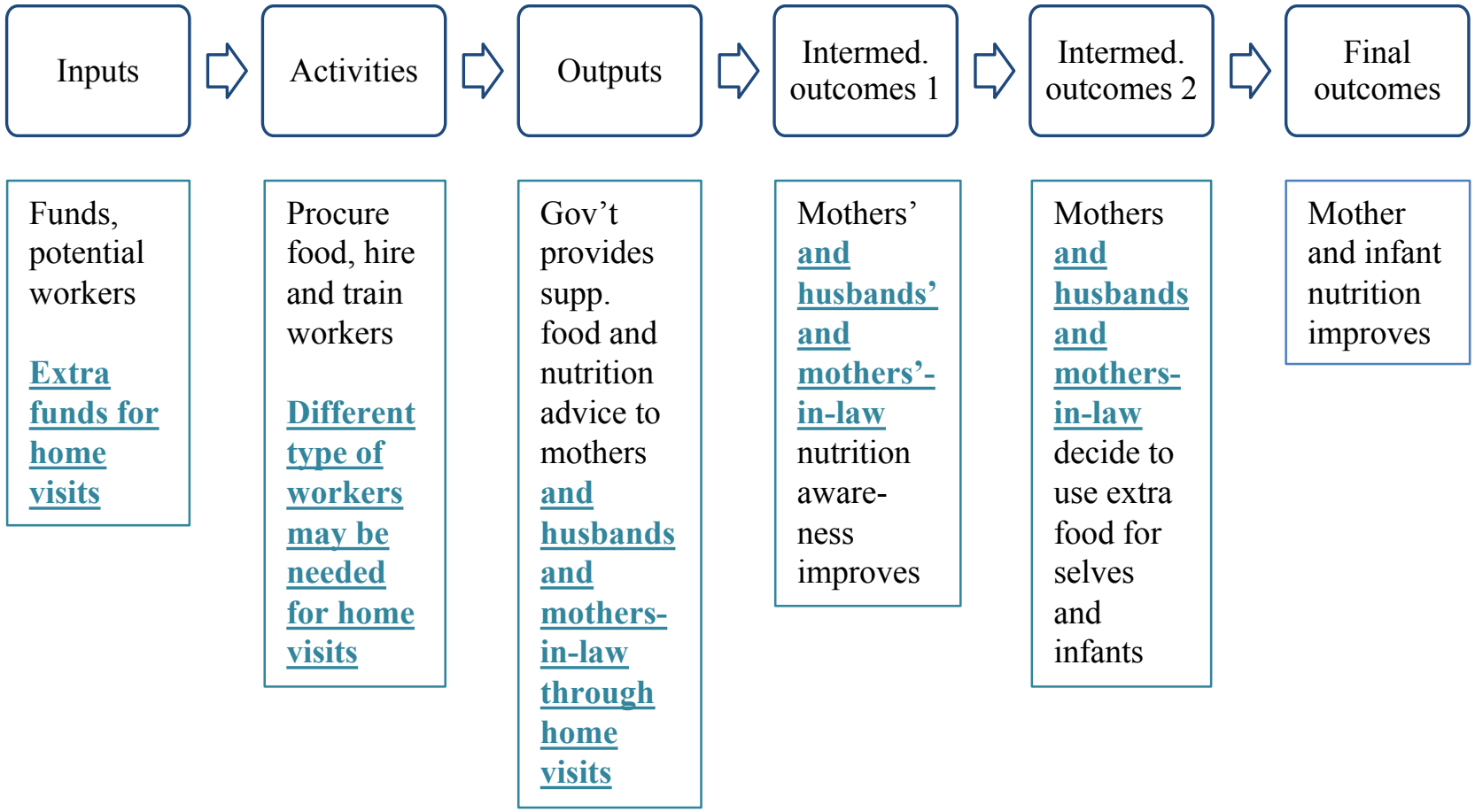
Adapted BINP Theory of Change



Adapted BINP Theory of Change



Adapted BINP Theory of Change

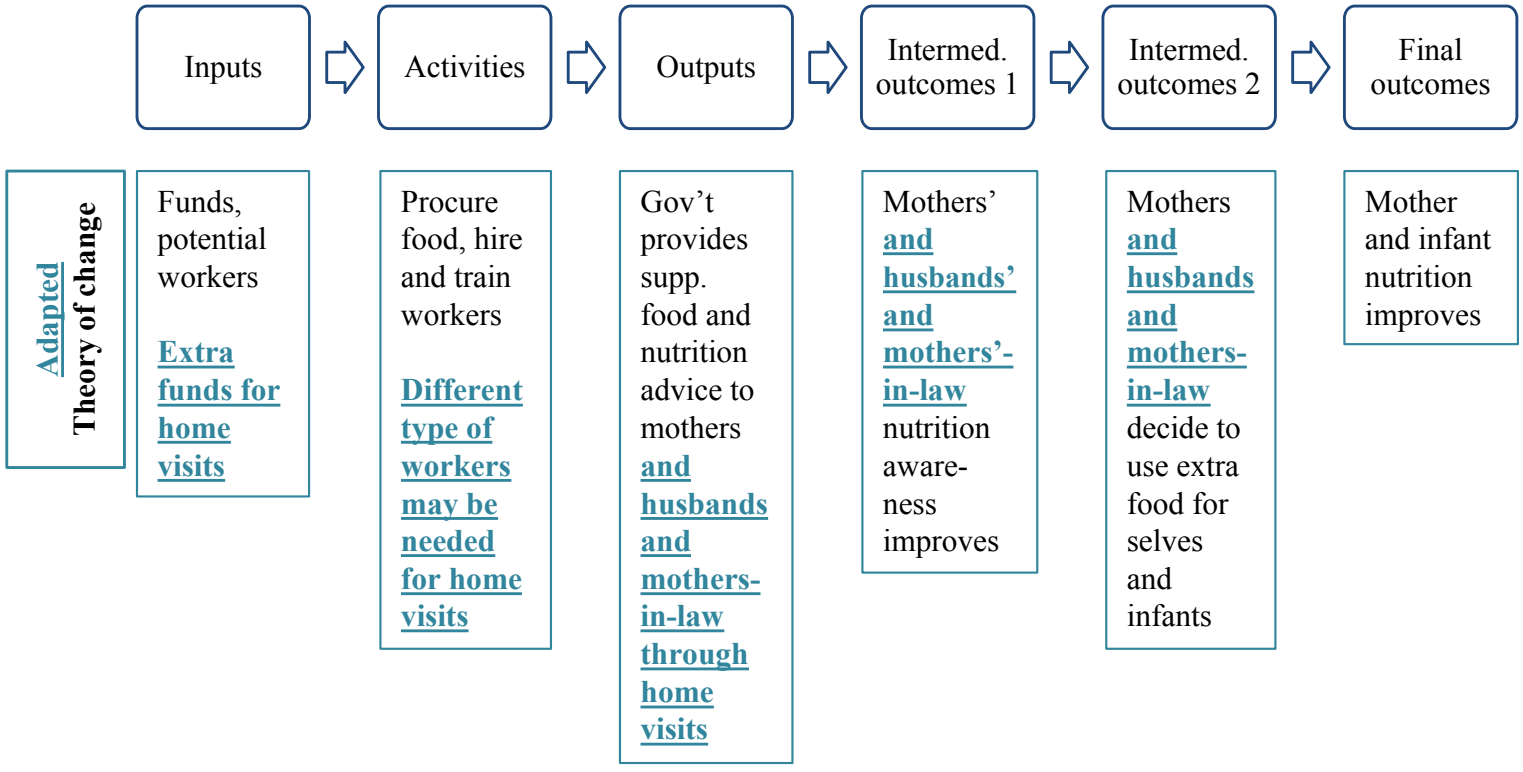


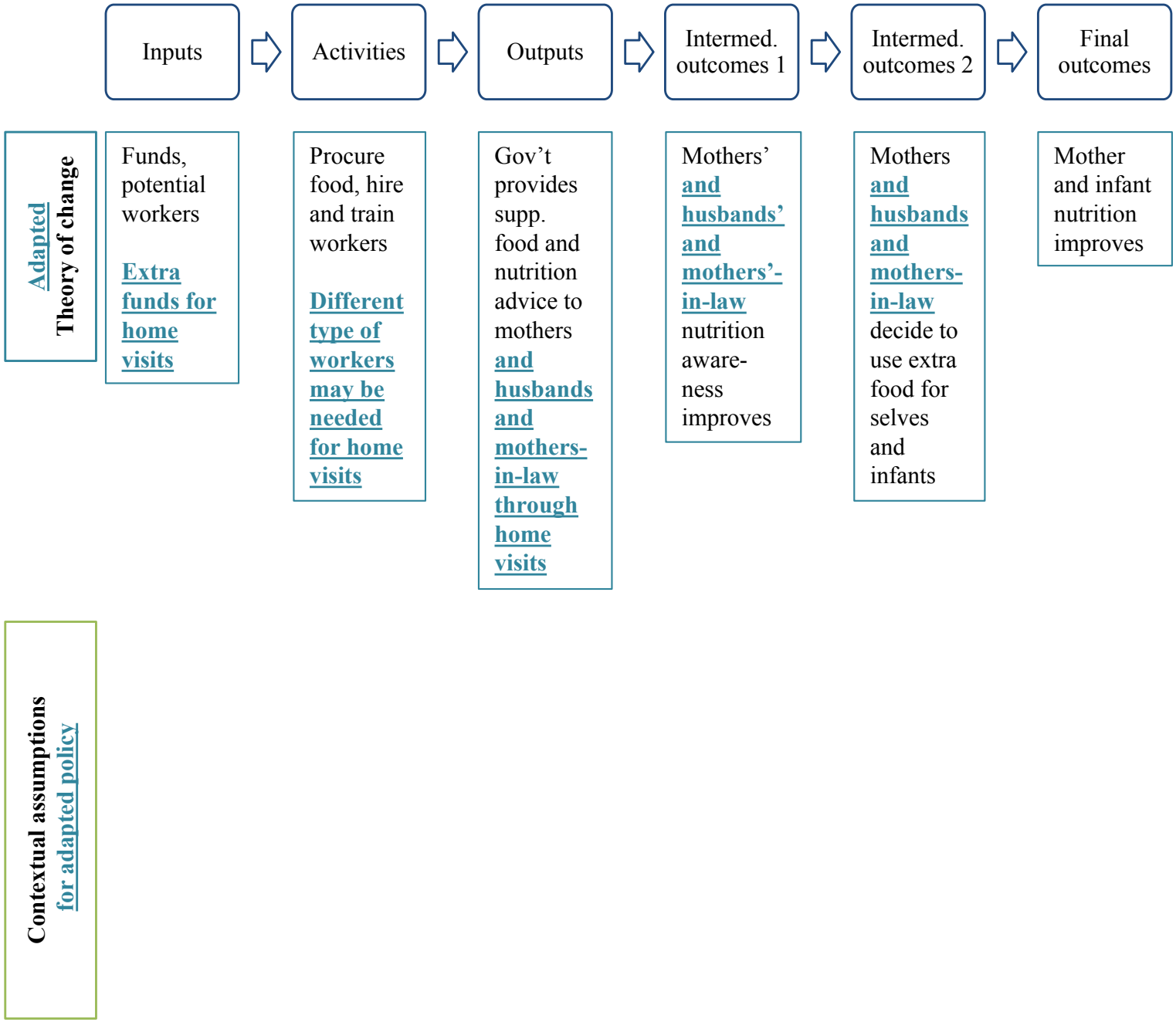
Step 5: Iterate

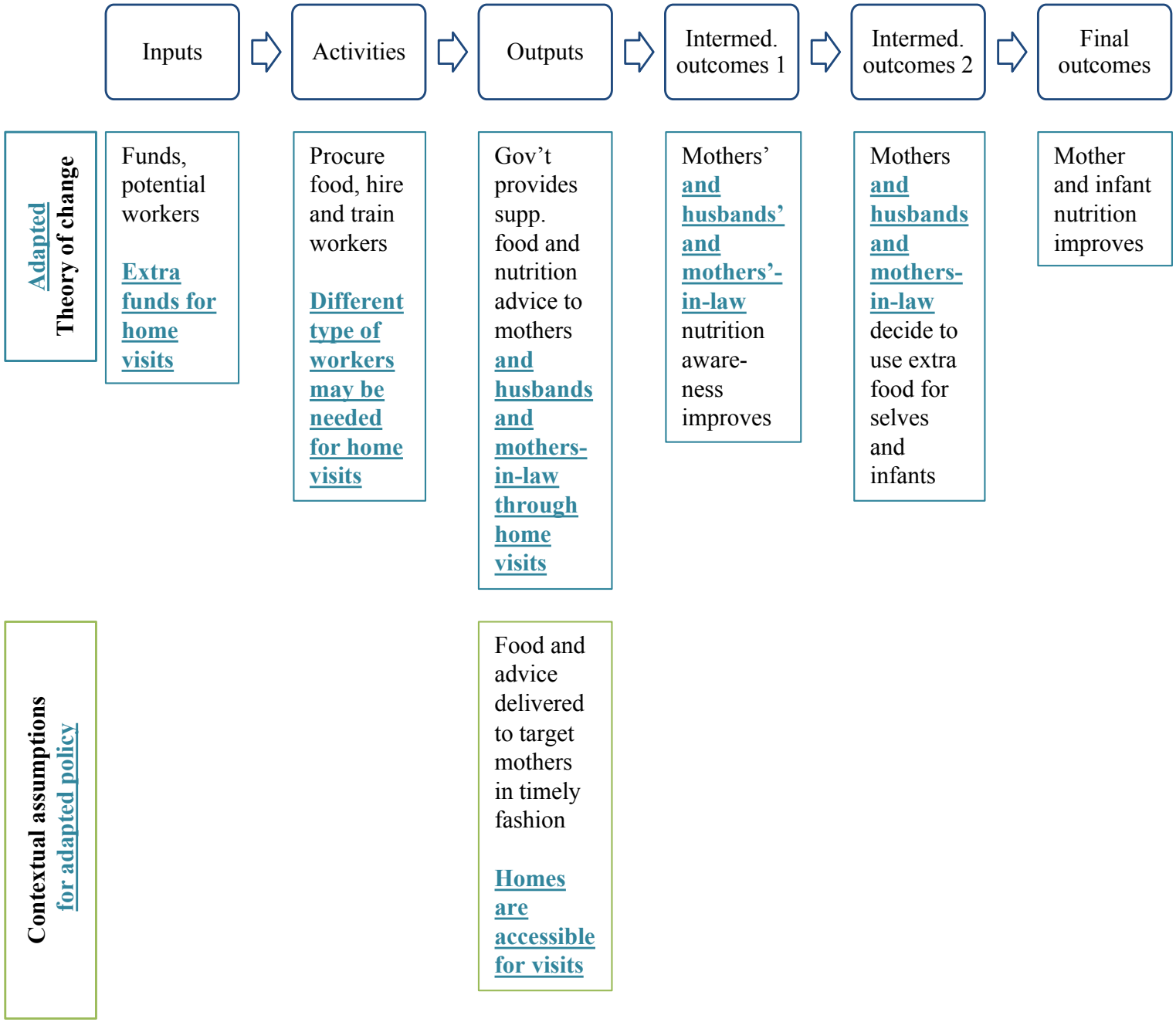
These adaptations to the theory of change also have contextual assumptions

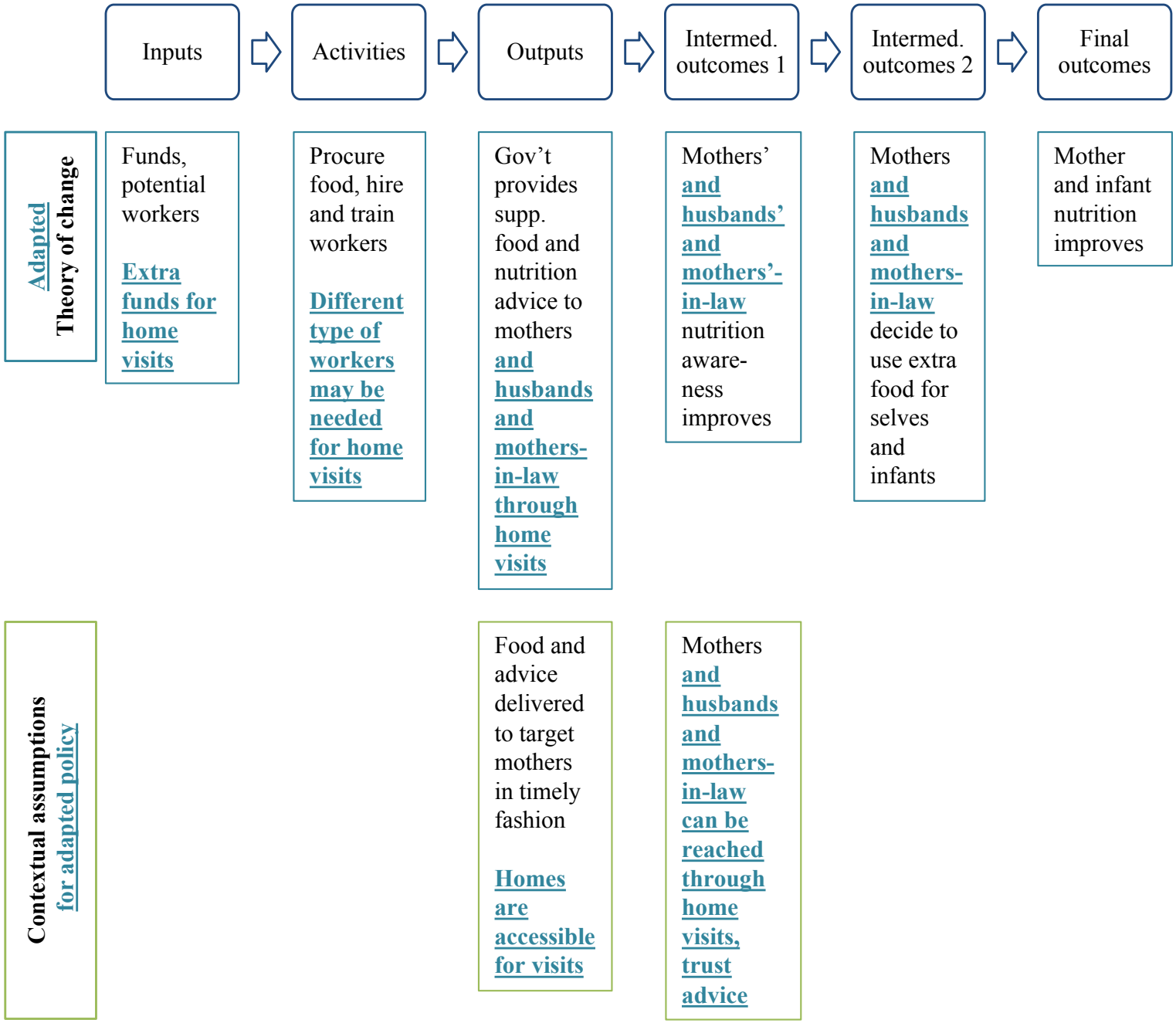
Iterate steps 1-4 on the adapted policy until contextual assumptions match actual contextual characteristics

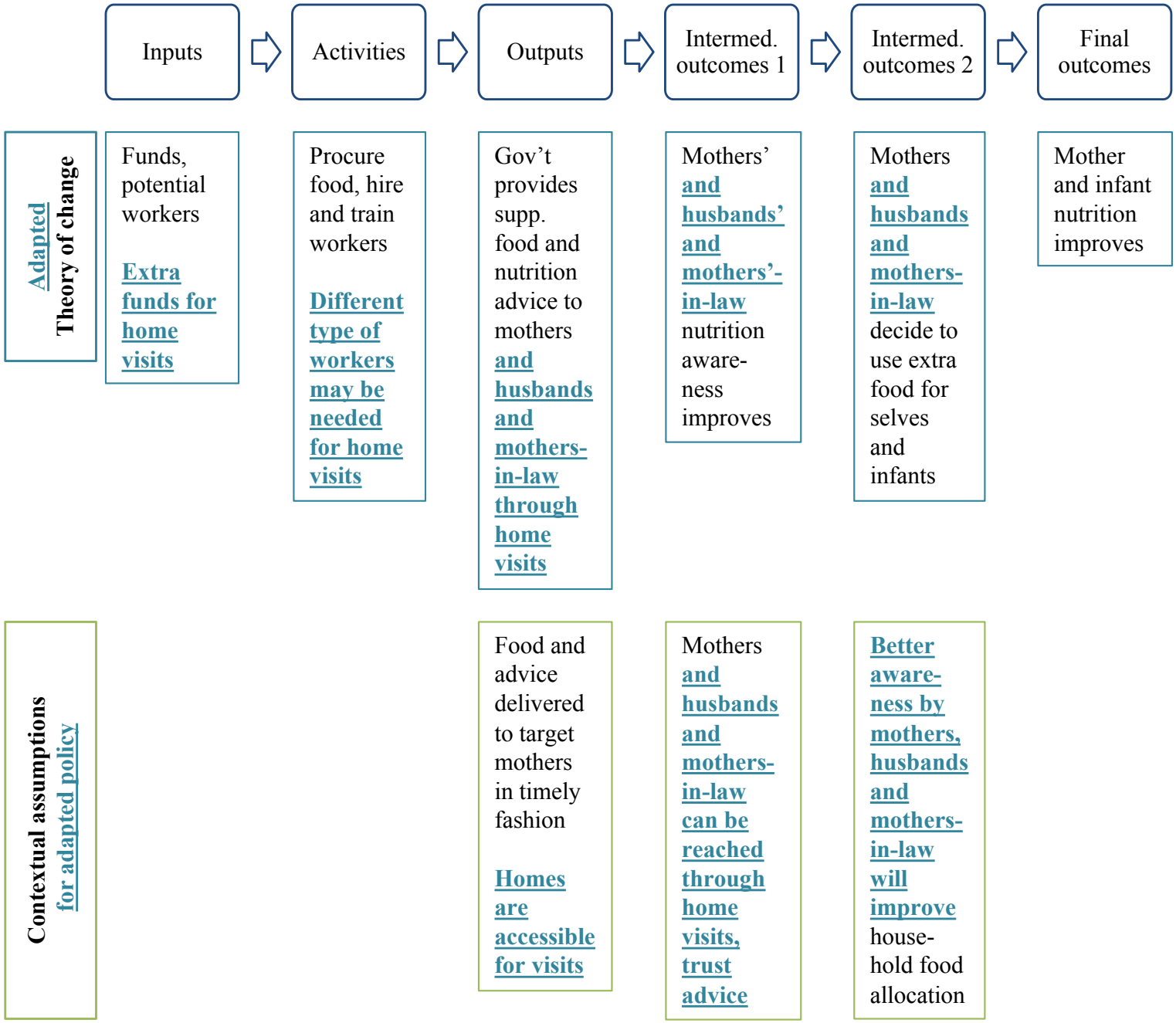
- In some cases even adapted policy may not fit context; give up and design a new policy from scratch

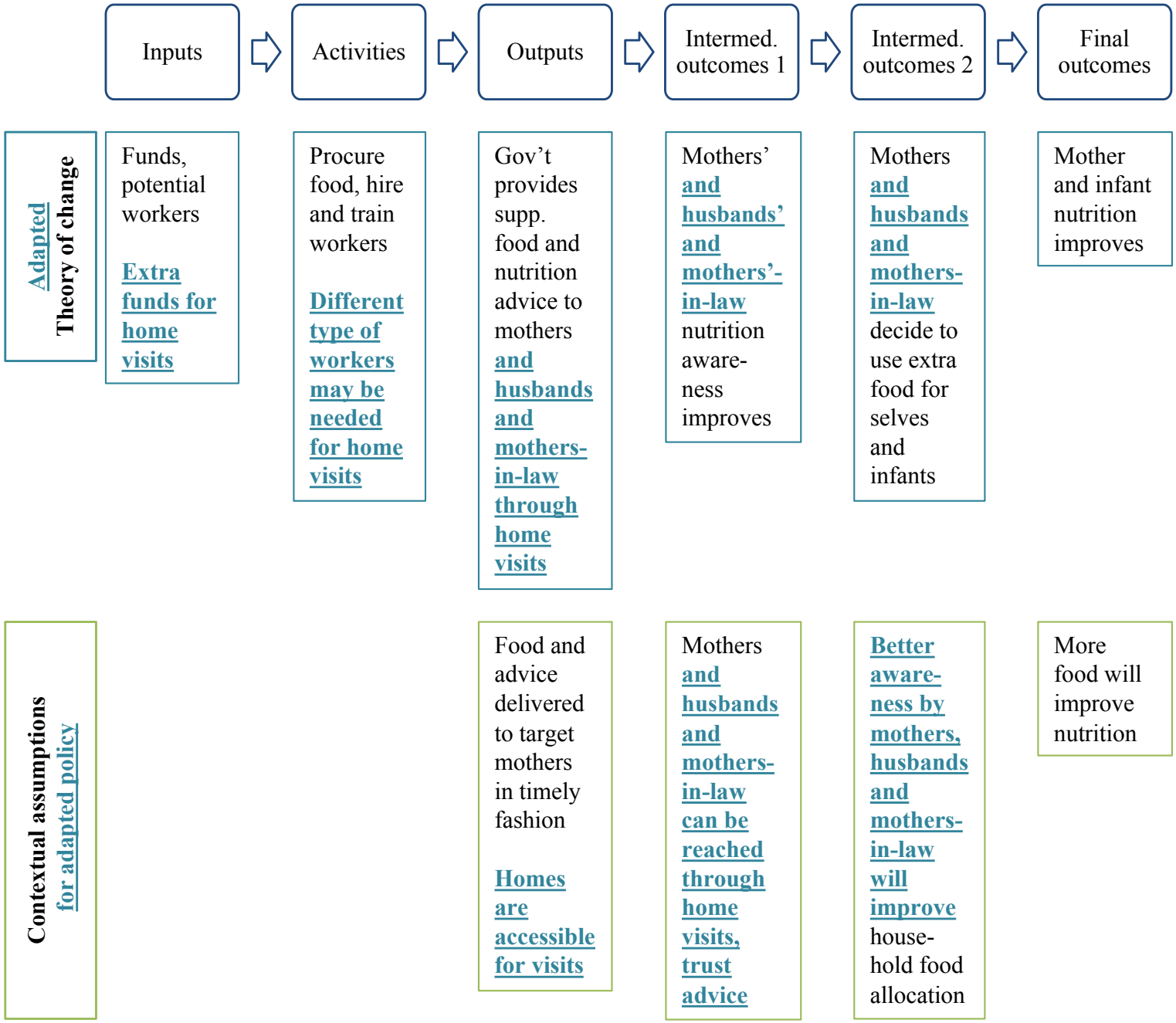

Adapted Theory of change
Inputs
Activities
Outputs
Intermed. outcomes 1
Intermed. outcomes 2
Final outcomes
Funds, potential workers
Extra funds for home visits
Procure food, hire and train workers
Different type of workers may be needed for home visits
Gov't provides supp. food and nutrition advice to mothers and husbands and mothers-in-law through home visits
Mothers' and husbands' and mothers'-in-law nutrition aware-ness improves
Mothers and husbands and mothers-in-law decide to use extra food for selves and infants
Mother and infant nutrition improves

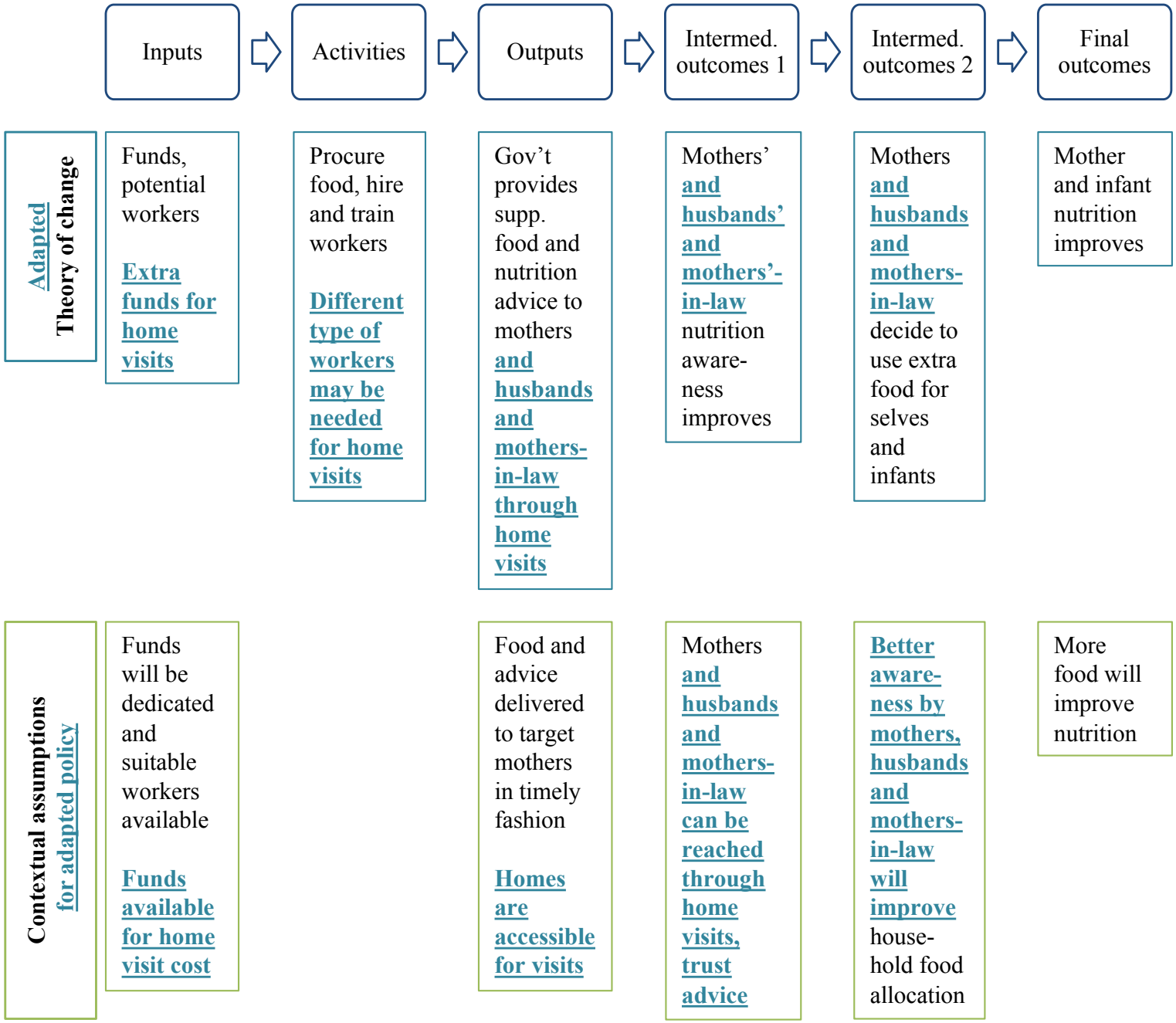

Inputs
Activities
Outputs
Intermed. outcomes 1
Intermed. outcomes 2
Final outcomes
Adapted Theory of change
Funds, potential workers
Extra funds for home visits
Procure food, hire and train workers
Different type of workers may be needed for home visits
Gov't provides supp. food and nutrition advice to mothers and husbands and mothers-in-law through home visits
Mothers' and husbands' and mothers'-in-law nutrition aware-ness improves
Mothers and husbands and mothers-in-law decide to use extra food for selves and infants
Mother and infant nutrition improves
Contextual assumptions for adapted policy

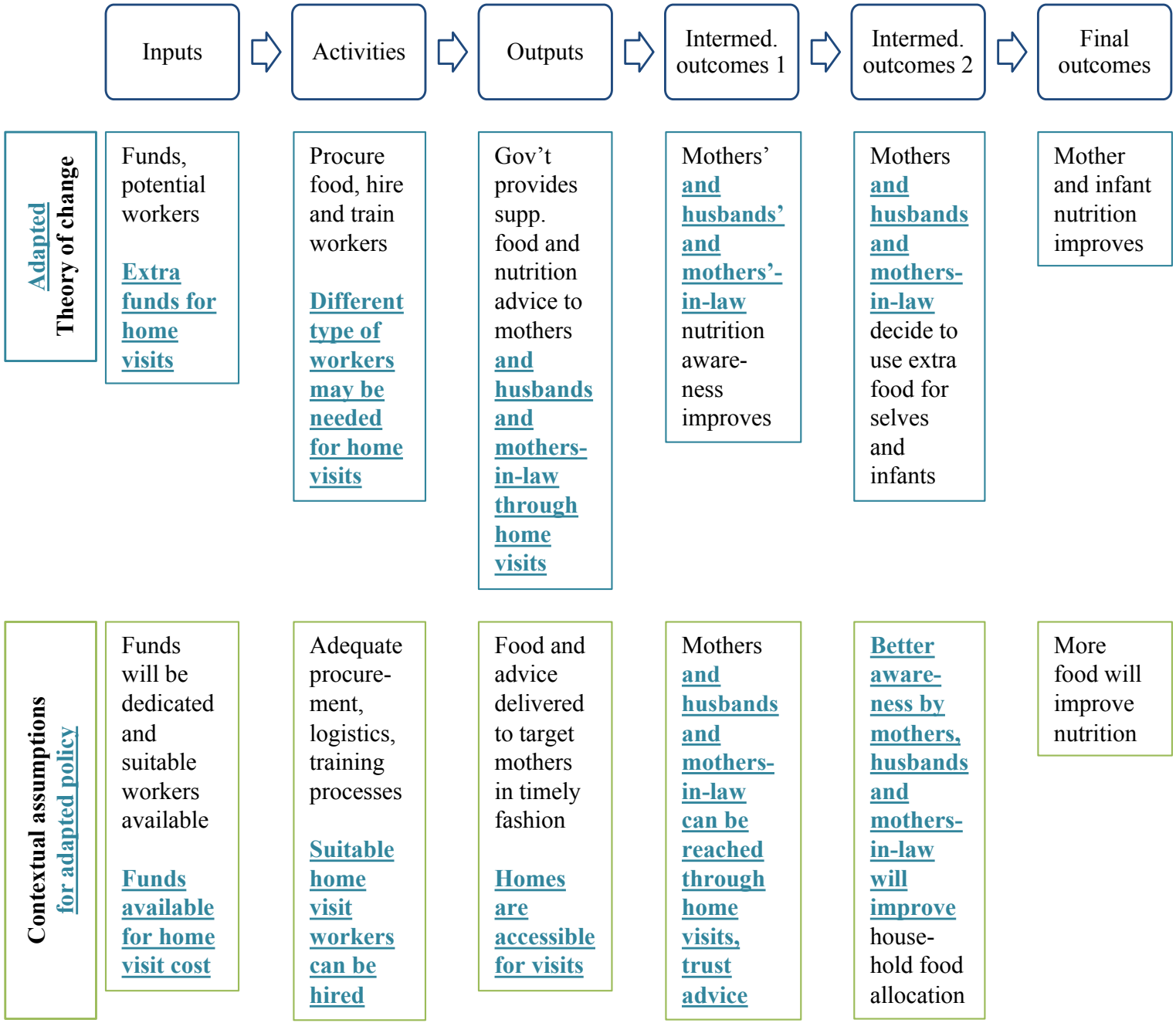

Inputs
Activities
Outputs
Intermed. outcomes 1
Intermed. outcomes 2
Final outcomes
Adapted Theory of change
Funds, potential workers
Extra funds for home visits
Procure food, hire and train workers
Different type of workers may be needed for home visits
Gov't provides supp. food and nutrition advice to mothers and husbands and mothers-in-law through home visits
Mothers' and husbands' and mothers'-in-law nutrition aware-ness improves
Mothers and husbands and mothers-in-law decide to use extra food for selves and infants
Mother and infant nutrition improves
Contextual assumptions for adapted policy
Food and advice delivered to target mothers in timely fashion
Homes are accessible for visits


Inputs
Activities
Outputs
Intermed. outcomes 1
Intermed. outcomes 2
Final outcomes
Adapted Theory of change
Funds, potential workers
Extra funds for home visits
Procure food, hire and train workers
Different type of workers may be needed for home visits
Gov't provides supp. food and nutrition advice to mothers and husbands and mothers-in-law through home visits
Mothers' and husbands' and mothers'-in-law nutrition aware-ness improves
Mothers and husbands and mothers-in-law decide to use extra food for selves and infants
Mother and infant nutrition improves
Contextual assumptions for adapted policy
Food and advice delivered to target mothers in timely fashion
Homes are accessible for visits
Mothers and husbands and mothers-in-law can be reached through home visits, trust advice


Inputs
Activities
Outputs
Intermed. outcomes 1
Intermed. outcomes 2
Final outcomes
Adapted Theory of change
Funds, potential workers
Extra funds for home visits
Procure food, hire and train workers
Different type of workers may be needed for home visits
Gov't provides supp. food and nutrition advice to mothers and husbands and mothers-in-law through home visits
Mothers' and husbands' and mothers'-in-law nutrition aware-ness improves
Mothers and husbands and mothers-in-law decide to use extra food for selves and infants
Mother and infant nutrition improves
Contextual assumptions for adapted policy
Food and advice delivered to target mothers in timely fashion
Homes are accessible for visits
Mothers and husbands and mothers-in-law can be reached through home visits, trust advice
Better aware-ness by mothers, husbands and mothers-in-law will improve house-hold food allocation


Inputs
Activities
Outputs
Intermed. outcomes 1
Intermed. outcomes 2
Final outcomes
Adapted Theory of change
Funds, potential workers
Extra funds for home visits
Procure food, hire and train workers
Different type of workers may be needed for home visits
Gov't provides supp. food and nutrition advice to mothers and husbands and mothers-in-law through home visits
Mothers' and husbands' and mothers'-in-law nutrition aware-ness improves
Mothers and husbands and mothers-in-law decide to use extra food for selves and infants
Mother and infant nutrition improves
Contextual assumptions for adapted policy
Food and advice delivered to target mothers in timely fashion
Homes are accessible for visits
Mothers and husbands and mothers-in-law can be reached through home visits, trust advice
Better aware-ness by mothers, husbands and mothers-in-law will improve house-hold food allocation
More food will improve nutrition


Inputs
Activities
Outputs
Intermed. outcomes 1
Intermed. outcomes 2
Final outcomes
Adapted Theory of change
Funds, potential workers
Extra funds for home visits
Procure food, hire and train workers
Different type of workers may be needed for home visits
Gov't provides supp. food and nutrition advice to mothers and husbands and mothers-in-law through home visits
Mothers' and husbands' and mothers'-in-law nutrition aware-ness improves
Mothers and husbands and mothers-in-law decide to use extra food for selves and infants
Mother and infant nutrition improves
Contextual assumptions for adapted policy
Funds will be dedicated and suitable workers available
Funds available for home visit cost
Food and advice delivered to target mothers in timely fashion
Homes are accessible for visits
Mothers and husbands and mothers-in-law can be reached through home visits, trust advice
Better aware-ness by mothers, husbands and mothers-in-law will improve house-hold food allocation
More food will improve nutrition


Inputs
Activities
Outputs
Intermed. outcomes 1
Intermed. outcomes 2
Final outcomes
Adapted Theory of change
Funds, potential workers
Extra funds for home visits
Procure food, hire and train workers
Different type of workers may be needed for home visits
Gov't provides supp. food and nutrition advice to mothers and husbands and mothers-in-law through home visits
Mothers' and husbands' and mothers'-in-law nutrition aware-ness improves
Mothers and husbands and mothers-in-law decide to use extra food for selves and infants
Mother and infant nutrition improves
Contextual assumptions for adapted policy
Funds will be dedicated and suitable workers available
Funds available for home visit cost
Adequate procure-ment, logistics, training processes
Suitable home visit workers can be hired
Food and advice delivered to target mothers in timely fashion
Homes are accessible for visits
Mothers and husbands and mothers-in-law can be reached through home visits, trust advice
Better aware-ness by mothers, husbands and mothers-in-law will improve house-hold food allocation
More food will improve nutrition

Limitations of mechanism mapping

Only gives qualitative predictions, not quantitative estimates

- Overall effect may be ambiguous

Requires ability to map ToC, identify salient assumptions

- Not-so-salient things can still be important

Without evidence, end up relying on intuition a lot

Ideal approach combines meta-analysis, sub-group analysis, and mechanism mapping (with empirical evidence)

How much to adapt?

We can propose many adaptations. But should we actually make them?

- Designing effective policies is hard. Why should we assume our ideas will be better than what has worked elsewhere (even with tools like mechanism mapping)?
- Our adaptations might even make things worse

Remember the policy quiz: our intuitions are not always accurate

The Fundamental Informational Trade-off

Evaluation evidence on a policy's effectiveness in other contexts is usually more rigorous than information about the local context

- But relying on this evidence requires strict fidelity to original policy design
- Implies that policymakers should make minimal adaptations

But using mechanism mapping to identify potential adaptations makes efficient use of local information

- It seems crazy to ignore what we know about our local context!
- But using this local information to make adaptations decreases the relevance of evaluation evidence from elsewhere

How should we make this trade-off?

Lean towards more fidelity when:

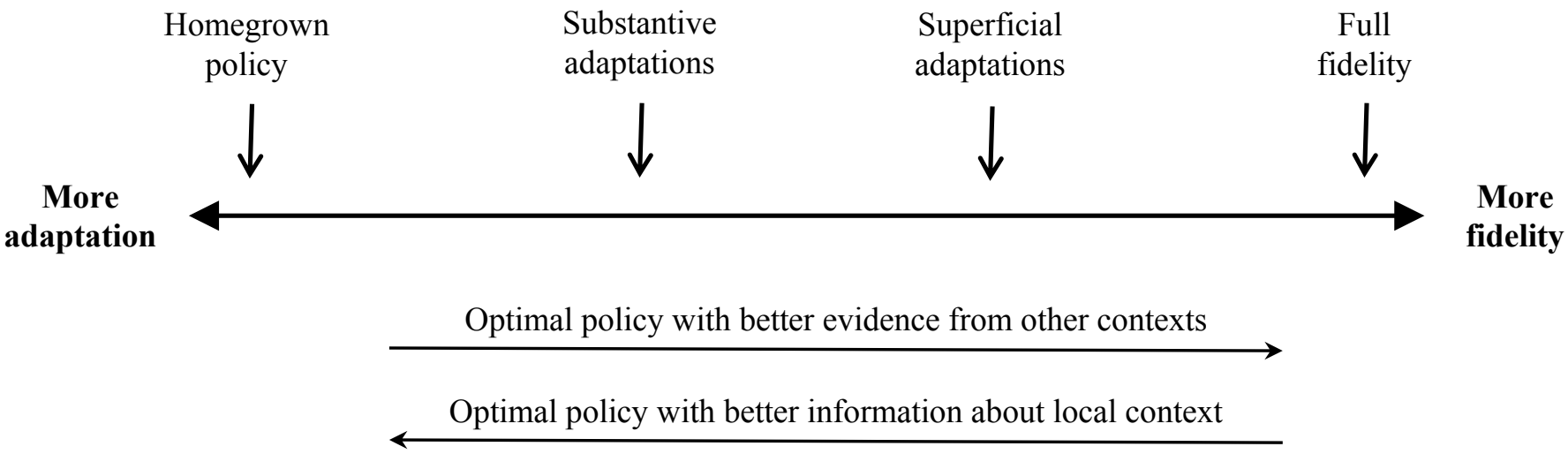
- Evidence on policy's effectiveness is strong, consistent, and from similar contexts
- We don't have good information about local context

Lean towards more adaptation when:

- Evidence on policy's effectiveness is weak, variable, or not relevant
- We have very good information about local context

Mechanism mapping can help weigh balance of evidence, areas of uncertainty

The Fundamental Informational Trade-off



Recap

Concerns about external validity and adaptation are central to policymaking

- Evidence provides a starting point, but judgment will always be crucial

Medical practice always based on combination of:

1. Rigorous evidence
2. Doctor's expertise
3. Doctor's judgment on each specific case

1) Which aspects to adapt and 2) how much to adapt?

Mechanism mapping in groups

Each group will have 8 minutes maximum to present:

- What the original policy is
- A mechanism map diagnosing its external validity
- Proposed adaptations, including:
  - A mechanism map of the adapted policy
  - How you would use evidence to inform these adaptations
- How you would evaluate whether the adapted policy is effective

Followed by 7 minutes questions and discussion

You now have time to prepare your presentation

Final course assignment

Final assignment: do a mechanism map for a policy of your choice

- A policy you know of from another country or another part of Brazil
- A policy your department is implementing or considering

Structured along the five steps in the policy memo

See written instructions for full details