Impact Evaluation: Randomisation

Development Studies Association of Ireland
May 09, 2016

Presentation Plan

1. Introduction
2. Overview of IE
3. NGO Considering an IE: A Guide
4. Example: Balsakhi Programme
5. Determining Programme Impact
6. Concerns about Randomised Evaluations
7. Final Remarks

1. Introduction

- We associate randomised trials with new drugs: Would you take a drug without evidence that it has positive effects?
- Until recently we didn't associate randomised trials with school textbooks, mosquito bed nets and governance interventions.
- Should we invest money earmarked for the poorest without robust evidence of positive impacts?

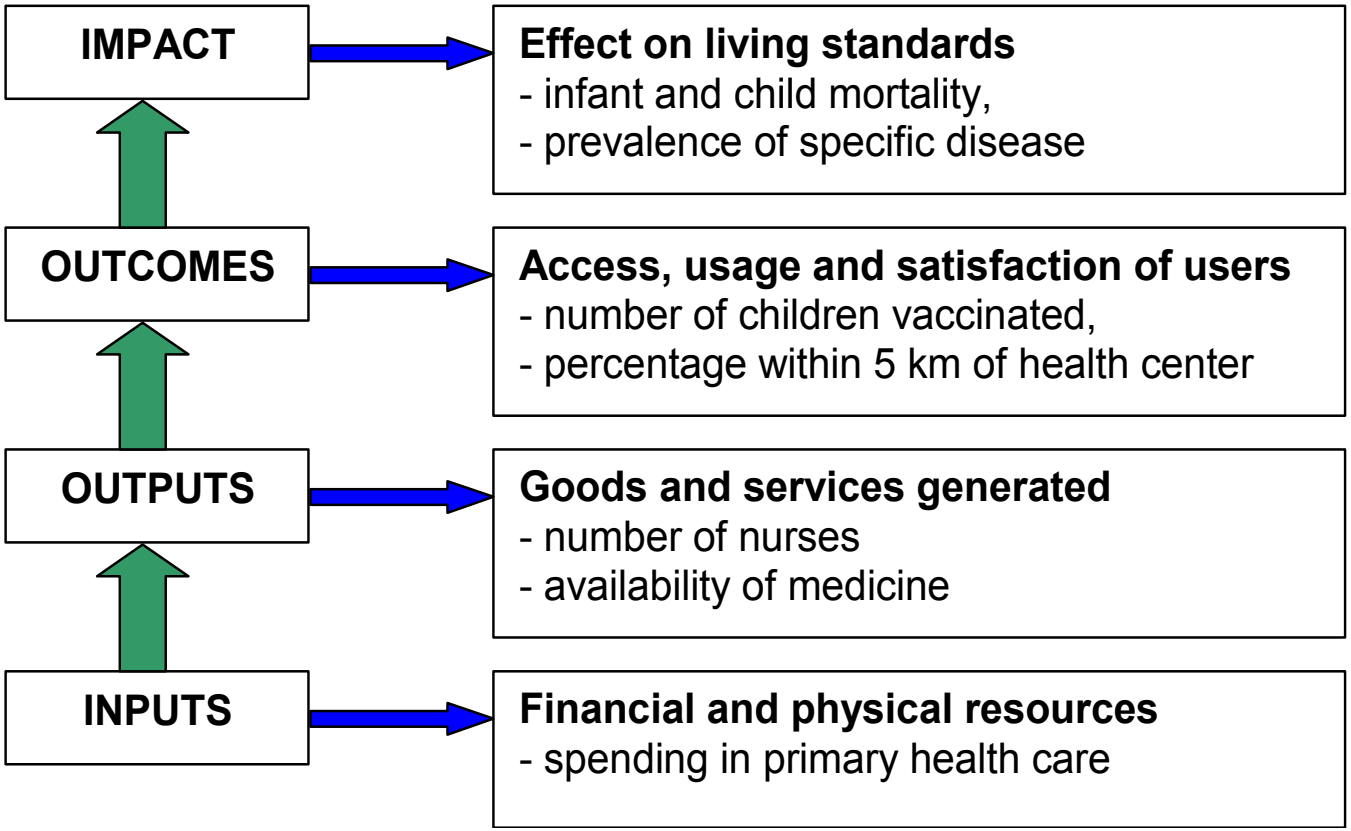
1. Introduction

- We associate randomised trials with new drugs: Would you take a drug without evidence that it has positive effects?
- Until recently we didn't associate randomised trials with school textbooks, mosquito bed nets and governance interventions
- Should we invest money earmarked for the poorest without robust evidence of positive impacts?

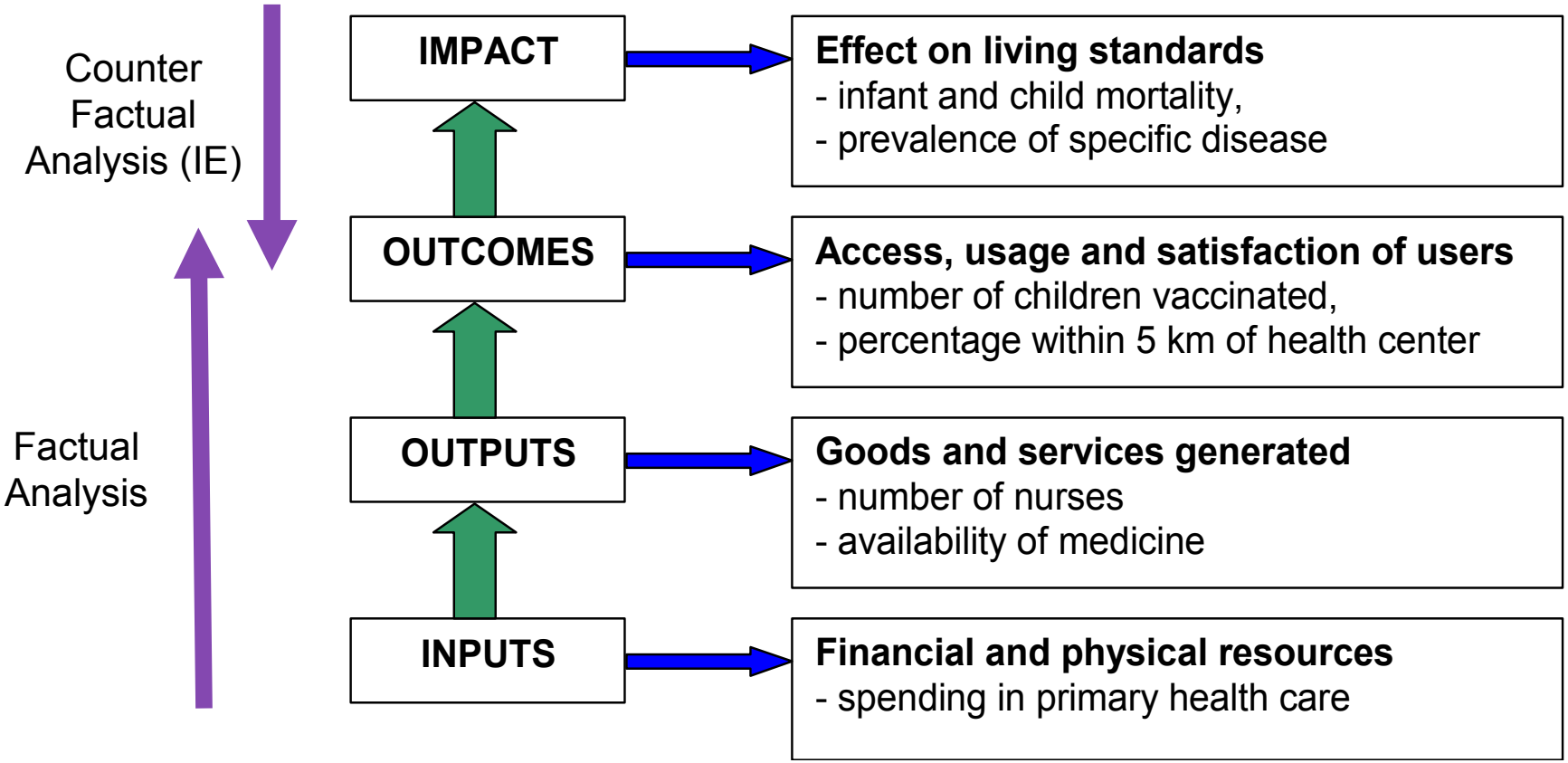
- We should agree that increasingly robust evidence should underpin/inform the allocation of aid monies.
- What form the evidence should take is where the debate occurs.

2. Overview of IE

IE in Project Design



Overview of IE
IE in Project Design



Overview of IE

How to measure impact

What would have happened in the absence of the programme?

Take the difference between

- what happened (with the programme) ...and
- what would have happened (without the programme)

=IMPACT of the programme

➢ Importance of constructing the best counterfactual will be illustrated with an example.

Overview of IE

How to measure impact

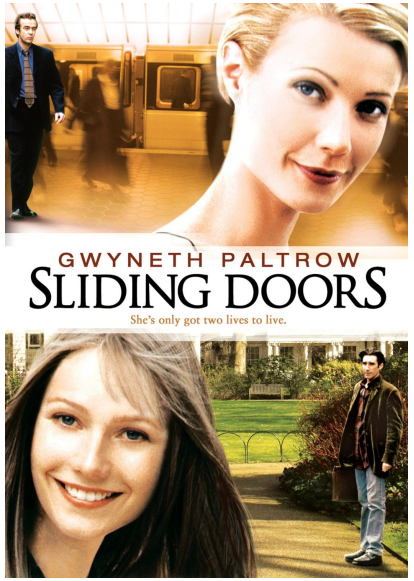
What would have happened in the absence of the programme?

Take the difference between

- what happened (with the programme) ...and
- what would have happened (without the programme)

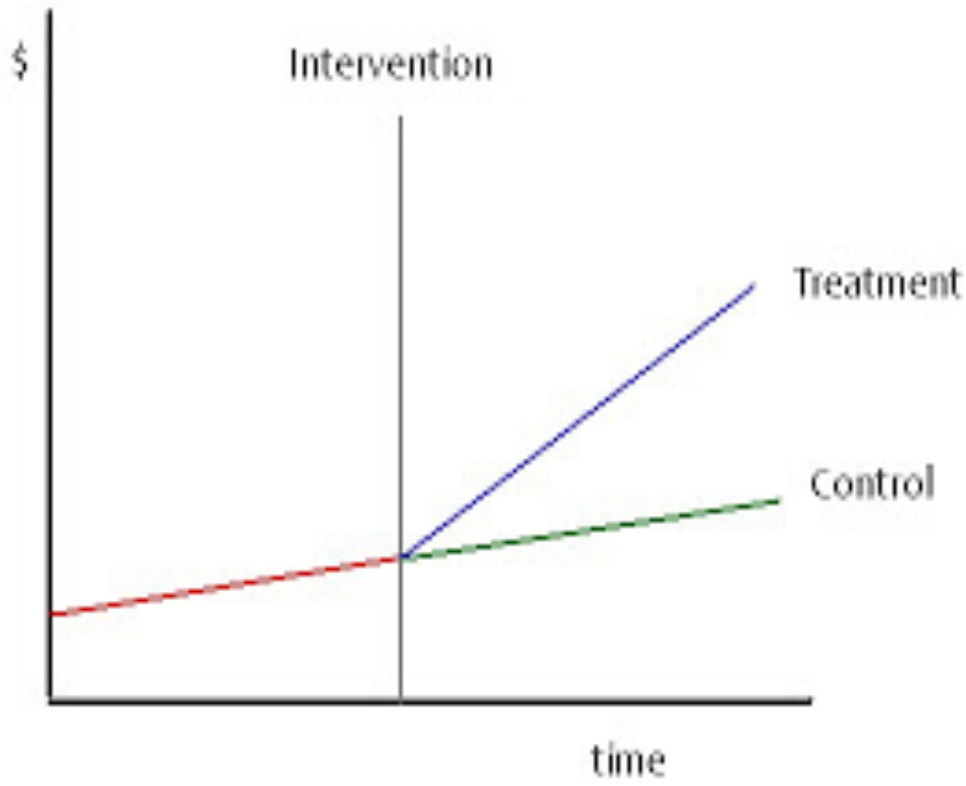
=IMPACT of the programme

➢ Importance of constructing the best counterfactual will be illustrated with an example.



Example of the counterfactual

The Counterfactual



Counterfactual is often constructed by selecting a group not affected by the programme.

Use random assignment of the programme to create a control group which mimics the counterfactual.

Non-randomised: Argue that a certain excluded group mimics the counterfactual.

Overview of IE

Options in Randomisation

1. **Simple Random Assignment.** When the number of intervention participants has been decided and additional eligible individuals are identified, simple random assignment through a coin flip or lottery can be used to select the treatment group and control groups.

2. **Phased-In Selection.** Even if an intervention plans to treat all eligible beneficiaries, there may be logistical reasons that prevent implementation from beginning everywhere at the same time. This type of schedule creates a natural opportunity for using an experimental design.

3. **Randomised Promotion (Encouragement Design).** In cases where randomised assignment is difficult, evaluators can randomise promotion of a particular intervention. For instance, a microfinance institution might be unwilling to turn potential clients away just because they are assigned to a control group, preferring to serve anyone who seeks to open a savings account.

Overview of IE

Additional Options

4. Randomisation near the threshold.
5. Multiple treatment groups: Perfectly possible but a control group still required.
6. Cross-cutting treatments: Useful for testing different components of treatment in different combinations.
7. Varying levels of treatment: Different dosage levels.

3. NGO Considering an IE: A Guide

When an NGO should consider an IE?

1. Is the intervention **innovative**? Is it testing a new, promising approach?
   Is the intervention **replicable**? Can it be scaled up or can it be applied to a different setting?
2. Is the intervention **strategically relevant**? Is it a flagship intervention that requires substantial resources; covers, or could be expanded to cover, a large number of people; or could generate substantial savings?
3. Is the intervention **untested**? That is, is very little known about the effectiveness of the intervention globally or in the specific context in which it is implemented?
4. Is the intervention **influential**? Will results be used to inform key policy decisions?

Practically when is a Randomised Evaluation Appropriate?

When to do a randomised evaluation

- When there is an important question you want/need to know the answer to.
- Programme is representative not gold plated.
- Time, expertise, and money to do it right.

When not to do a randomised evaluation

- When the programme is premature and still requires considerable "tinkering" to work well.
- When the project is on too small a scale to randomise into two "representative groups".
- If a positive impact has been proven using rigorous methodology and resources are sufficient to cover everyone.
- After the programme has already begun and you are not expanding elsewhere.

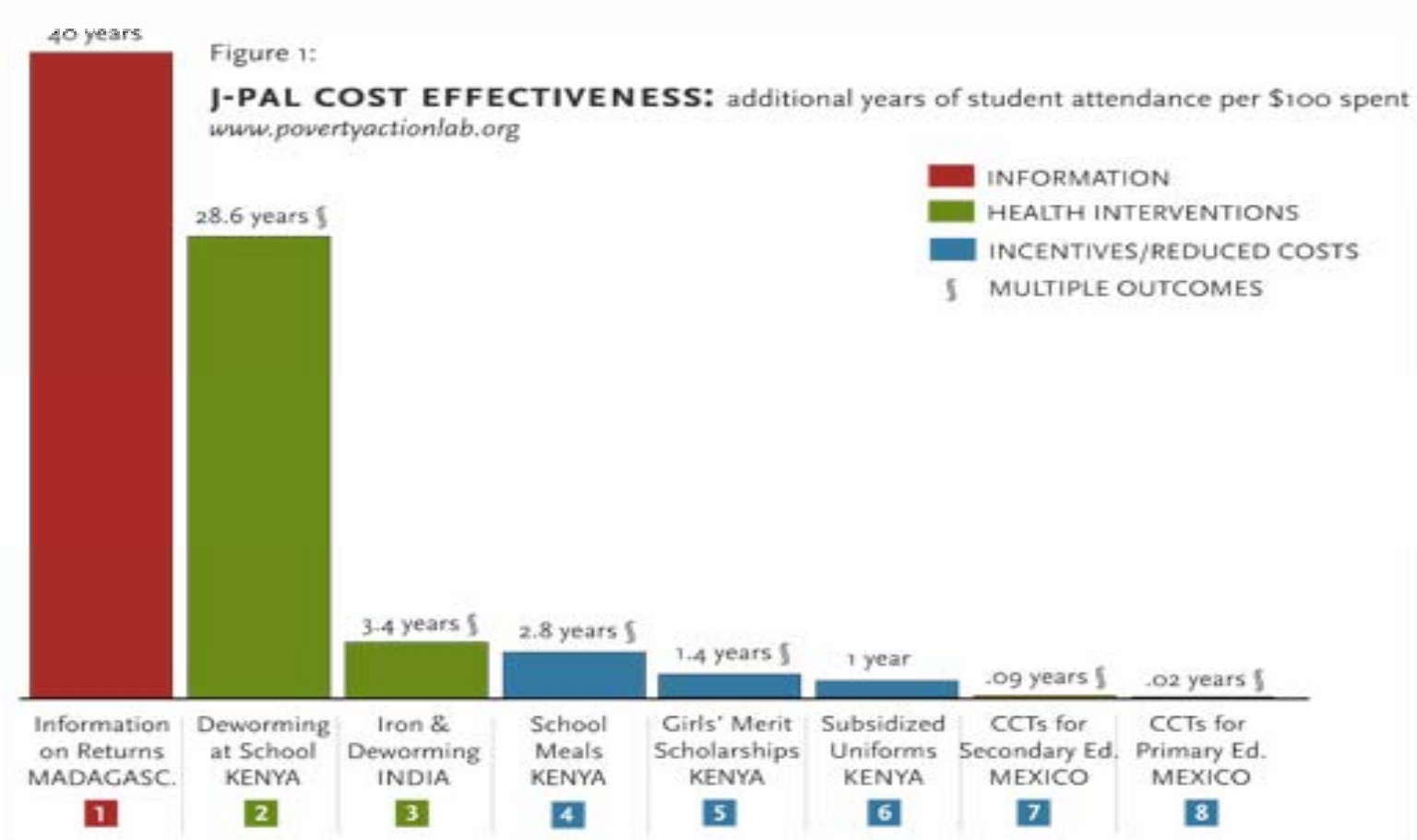
The Bigger Picture

What IE's tell us about ROI's?

1. Impact evaluations calculate an ROI by **compare programme to no programme**.
2. Another productive approach is been to focus on **relative ROIs**: the effect of programme 1 versus programme 2 versus programme 3 versus no programme.
3. A final, and supremely important, approach is the evaluation of programme 1 versus no programme in context A, context B, context C. Such **multi-country studies** not only validate relationships from a single place or pilot, but also enable us to calculate an average ROI.

Comparison of ROI's

Education Interventions



4. Example: Balsakhi Programme

- Implemented by Pratham, an NGO from India.
- Programme provided tutors (Balsakhi) to help at-risk children with school work.

Methods to estimate impacts

Let's look at different ways of estimating the impacts using the data from the schools that got a balsakhi

A. Pre — Post (Before vs. After)

B. Simple difference

C. Difference-in-difference

D. Other non-experimental methods

E. Randomised Experiment

A. Pre-post (Before vs. After)

Look at average change in test scores over the school year for the balsakhi children

Can this difference (26.42) be interpreted as the impact of the balsakhi programme?

Average **post-test** score for children with a balsakhi	51.22
Average **pretest** score for children with a balsakhi	24.80
Difference	**26.42**

B. Simple difference

Compare test scores of...

- Children that got balsakhi with
- Children that did not

Can this difference (-5.05) be interpreted as the impact of the balsakhi programme?

Average score for children with a balsakhi	**51.22**
Average score for children without a balsakhi	56.27
Difference	-5.05

C. Difference-in-Differences

Compare gains in test scores of...

- Children who got balsakhi

With gains in test scores of...

- Children who did not get balsakhi

Can 6.82 be interpreted as the impact of the balsakhi programme?

	Pretest	Post-test	Difference
Average score for children **with** a balsakhi	24.80	51.22	26.42
Average score for children **without** a balsakhi	36.67	56.27	19.60
Difference			6.82

D. Other Methods

There are more sophisticated non-experimental methods to estimate programme impacts:

- Regression
- Matching
- Instrumental Variables
- Regression Discontinuity

These methods rely on being able to "mimic" the counterfactual under certain assumptions

Problem: Assumptions are not testable

E. Randomised Experiment

Suppose we evaluated the balsakhi programme using a randomised experiment

- What would be the advantage of using this method to evaluate the impact of the balsakhi programme?

Impact of Balsakhi - Summary

Method	Impact Estimate
(1) Pre-post	26.42*
(2) Simple Difference	-5.05*
(3) Difference-in-Difference	6.82*
(4) Regression	1.92
(5)Randomized Experiment	5.87*

*: Statistically significant at the 5% level

Bottom Line: Which method we use matters!

5. Determining programme impact

Hypothesis Testing

If our results show a difference between the treatment and control group, this could be because:

A. The impact of the programme is real

B. The difference is due to chance and we can not reject null hypothesis of no impact.

Use statistics to calculate whether the difference is real or due to random chance

If it is very unlikely (less than a 5% probability) that the difference is solely due to chance:

– We “reject our null hypothesis”

We may now say:

– “our programme has a statistically significant impact”

Hypothesis testing: conclusions

If it is very unlikely (less than a 5% probability) that the difference is solely due to chance:

- We “reject our null hypothesis”

We may now say:

- “our programme has a statistically significant impact”

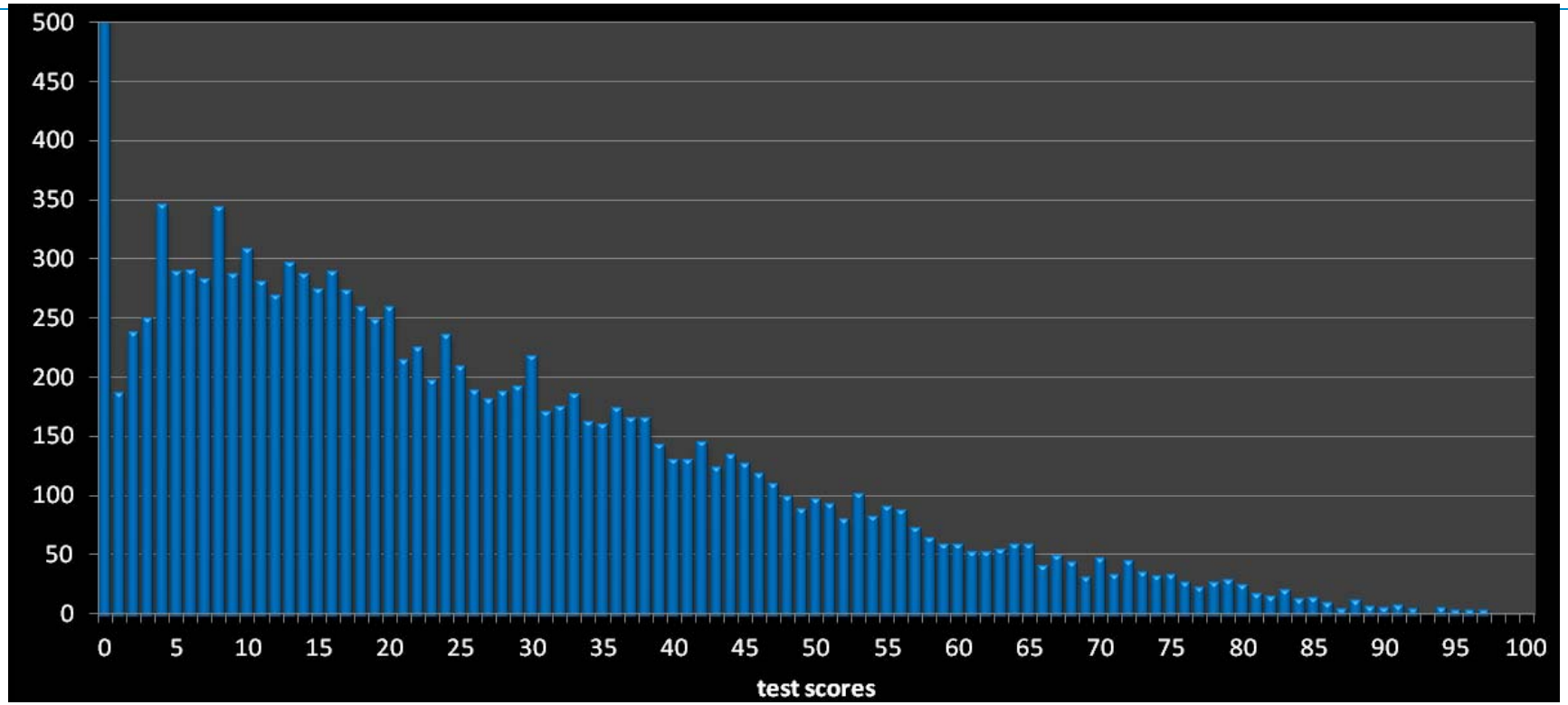
Hypothesis testing: conclusions

Are we now 100 percent certain there is an impact?

- No, we may be only 95% confident
- And we accept that if we use that 5% threshold, this conclusion may be wrong 5% of the time
- That is the price we're willing to pay since we can never be 100% certain

Because we can never see the counterfactual, We must use random sampling and random assignment, and rely on statistical probabilities

Back to the Balsakhi example: Baseline test score data



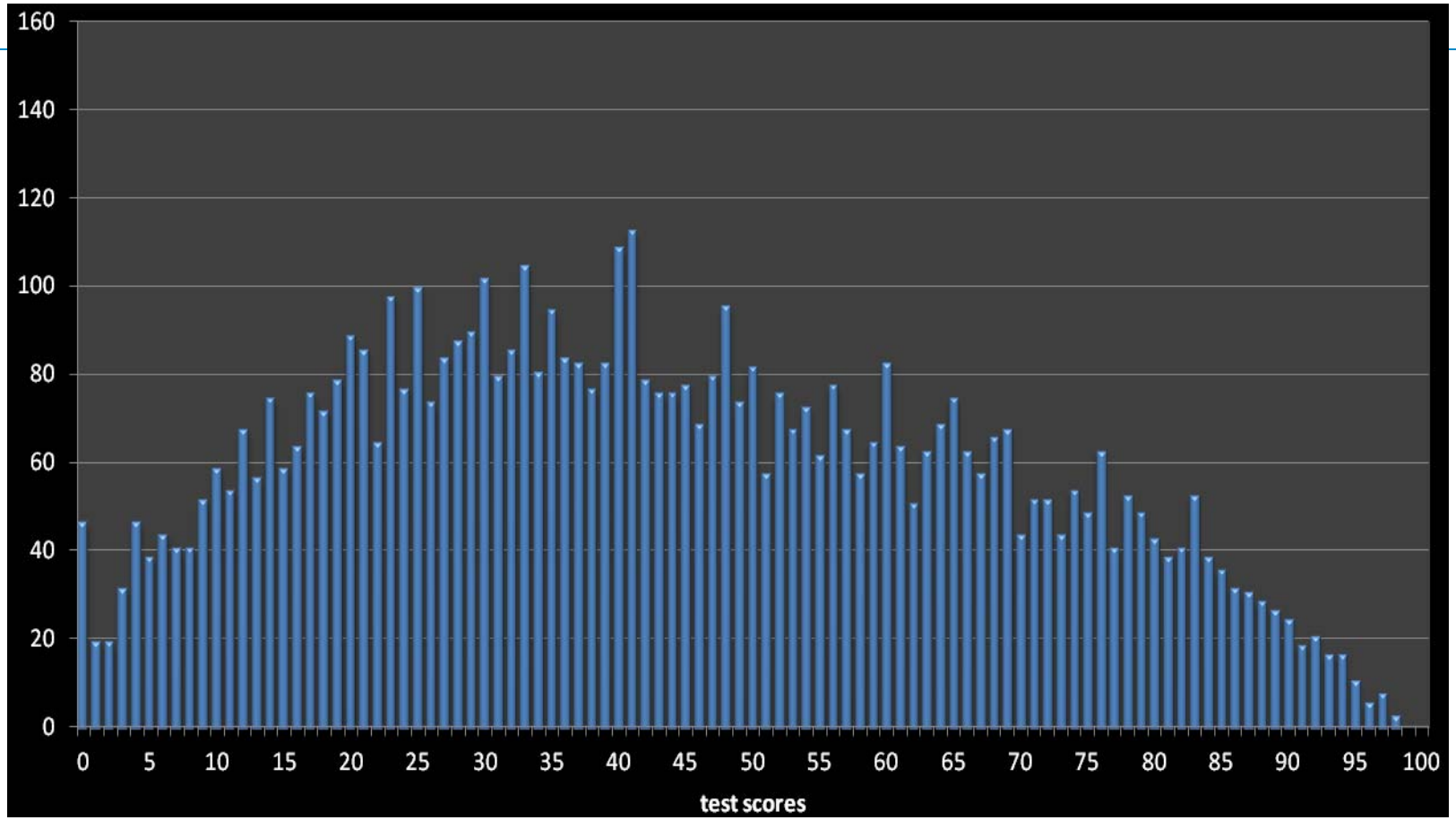
This was the distribution of test scores in the baseline.

The test was out of 100.

Some students did really well, most, not so well.

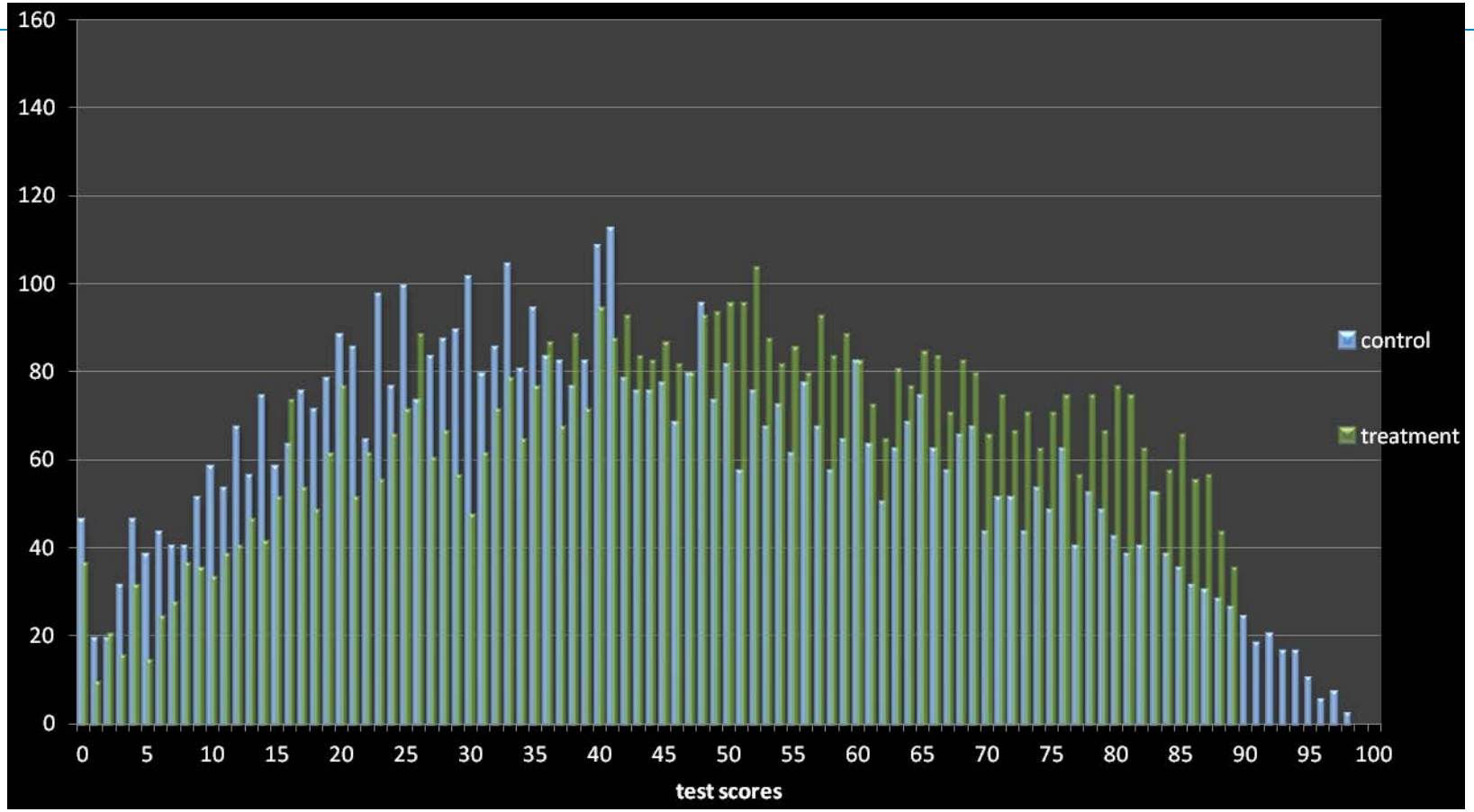
Many actually scored zero.

Endline test scores

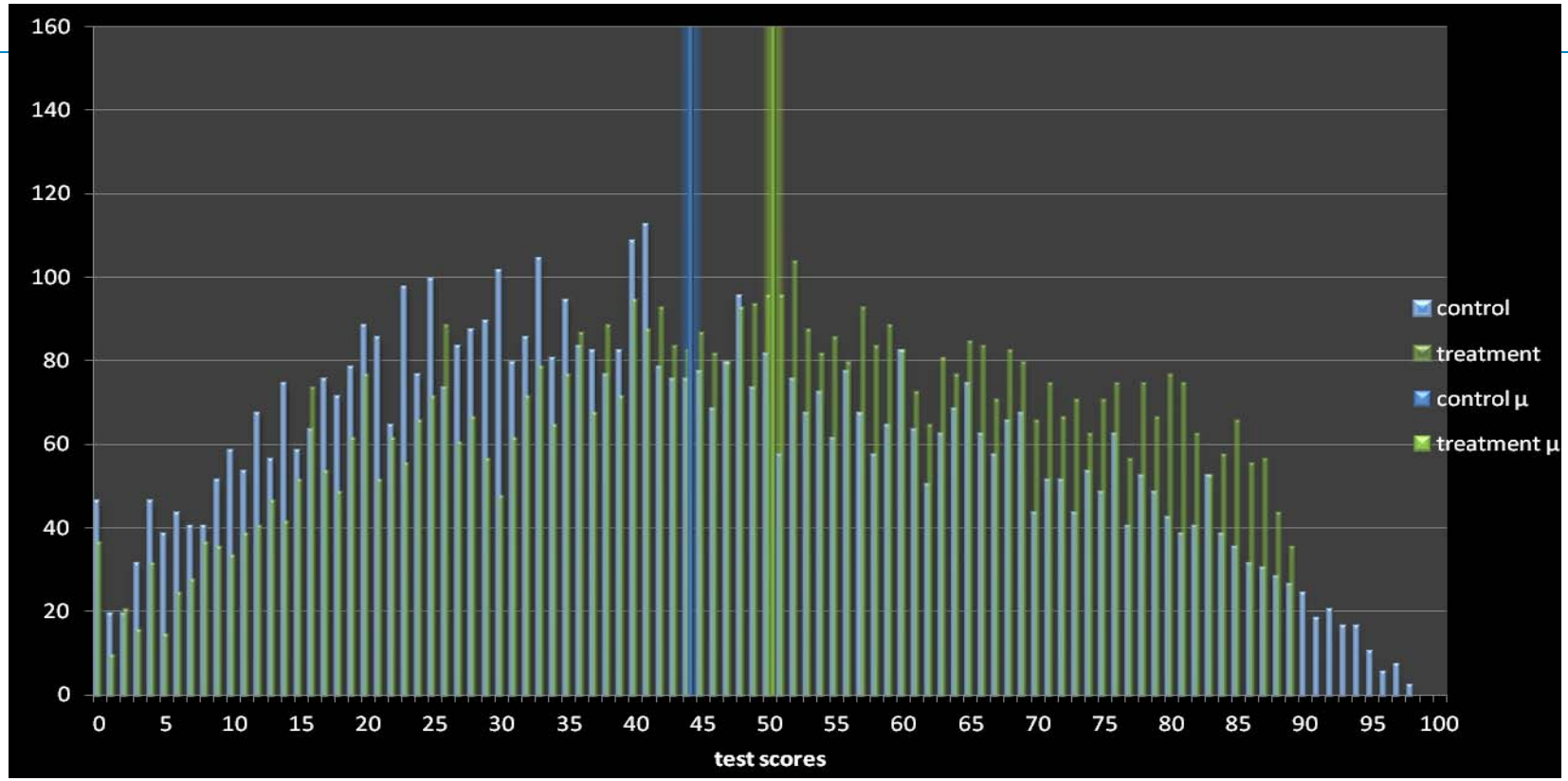


Now, look at the improvement. Very few scored zero, and many scored much closer to the 40-point range...

Post-test: control & treatment

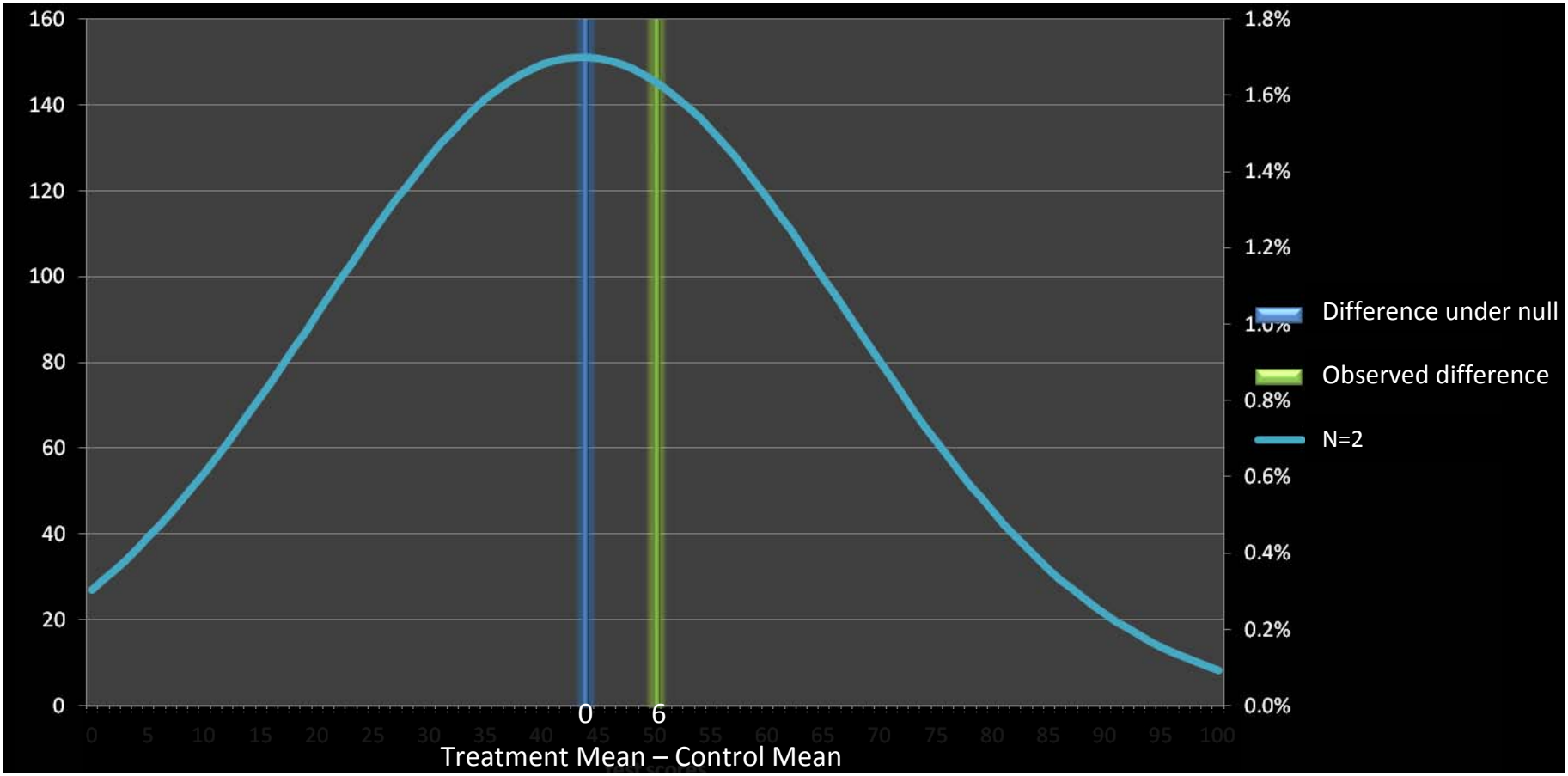


Average difference: 6 points



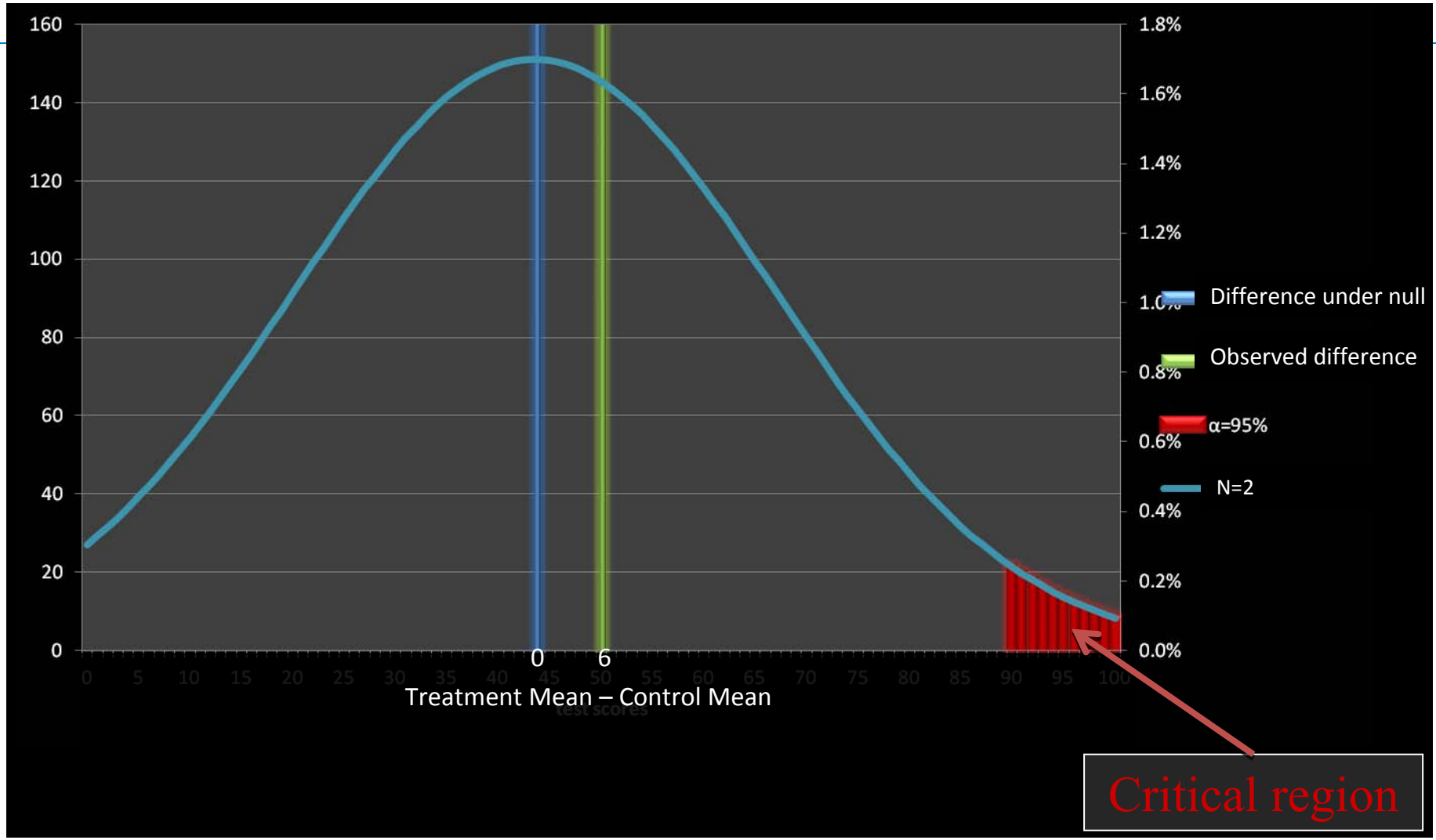
What's the probability that the 6 point difference is due to chance? (Testing statistical significance)

That probability of observing a 6 point difference when in fact Ho is true depends on sample size (here: N=2)

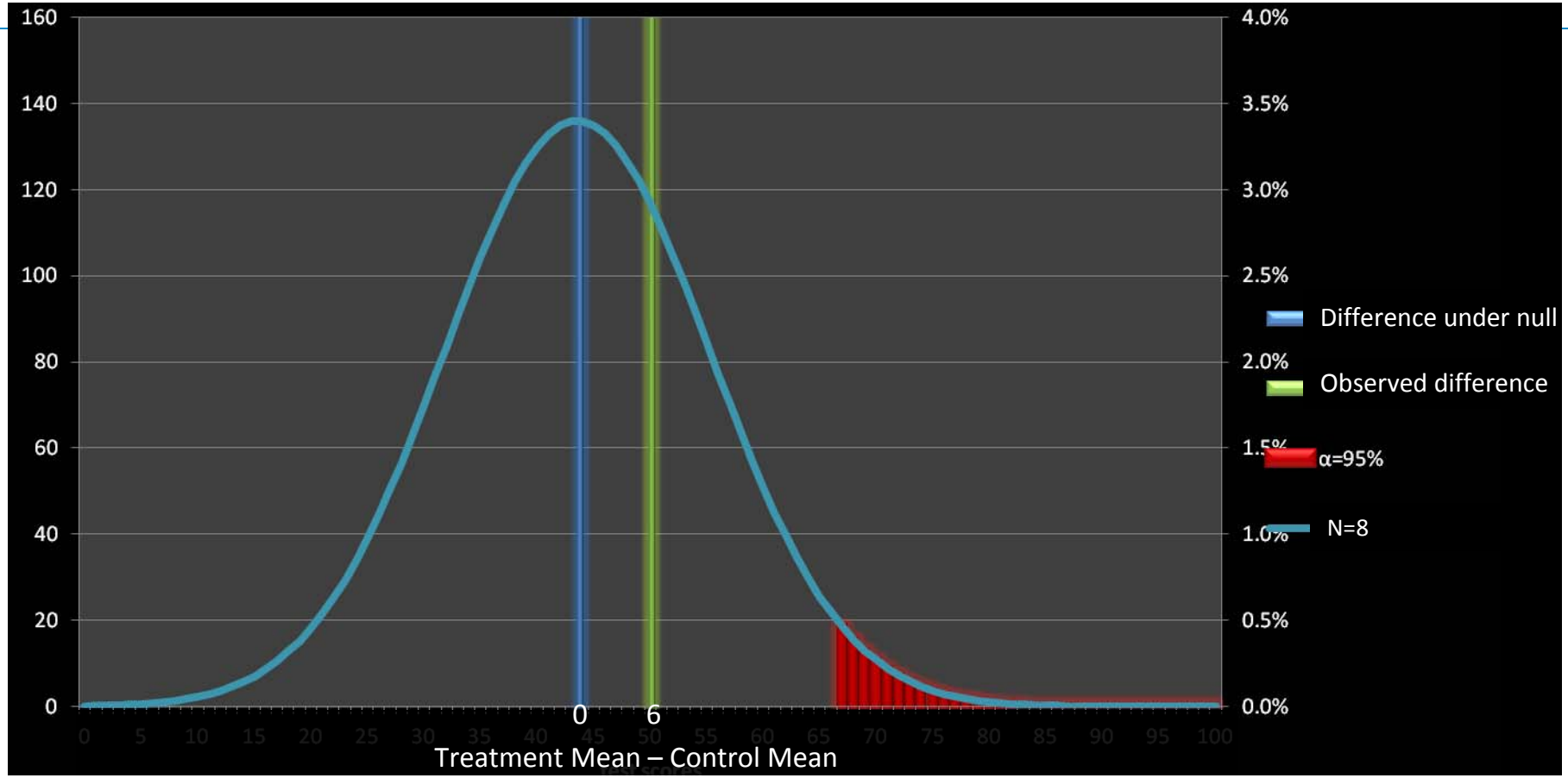


Q: How many children would we need to randomly sample to detect that the difference between the two groups is statistically significantly *different from zero*?

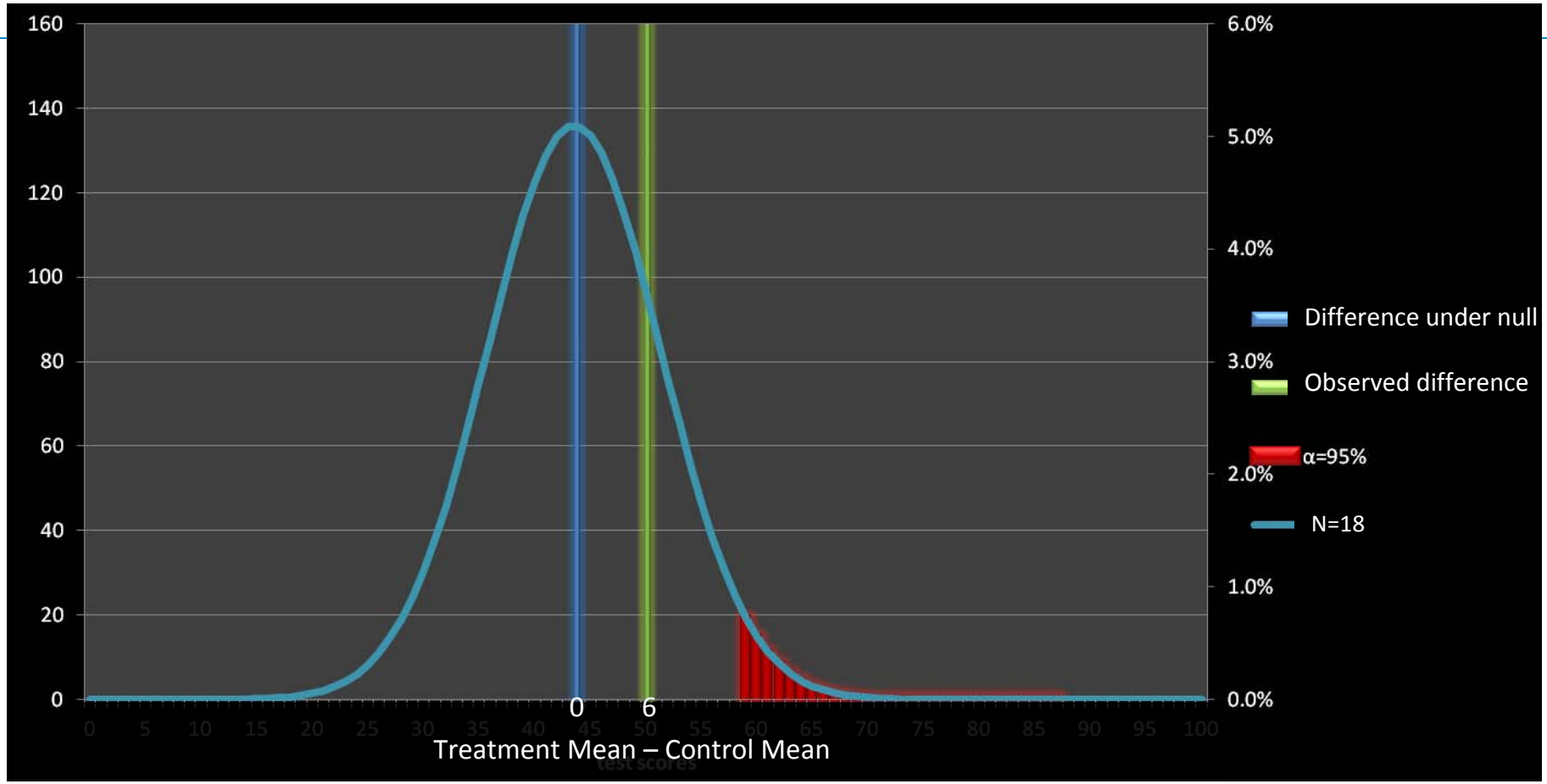
“Significance level” (5%)



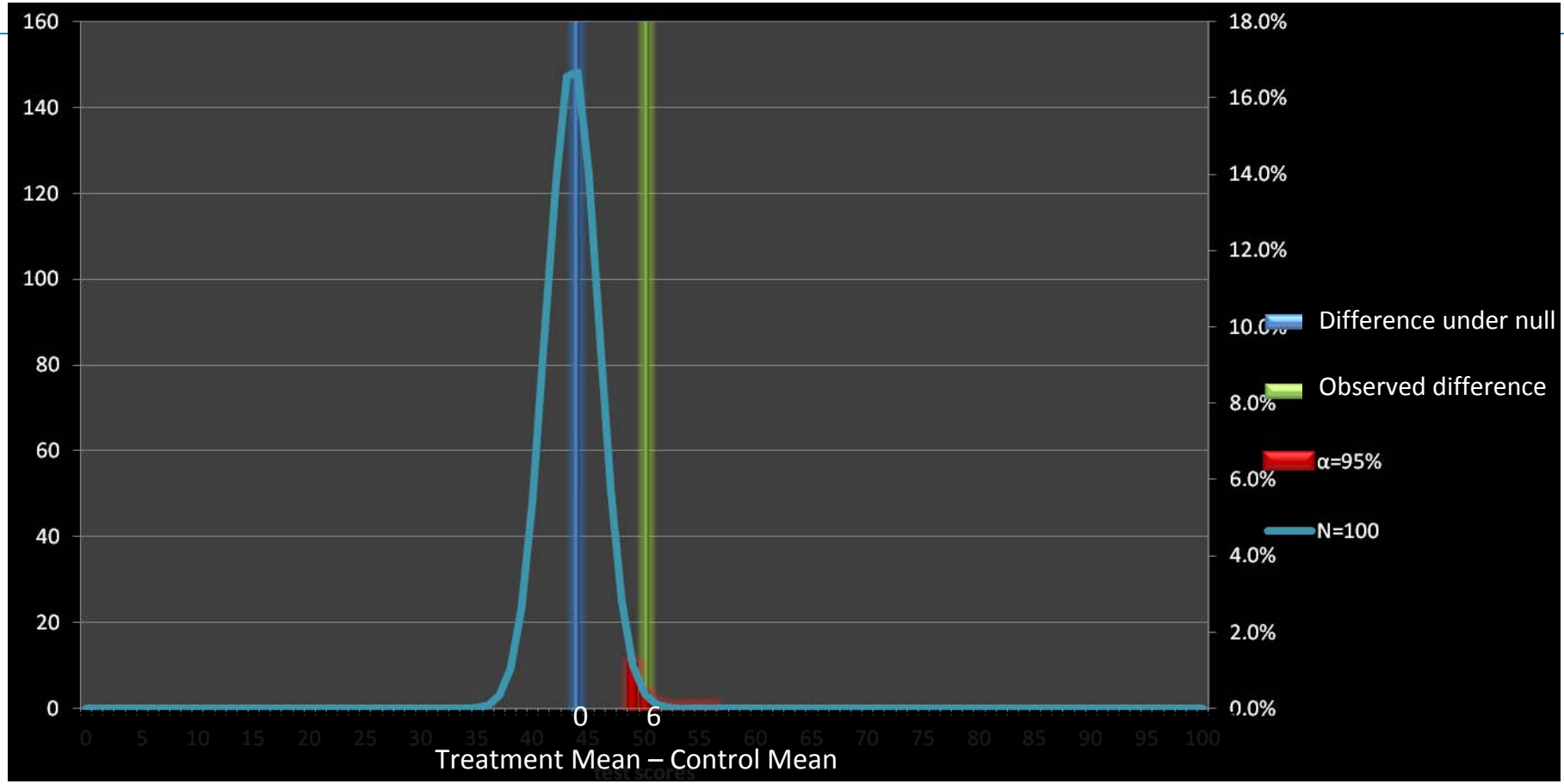
Significance: Sample size = 8



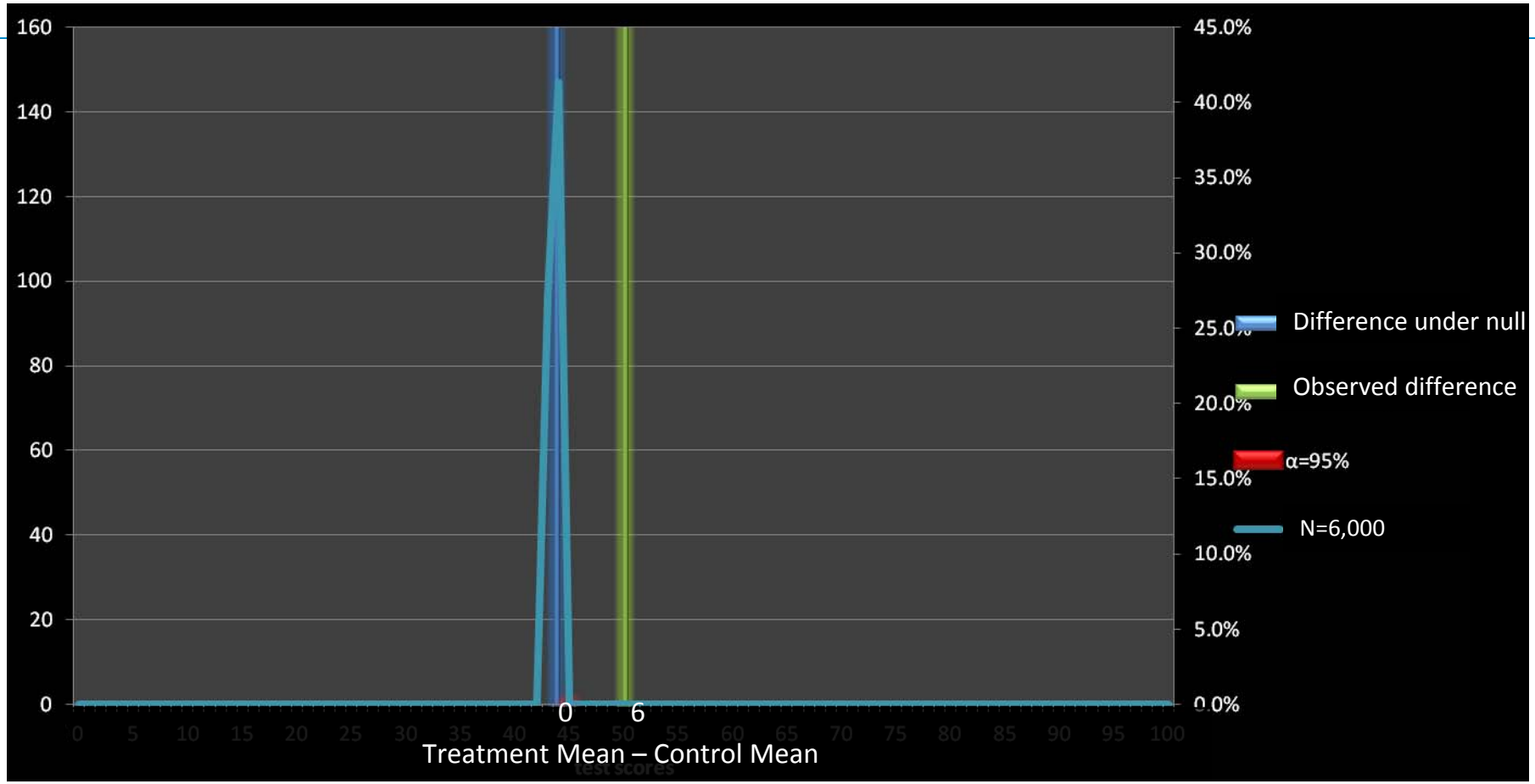
Significance: Sample size = 18



Significance: Sample size = 100



Significance: Sample size = 6,000



6. Concerns about Randomised Evaluations

1. **External Validity:** Programme impacts are so contextual that the results from one trial may be of little assistance to decision-makers in other times and places.
2. **The Why Question:** What part of the programme (processes) lead to the positive result? What people, processes, and programme factors were effective in achieving the result?
3. **Publication/Policy Interest Bias:** Papers with non-significant findings tend to have a harder time publishing while policy makers can often cherry pick evidence and ignore non-significant findings.
4. **Timeliness of Results:** Knowing the result of a programme after 1-3 years may not be helpful for decision making.
5. **Numbers v's stories:** Handing over the evaluation to a highly quantitative evaluation process can be difficult and the methodologies can be scary. Qualitative findings are more intuitive and can be more powerful tools of persuasion.

A Discussion about the Concerns

1. **External Validity:**
   - Understanding heterogeneous effects.
   - Making sure participants in the study are appropriate (representative of target group, globally relevant).
   - Test important theories/ideas not the programme.
   - Sector wide investment in core questions and systemic reviews.
   - Remains an issue but this is an issue for all social science research.
2. **The Why Question:**
   - Vary the programme components themselves experimentally to understand why a programme works, and what it reveals about the process of development.
   - Admittedly only a small number of programme components can be varied.

A Discussion about the Concerns

Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

3. **Publication/Policy Interest Bias:**
   - This is an issue the policymakers and academics and improve on. All stakeholders should be open to non results.
4. **Timeliness of Results:**
   - Development is a gradual process so intervention impacts on household consumption, micro entrepreneur profits and infection rates take time.
   - Easy to build in (a) monitoring and evaluation of administrative data, (b) qualitative activities and (c) outcome focused midlines.
5. **Numbers v's stories:**
   - Stories no longer enough for donors in an increasingly competitive environment.
   - Improvements required in the impact evaluation (quantitative and qualitative) skill set of the sector required.

7. Final Remarks

- Ireland lags behind the move towards more rigorous evaluation in development.
- Randomisation is something to be embraced, not feared.
- Certain questions lend themselves to RE, others don't. RE not a gold plated method but the best way to get precise estimates of impact (appropriate counterfactual) for certain types of questions.
- RE is not M&E, it's R&D: A means to innovate, learn and improve programmes.
- RE can help understand which part of an intervention is most important but only for a limited set of alterations.