PROGETTO VALUTAZIONE

APPLICA
KNOW-HOW EXPERIENCE APPLICATION

Developing Quality Terms of Reference for Impact Evaluation

Training seminar

Valletta, Malta – 5-7 October 2016

Three-days on how to write BETTER Terms of Reference (ToRs) for Evaluation

European Commission

OFFICE OF THE DEPUTY PRIME MINISTER
MINISTRY FOR EUROPEAN AFFAIRS

PARLIAMENTARY SECRETARIAT
FOR THE EU PRESIDENCY 2017 AND EU FUNDS

DO WE NEED BETTER TORs FOR THE EVALUATION OF STRUCTURAL FUNDS?

YES, WE DO

What is so special about TORs?

TORs perform two distinct roles, involving different actors and requiring different expertise

1. Apply a mechanism to **choose one** evaluator among the bidders
2. Make sure that **the winning bidder delivers** useful evaluation products

THUS WRITING A TOR IS BOTH

A **STEP IN THE PROCUREMENT**
FOR EVALUATION SERVICES

AND **A STEP TOWARD THE DESIGN**
OF AN EVALUATION

ideally, these steps should be separate and done by different people with very different expertise: in practice the viewpoints of the procurement people tend to prevail.

The often inappropriate mix of procurement procedure and social science design is not **the only source of complaint** about ToRs

Often the social science part alone is really bad

For example, Patricia Rogers

of the University of Melbourne, where she is Professor of Public Sector Evaluation

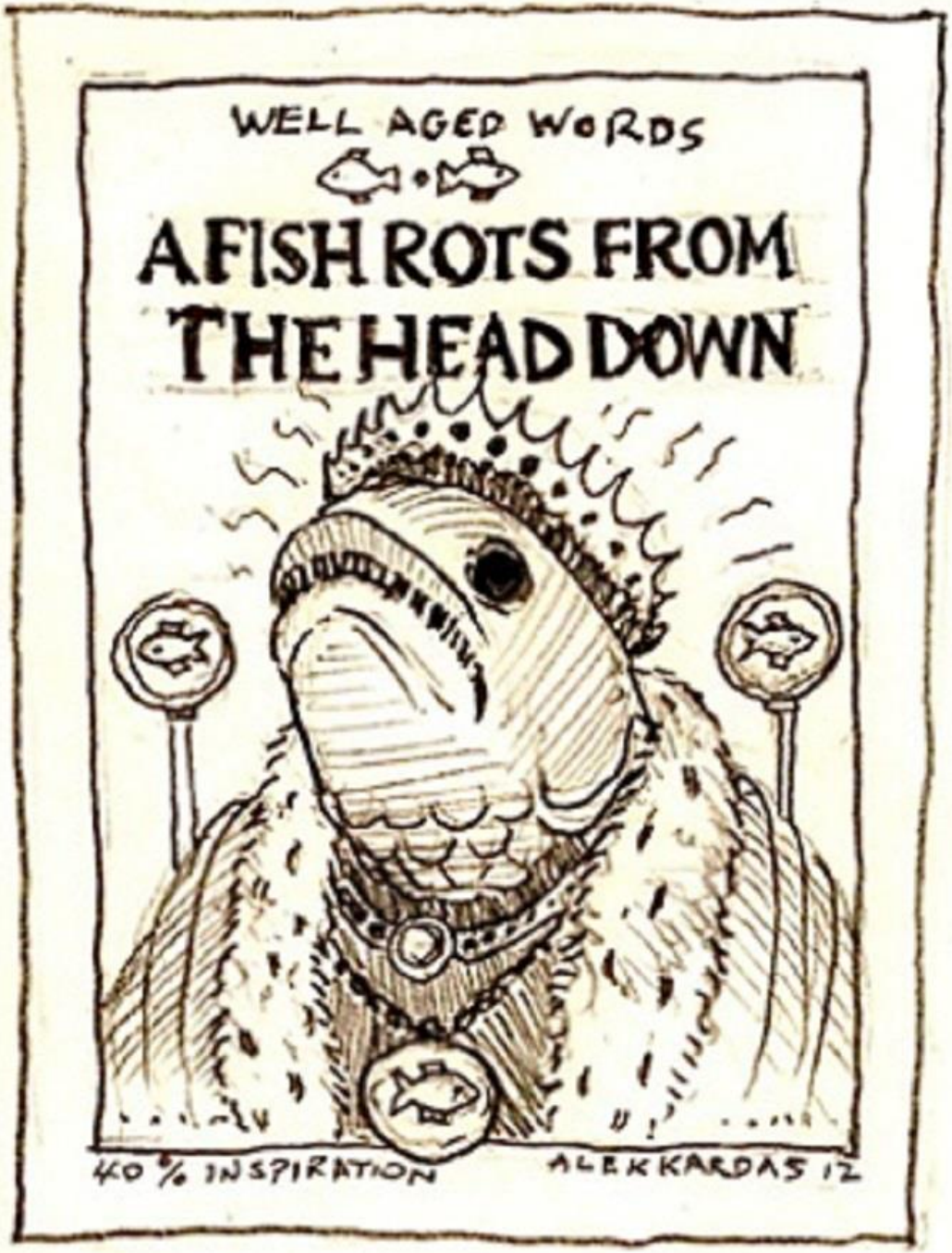
wrote about TORs

Many problems with evaluation can be traced back to the TOR:

“Many TORs are
too ambitious,
too vague,
inaccurate,
or not appropriate”

A TOR CANNOT BE BETTER THAN THE IDEA OF EVALUATION BEHIND IT.

Many definitions of the purpose of evaluation are too ambitious, too vague, inaccurate, or not appropriate.


WELL AGED WORDS
A FISH ROTS FROM THE HEAD DOWN
40% INSPIRATION
ALEKKARDAS 12

MOST DEFINITIONS OF EVALUATION ARE

SEQUENCES OF IMPRESSIVE BUT OFTEN VAGUE BUZZWORDS

TAKE THE DEFINITION OF EVALUATION PRODUCED BY OECD

"The systematic and objective assessment of an on-going or completed project or programme, its design, implementation and results. The aim is to determine the ***relevance*** *and fulfillment of objectives:* ***efficiency, effectiveness, impact, sustainability****."*

WE CAN REFER TO THESE AS THE "PARIS FIVE"

The EU toolbox FOR BETTER REGULATIONS identifies these

efficiency, effectiveness, relevance, EU-added value, coherence

Your first assignment will be to write a short TOR

we provide you with a short description of the program to be evaluated

your job is to give a definition of each of the
Paris-five or of the Brussels five

And show *how* and *whether*
they are useful in writing a TOR and specifically
in formulating
some evaluation question

The following slides are for closing the first day, after the plenary discussion of the first drafts of the group–level TORs

How do you get from 5 to 1
Semiserious conclusions

IF YOU DROP RELEVANCE, WHICH LIVES IN A WORLD OF ITS OWN
CAN REFER TO THE REMAINING FOUR AS THE “FAB FOUR”

efficiency, effectiveness,
impact, sustainability

IF YOU THINK THAT SUSTAINABILITY DOES NOT BELONG
WITH THE OTHER THREE
DROP IT AND YOU HAVE

“THE HOLY TRINITY OF EVALUATION”

efficiency, effectiveness, impact

efficiency, effectiveness……, and impact

Efficiency and effectiveness are the **inseparable duo** in every document in every Public Administration the world over.

Since they should mean everything desirable, they end up meaning nothing.

Most people think they must mention the two names together.

THERE IS SOME PARALLELISM with Ogino-Knaus. Nothing more wrong.

efficiency, effectiveness……, and impact

Ogino was a japanese doctor interested in increasing the *possibility of conceiving*.

Knaus was a German doctor interested in contraception.

The numbers of fertile days calculated by the two researchers were consistent with each other, so counting the days was called **OGINO-KNAUS** , but the two doctors never worked together.

Thus if we also eliminate

efficiency, effectiveness we are left with impact

So the topic is indeed **writing a TOR for IMPACT EVALUATION**

The ‘old school’ and the
IMPACT INDICATORS

THE TROUBLE WITH INDICATORS 1

Sky is the limit…

In fact you can find indicators just about anything,
coherence indicators, efficacy indicators,
economy, validity, reliability,
transparency indicators

22,600 indicators!

In 2006, for ERDF programmes (Objective 1 and 2), a total of 22,600 indicators were reported across 227 programmes, with an average of 106 indicators per programme, ranging from 25 in Denmark to 192 in Italy.

Over 100 indicators suggests a lot of counting but also dispersed effort and a lack of concentration.

- Of these 22,600 indicators
- 94% had final achievements
- 58% had targets;
- 6% had baselines;
- 55% had targets and achievements;
- 5% had baselines, targets and achievement

THE TROUBLE WITH INDICATORS 2

If you don't compare an indicator
with something,
it won't mean anything

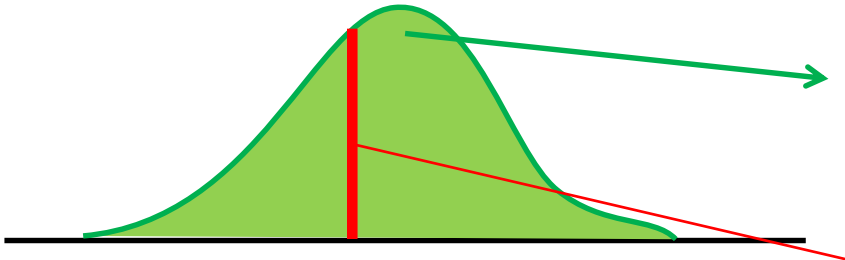
- 58% had targets;
- 6% had baselines;
- 55% had targets and achievements;
- 5% had baselines, targets and achievement

Some comparisons are more problematic than others..

Goodhart‘s Law

“Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes”.

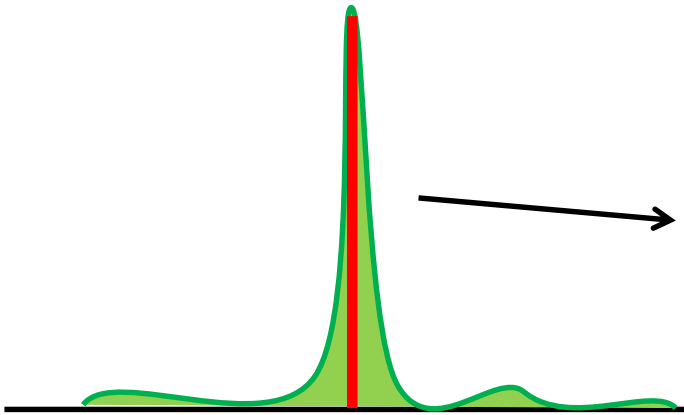
Goodhart‘s Law



Field of interest of our policy (eg. education)

Indicator (eg. test results)

When not incentivised, test results could be a good proxy of education.



When incentivised, teachers start to teach only what is in the tests, help students with the tests or openly cheat on the results.

Test results are not a good proxy of education anymore. Here test results grow, when the total amount of education shrinks.

Multi-purpose use of information

There are many reasons why we collect (monitoring or in general performance) informations:

- To budget/plan, to control, to evaluate, to promote, to learn, to motivate…
- …ultimately, to improve, to do our job in a better way.

As a consequence of Goodhart‘s law we cannot use the same information for some combinations of above mentioned tasks.

Essentially, measurement for budgeting/planning and for control/reporting should be totally separate from measurement for learning (*including via evaluation*) and improvement!

VERONICA GAFFEY,
WHO WROTE ONE OF THE ARTICLE WE WILL DISCUSS
AT SOME LENGTH (AND DEPTH) TOMORROW

AT SOME POINT MAKES THESE DISARMING REMARKS

But we need to ask why we are gathering all this data. Much of the data are not meaningful.
Their purpose seems primarily to provide a figure which will unlock the final payment from the Structural Funds, rather than being an exercise in accountability and learning.
We knew that impact indicators were not delivering much meaningful information – but we did not have sufficient knowledge at that stage to challenge them more radically.

We needed indicators so that some aggregate figures could be generated at EU level to communicate the achievements of the policy.

AH-HA!

DESPITE ALL THE RHETORIC AND THE HYPE, ALL IT WAS REALLY NEEDED WERE SOME AGGREGATE FIGURES TO COMMUNICATE THE ACHIEVEMENTS OF THE POLICY.

ONE NEEDS NO EVALUATION FOR THAT, WE COULD ALL HAVE AVOIDED MAKING SO COMPLICATED TORS

NOW THINGS HAVE CHANGED AGAIN AND NOW WE NEED SERIOUS STUFF, LIKE IMPACT EVALUATIONS

TOMORROW we will start with a group discussion of Howard White's article
A Contribution to Current Debates in Impact Evaluation