Replicability and generalizability of economic experiments results

Maria Espinosa, Marie Ferré

Outline of the session

- What does it mean replicability and generalizability?

- Presentation of the studies:

  External validity of experiments in environmental economics: framing and subject pool effects among students and professionals (Ferre et al,.)

  Feasibility Study on the Valuation of Public Goods and Externalities in EU Agriculture (Madureira et al., 2015)

- Open discussion on how to improve the replicability/generalizability of economic experiments?

What does it mean replicability/generalizability?

- **Reproducibility**

Data reproducibility means that Researcher B (e.g. the reviewer of a paper) obtains exactly the same results (e.g. statistics and parameter estimates) that were originally reported by Researcher A (e.g. the author of that paper) from A's data when following the same methodology (Asendorpf et al., 2013)

- **Replicability**

Replicability means that the finding can be obtained with other random samples drawn from a multidimensional space that captures the most important facets of the research design. In psychology, the facets typically include the following: (a) individuals (or dyads or groups); (b) situations (natural or experimental); (c) operationalizations (experimental manipulations, methods, and measures); and (d) time points.

- **How to measure replicability?**

- *"Establishing whether a finding is quantitatively replicated is more complex than it might appear (Valentine et al., 2011)".*

When both studies show significant effects, but effect sizes are very different, has the effect been replicated?

- There is no single standard for evaluating replication success. However more often it is used: significance and *P* values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes.

- **Evidence of replicability**
- **What is your experience in replicating other studies/ in having your research replicated?**

Evaluating replicability of laboratory experiments in economics

Colin F. Camerer[1,*,†], Anna Dreber[2,†], Eskil Forsell[2,†], Teck-Hua Ho[3,4,†], Jürgen Huber[5,†], Magnus Johannesson[2,†], Mi...
+ See all authors and affiliations

"The reproducibility of scientific findings has been called into question. To contribute data about reproducibility in economics, we replicate 18 studies published in the American Economic Review and the Quarterly Journal of Economics in 2011-2014. All replications follow predefined analysis plans publicly posted prior to the replications, and have a statistical power of at least 90% to detect the original effect size at the 5% significance level. We find a significant effect in the same direction as the original study for 11 replications (61%); on average the replicated effect size is 66% of the original. The reproducibility rate varies between 67% and 78% for four additional reproducibility indicators, including a prediction market measure of peer beliefs".

- **Generalizability**

  Generalizability of a research finding means that it does not depend on an originally unmeasured variable that has a systematic effect.

"To summarize, data reproducibility is necessary but not sufficient for replicability, and replicability is necessary but not sufficient for generalizability"

Presentation of the studies

JRC SCIENTIFIC AND POLICY REPORTS

Feasibility Study on the Valuation of Public Goods and Externalities in EU Agriculture

http://ftp.jrc.es/EURdoc/JRC83468.pdf

Authors

Lívia Madureira
Universidade de Trás-os-Montes e Alto Douro (UTAD)

Vila Real, Portugal
Email - lmadurei@utad.pt

José Lima Santos
Instituto Superior de Agronomia, Universidade Técnica de Lisboa (ISA/UTL)
Lisboa, Portugal
Email - jlsantos@isa.utl.pt

Editors

Maria Espinosa
Sergio Gomez y Paloma

Objetive of the study

- To avoid the policy failure of public goods , there is an increasing demand for the economic valuation of changes in multiple PGaE of agriculture

To develop a methodological approach for estimating the society's value of Public Goods and Externalities (PGaE) produced by EU Agriculture

- This policy need was raised again (in 2015) where the European Court of Auditors (ECA) carried out an audit of DG AGRI and ESTAT, entitled: "*Is the Commission's system for performance measurement in relation to farmers' incomes well designed and based on sound data?*"

- examine whether the EAAs can be further developed to provide a reasonable estimate of the economic value of the public goods that are produced by farmers => on-going

Challenges in developing frame-work

To address this policy demand, the required valuation framework needs to:

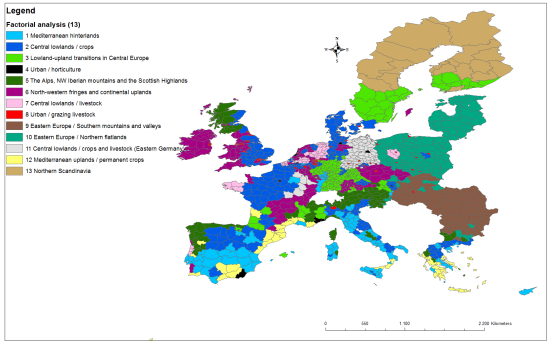
- be empirically-based and policy-relevant, that is focused on available policy options at this broad, supranational scale;
- be understandable by the general public of many involved countries whose values are to be surveyed;
- provide context-rich valuation scenarios, which lead people to engage in economic trade-offs instead of giving symbolic reactions to abstract scenarios
- take into account substitution effects across goods and services, in order to avoid aggregation biases when valuing changes in multiple PGaE.

- Example of choice set (micro=my thesis)

	Alternat. A	Alternat. B	Alternat. C
Surface	50 % eligible surface	Free to choose	Status Quo option
Grazing	Allowed	Not allowed	
Technical training & advisory service compulsory & free of charge	No	Yes	
Fixed payment:1000 €	No	Yes	
Premium (€/ha & year)	60	80	

- Example of choice set (macro=this study)

Programme providing services …	No application	Option A	Option B
 Landscape conservation	0 %	0 %	100 %
 Biodiversity conservation	0 %	100 %	0 %
 Soil erosion control	0 %	50 %	50 %
 Fire risk reduction	0 %	100 %	0 %
 Increase in taxes payments (annually for 5 years)	0 €	3 €	21 €



Land Use Policy 53 (2016) 56–70

Contents lists available at ScienceDirect

Land Use Policy

journal homepage: www.elsevier.com/locate/landusepol

Building an empirically-based framework to value multiple public goods of agriculture at broad supranational scales

José Lima Santos[a,*], Lívia Madureira[b], Ana C Ferreira[b], Maria Espinosa[c], Sergio Gomez y Paloma[c]

STEP 1: Select Public Goods & Externalities delivered by the EU Agricultural Sector

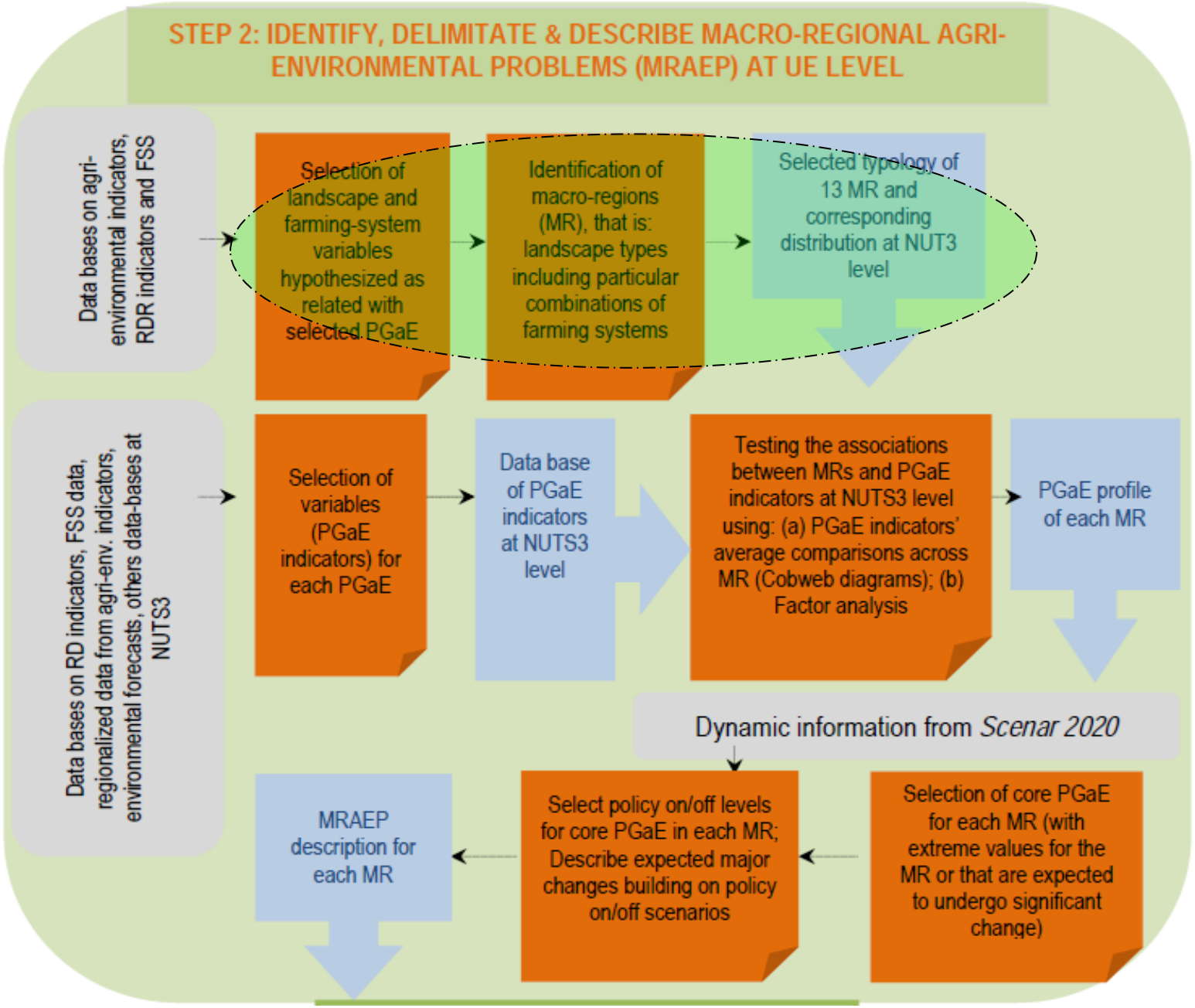
STEP 2: Identify, delimitate & describe macro-regional agri-environmental problems (MRAEP) at EU LEVEL

STEP 3: Design choice experiment survey and test it at pilot scale

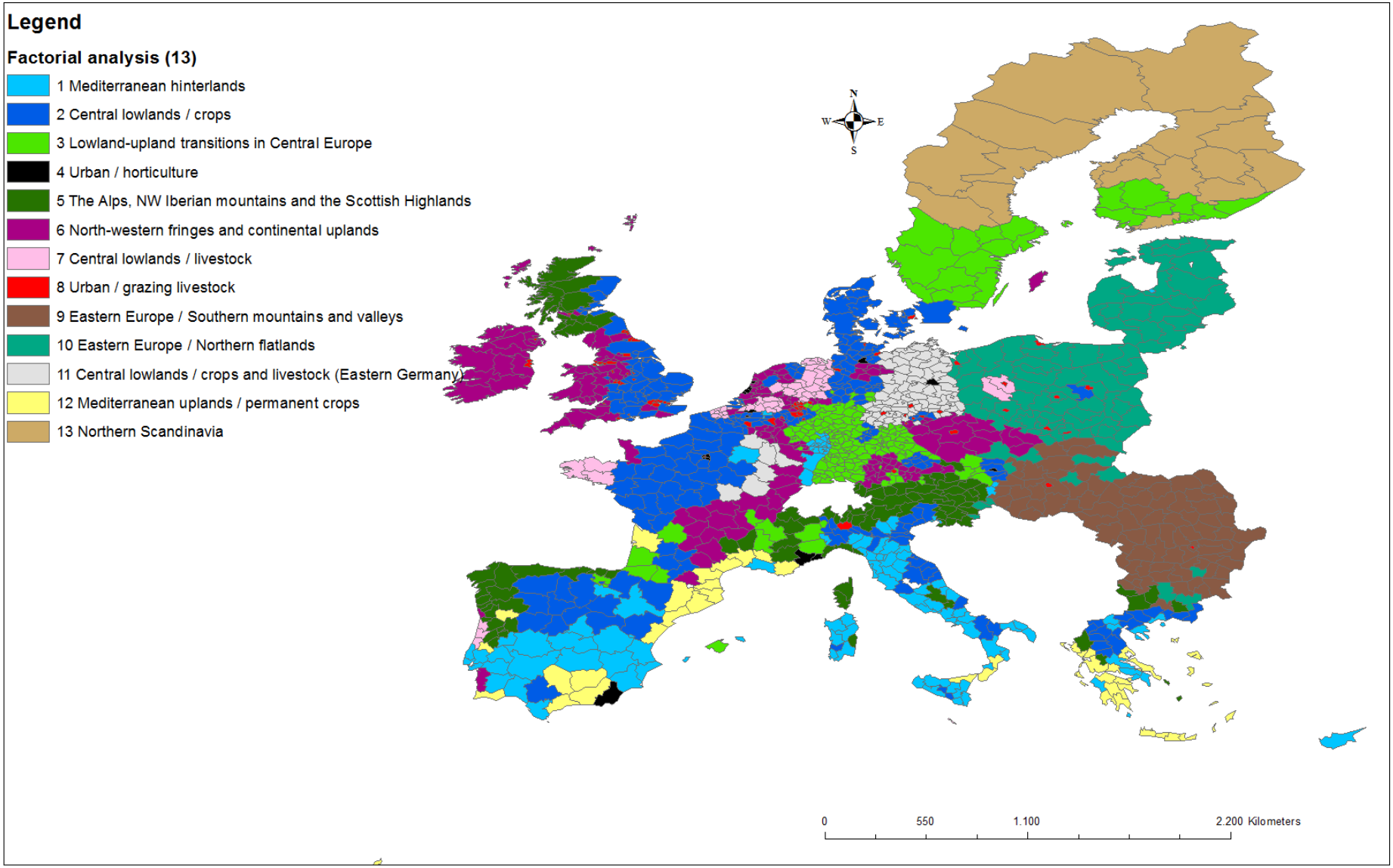
STEP 1: Select Public Goods & Externalities delivered by the EU Agricultural Sector

PGaE	PGaE indicators
Cultural landscape	Recreation potential index Cultural heritage
Farmland biodiversity	HNVF
Water quality	Total N input
Water availability	Infiltration Irrigated UAA
Soil quality	Soil erosion
Air quality	Total NH_3 emissions
Climate stability	Soil carbon content
Flooding resilience	Flooding risk
Fire resilience	Fire risk

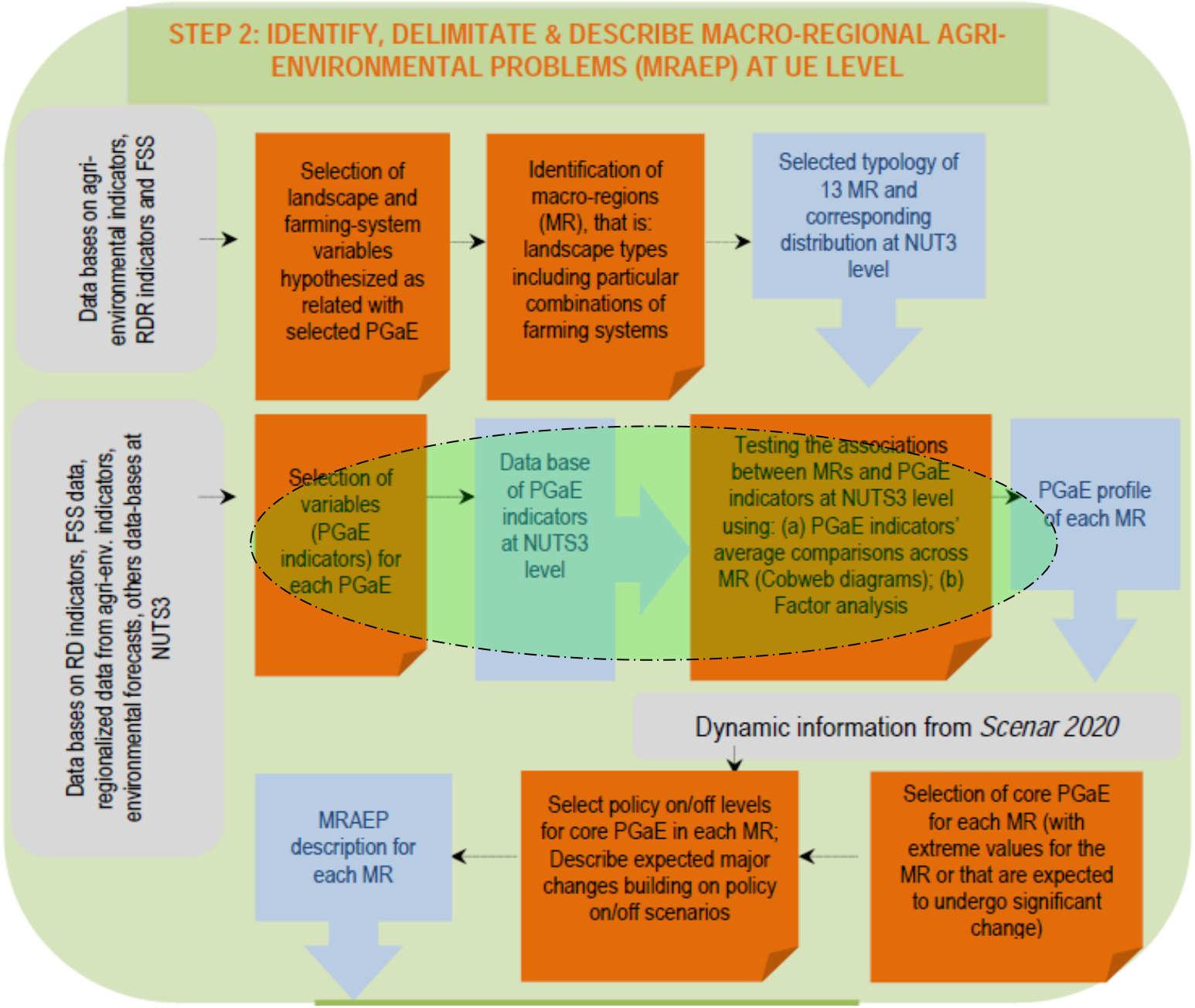
Designing MRAEP as valuation scenarios



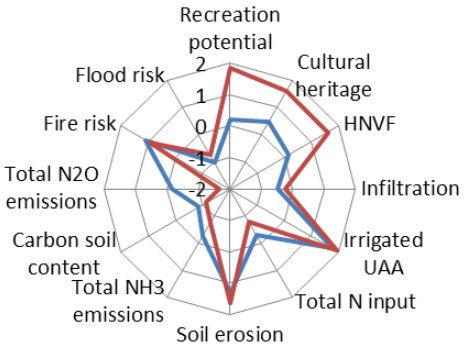
1. Identifying & describing MR



Designing MRAEP as valuation scenarios

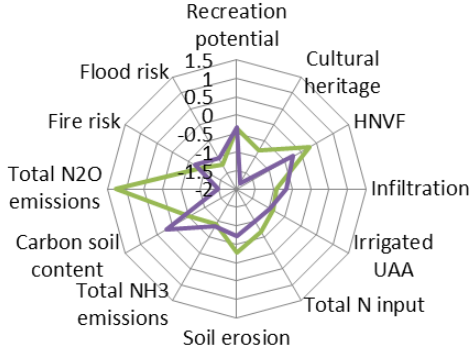


Identifying the current bundles of PgaE in each MR



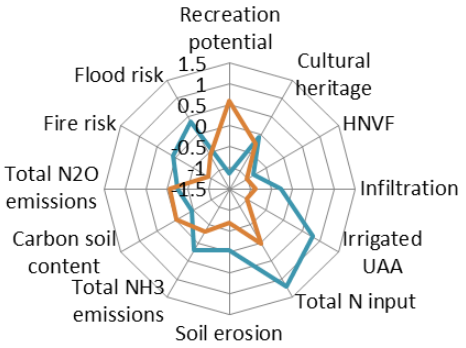
1 1 Mediterranean hinterlands

12 12 Mediterranean uplands / permanent crops



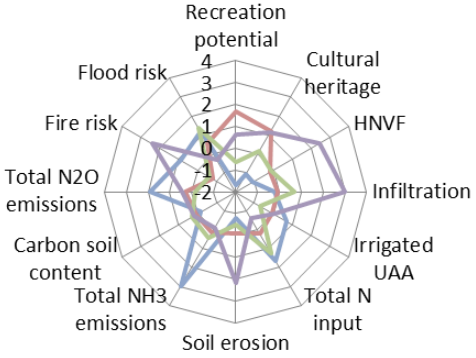
9 Eastern Europe / Southern mountains and valleys

10 Eastern Europe / Northern flatlands



2 Central lowlands / crops

11 Central lowlands / crops and livestock (Eastern Germany)

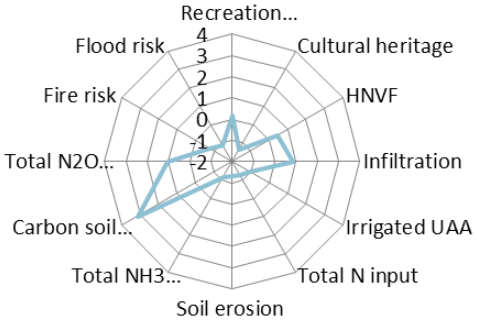


7 Central lowlands / livestock

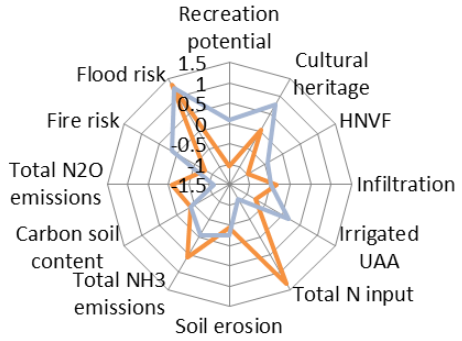
3 Lowland-upland transitions in Central Europe

6 North-western fringes and continental uplands

5 The Alps, NW Iberian mountains and the Scottish Highlands



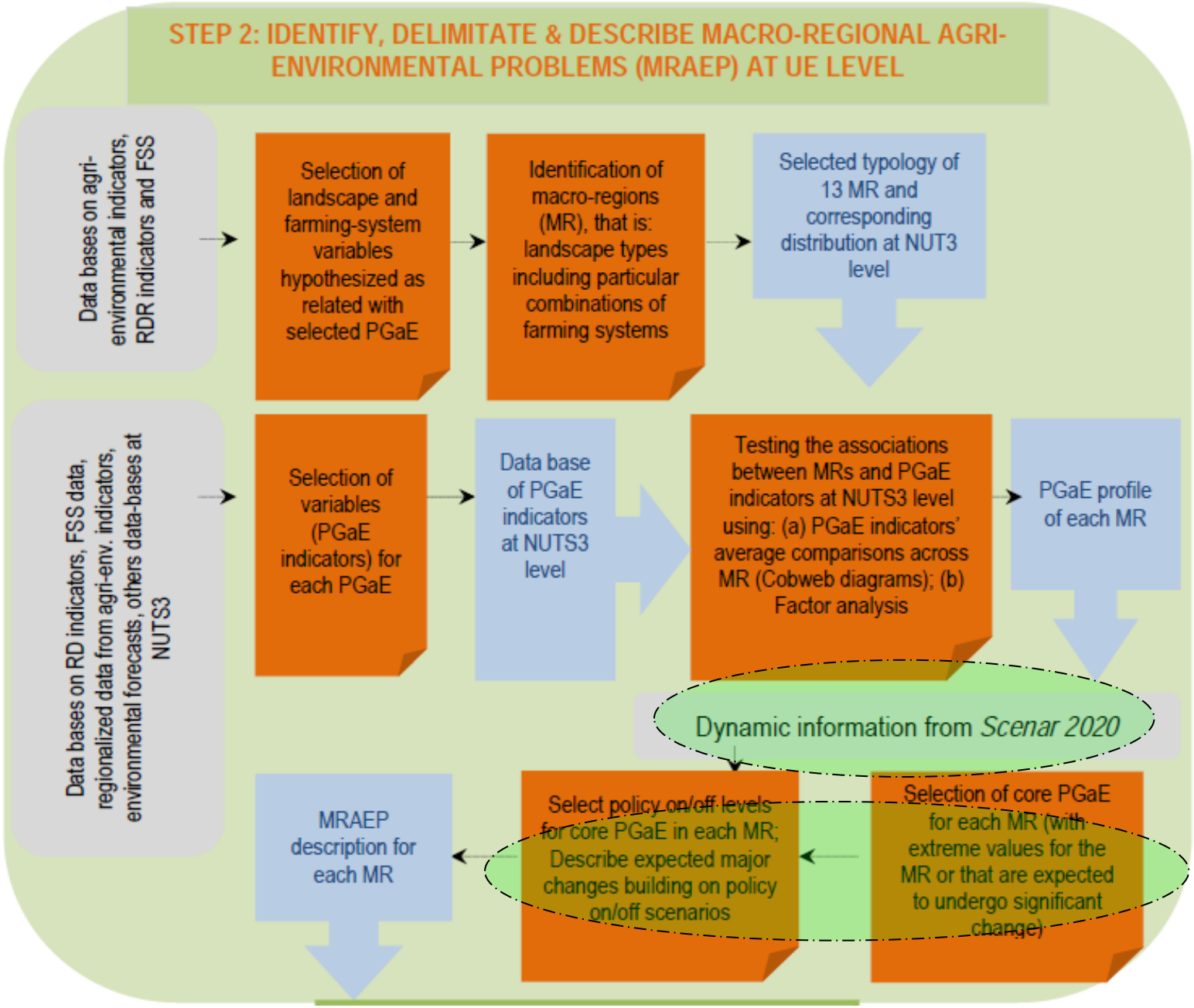
13 Northern Scandinavia



8 Urban / grazing livestock

4 Urban / horticulture

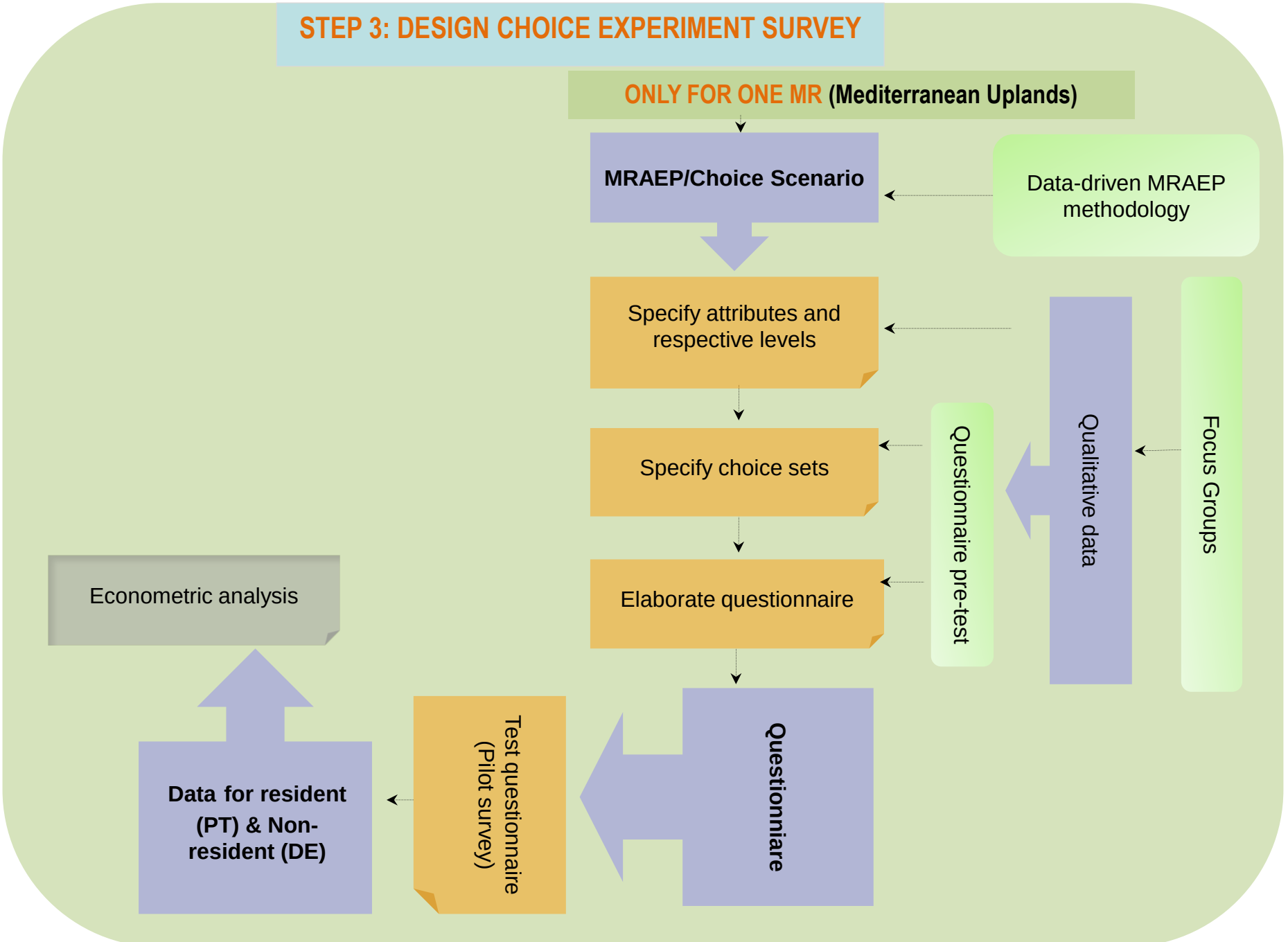
Designing MRAEP as valuation scenarios



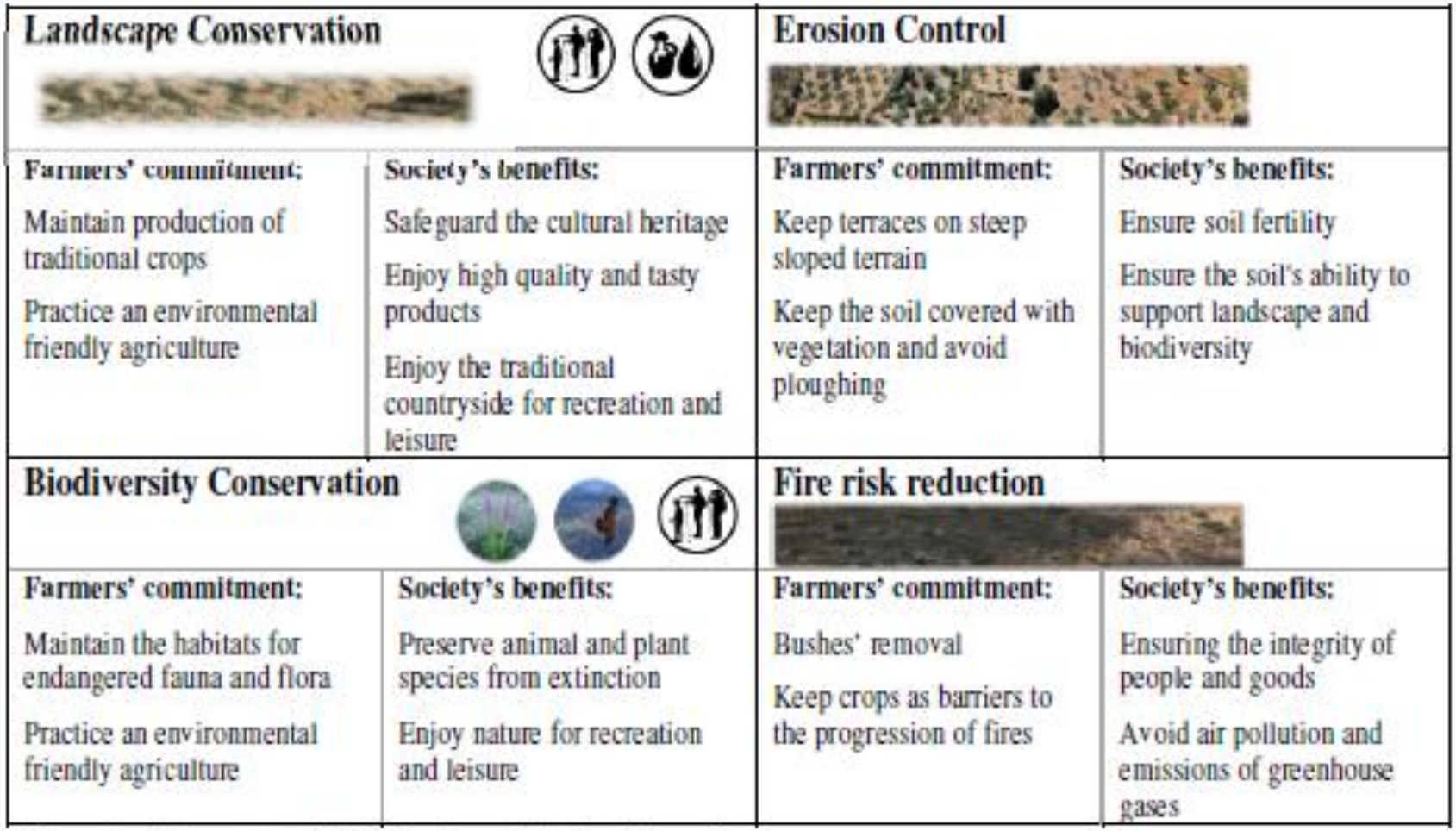
Set of PGaE to be valued in each MRAEP – selected example

Table 1 – Farmland abandonment in Mediterranean uplands/permanent crops (MR12)

PGaE	PGaE indicators	Core dynamic trend and its effect on PGaE indicators	Available policy options (i.e. PGaE programmes)	Selected PGaE
		Farmland abandonment	e.g. through payments to farmers to maintain current land use	
Landscape (cultural services)	- Very high recreation potential index - **Very high cultural heritage**	Decrease	Y	X
Biodiversity	- **Very high HNVF**	Decrease	Y	X
Water Quality	- Very low total N input			
Water Availability	- Medium-low infiltration - **Very high irrigated UAA**	Decrease		
Soil Quality	- **Very high risk of soil erosion***	Increase	Y	X
Air Quality	- Low total NH_3 emissions			
Climate Stability	- **Very low soil carbon content**	Increase		
Resilience to flooding	- Very-low flooding risk			
Resilience to fire	- **High fire risk**	Increase	Y	X


STEP 3: DESIGN CHOICE EXPERIMENT SURVEY
ONLY FOR ONE MR (Mediterranean Uplands)
MRAEP/Choice Scenario
Data-driven MRAEP methodology
Specify attributes and respective levels
Qualitative data
Focus Groups
Specify choice sets
Questionnaire pre-test
Elaborate questionnaire
Econometric analysis
Questionniare
Test questionnaire (Pilot survey)
Data for resident (PT) & Non-resident (DE)

- **Non-monetary attributes: selection and description –**
  Attributes for MRAEP "farmland abandonment in Mediterranean Upland"



Landscape Conservation		Erosion Control	
Farmers' commitment:	**Society's benefits:**	**Farmers' commitment:**	**Society's benefits:**
Maintain production of traditional crops Practice an environmental friendly agriculture	Safeguard the cultural heritage Enjoy high quality and tasty products Enjoy the traditional countryside for recreation and leisure	Keep terraces on steep sloped terrain Keep the soil covered with vegetation and avoid ploughing	Ensure soil fertility Ensure the soil's ability to support landscape and biodiversity
Biodiversity Conservation		**Fire risk reduction**	
Farmers' commitment:	**Society's benefits:**	**Farmers' commitment:**	**Society's benefits:**
Maintain the habitats for endangered fauna and flora Practice an environmental friendly agriculture	Preserve animal and plant species from extinction Enjoy nature for recreation and leisure	Bushes' removal Keep crops as barriers to the progression of fires	Ensuring the integrity of people and goods Avoid air pollution and emissions of greenhouse gases

Figure 5 – Programmes delivering the selected public goods

Non-monetary attributes levels

Attribute/PG Programme	%area benefited	%area benefited	%area benefited
PG Cultural landscape	0%	50%	100%
PG Farmland biodiversity	0%	50%	100%
PG Soil quality	0%	50%	100%
PG Fire resilience	0%	50%	100%

- *Applying each programme in 50/100% of the area of the region will ensure the **maintenance of the existing traditional landscape**, **preserve all currently endangered species** and **prevent the increase of the risk of erosion and fire risk compared to the current situation.***

Conclusion on Non-monetary attributes levels

- The SQ is a policy-off scenario => it is too drastic
- Very difficult to provide context-rich scenarios for EU survey
- It was very complex to determine results indicator as attributes for the whole EU => therefore the evaluation survey does not value the outcome of a policy, but which public good (overall) is more value by respondents.
- The intermediate (quality) levels were not appreciate by respondents (judged as unstable and transient by respondents)

Results should be interpret in relative terms and directional effects => determine the relative importance of each public good by respondents rather than determining the WTP for specific results/outcome in each public good.

- Monetary attribute –

  Payment vehicle, amounts and duration of payments

- This cost (programmes implementation costs) have to be support by the European citizens, including you, by higher taxes, or creating special rates on products or about visitors to this region,

- The amounts, duration of payments and unit that pays

Increase in tax payments per household (annually for 5 years)	$	3 €	12 €	21 €	39 €

- Choice set: number of alternatives (baseline and reference levels) – MRAEP "farmland abandonment in Mediterranean Upland"
  - Example of choice set (experimental design=efficient design, priors=0)

Programme providing services …	No application	Option A	Option B
 Landscape conservation	0 %	0 %	100 %
 Biodiversity conservation	0 %	100 %	0 %
 Soil erosion control	0 %	50 %	50 %
 Fire risk reduction	0 %	100 %	0 %
 Increase in taxes payments (annually for 5 years)	0 €	3 €	21 €

- Test at pilot level of the questionnaire for–Samples MRAEP "farmland abandonment in Mediterranean Upland"

Three samples for 300 valid interviews each have been selected

a) Face-to-face interviews with CAPI (computer assisted personal interviews), carry out in the Metropolitan area of Lisbon (PT)

b) Panel web-base (on-line) administrated for national population of Portugal

c) Panel web-base (on-line) administrated for national population of Germany

Criteria for sampling selection:

1. Stratified samples have been selected

a) Age, gender

b) & c) Age, Gender, NUTS2

2. Individual with 18 years old in charge of household expenses

- Estimates for the attributes WTP (based on the models MNL and RPL with socioeconomics), values are in € per 100% of the area in the MR.

PGaE	PT_F2F		PT_WEB		DE_WEB	
	MNL	RPL	MNL	RPL	MNL	RPL
Landscape (cultural)	28	37	30	37	38	39
Farmland Biodiversity	32	37	48	55	54	62
Erosion control	14	13	23	24	11	17
Fire risk reduction	32	37	37	51	14	23

!!! Results should be interpret in relative terms as relative importance of each public good

- ## Implement EU large-scale survey - definitions

1. To implement the CM strategy at EU scale, alternative sampling plans (budget) were developed.
2. The range of surveys 14.400 – 42.200 with a cost range of 108.000-2.911.800 Euros.
3. The budgetary cost depends on:

a) number of MRAEP to be valued in each MS (nr surveys). As well important to value non-resident population.

b) Samples size/sampling error of each survey

c) Survey administration mode (FTF=45-69 Euros/interview), Web-based (6-9 euros/interview)

Main conclusions replicability/generalizability

- The aim of this study is to have a replicable methodology in the whole EU aiming to evaluate demand PGaE.
- Replicability:
  - The larger the sample size => better replicability
  - Important to consider the same stratified sample
- Generalizability:
  - The study is designed to cover all EU & main ecosystem services
  - It cannot be generalize to other ecosystems/regions=> very context-dependent

Open discussion on how to improve replicability/generalizability of economic experiments?

- **Improve replicability => design and analysis** (Asendorpf et al., 2013)
  1. Increase sample size
  2. Increase reliability of measures: less measurement error
  3. Increase study design sensitivity: distinguish systematic/random error
  4. Increase adequacy of statistical analyses.
  5. Avoid multiple underpowered studies
  6. Consider error introduced by multiple testing
  7. ???
  8. ???

- **Improve replicability => Publication process**
  - As scientists we should design and document our methods to anticipate replication and make it easy to do.
  - It is important to provide information not only on significant results, but also on those that either have no effect or had unintended effect=> transparency.
  - As compulsory practice the scientific journals should require the posting of data and computer code to make easy the process of replication
  - …
  - …

Iniciatives to track replicability

Journal of Experimental Social Psychology 50 (2014) 217–224

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

The Replication Recipe: What makes for a convincing replication?

Mark J. Brandt [a,*,1], Hans IJzerman [a,1], Ap Dijksterhuis [b,2], Frank J. Farach [c,2], Jason Geller [d,2], Roger Giner-Sorolla [e,2], James A. Grange [f,2], Marco Perugini [g,2], Jeffrey R. Spies [h,2], Anna van 't V

[a] Tilburg University, Netherlands
[b] Radboud University Nijmegen, Netherlands
[c] University of Washington, USA
[d] Iowa State University, USA
[e] University of Kent, UK
[f] Keele University, UK
[g] University of Milano-Bicocca, Italy
[h] Center for Open Science, USA
[i] TIBER (Tilburg Institute of Behavioral Economics), Netherlands

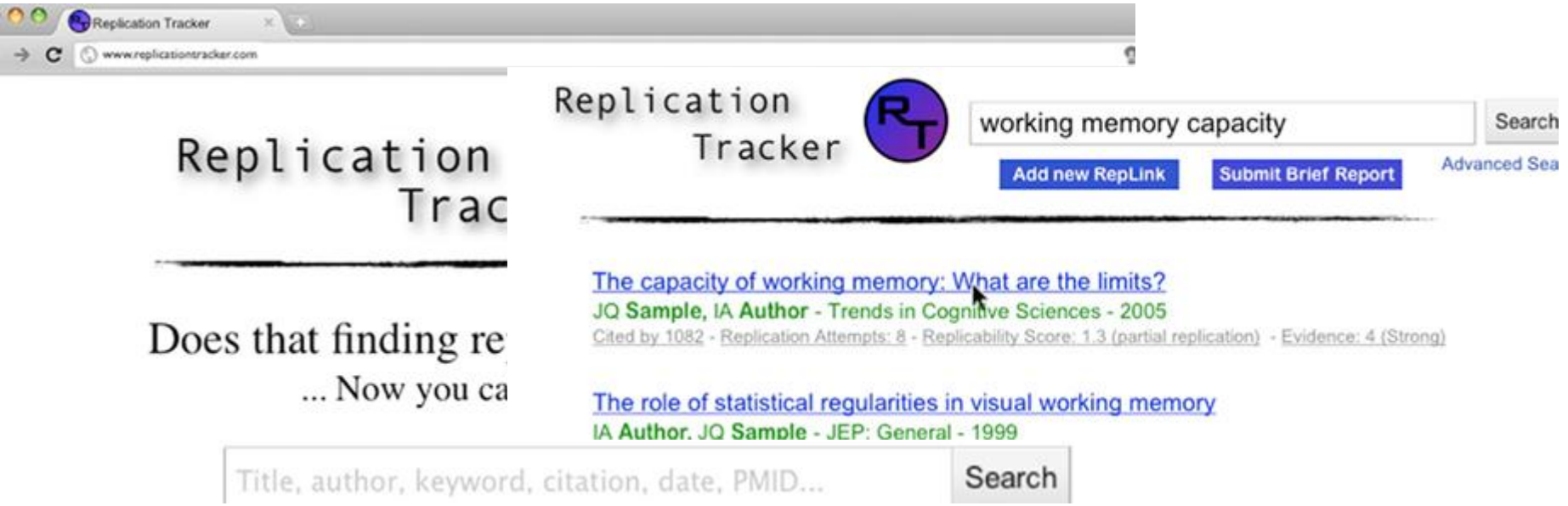
1. Carefully defining the effects and methods that the researcher intends to replicate;
2. Following as exactly as possible the methods of the original study (including participant recruitment, instructions, stimuli, measures, procedures, and analyses);
3. Having high statistical power;
4. Making complete details about the replication available, so that interested experts can fully evaluate the replication attempt (or attempt another replication themselves);
5. Evaluating replication results, and comparing them critically to the results of the original study.

- **Iniciatives to track replicability**

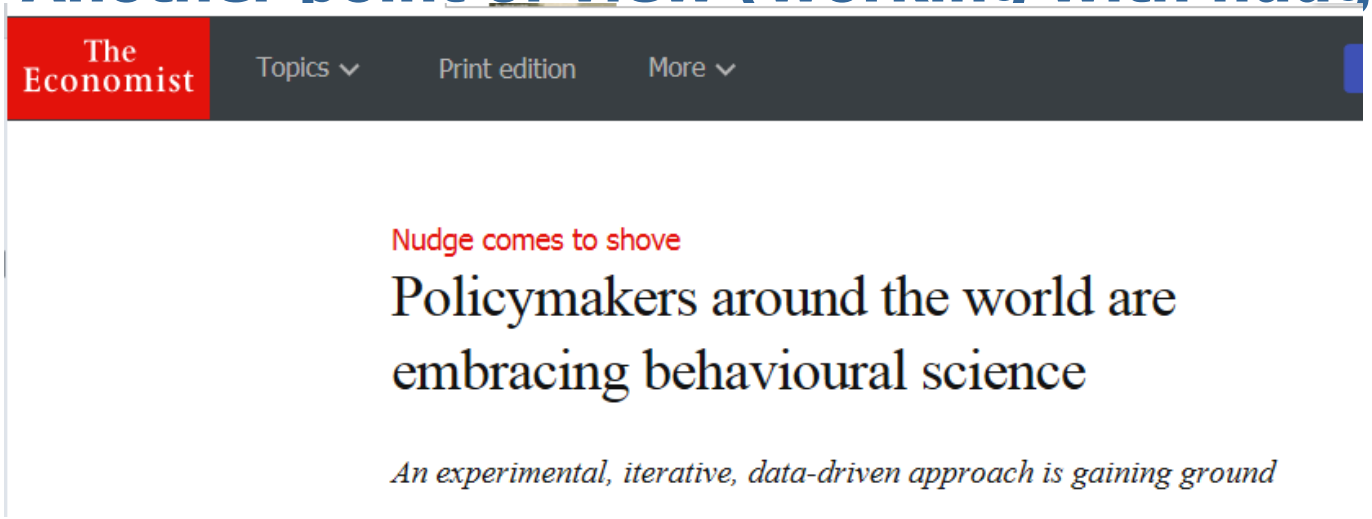
- http://journal.frontiersin.org/article/10.3389/fncom.2012.00008/full

- Proposal of tracking establishing an open-access journal dedicated to publishing replication attempts.

- We propose tracking replications as a means of post-publication evaluation, both to help researchers identify reliable findings and to incentivize the publication of reliable results.



Another point of view (working with nudges):



http://www.economist.com/news/international/21722163-experimental-iterative-data-driven-approach-gaining-ground-policymakers-around

A "replication crisis", in which scientists in many fields have repeated published experiments and failed to find the same results, has hit particularly hard in the behavioural sciences, with some much-cited findings now open to question. But the approach taken by nudge units and their kind already incorporates the remedy. It has nudged policymakers towards a new way of thinking about policy that involves trial and error, and step-by-step improvement. The theories of behavioural science can only suggest which nudges to try; it is for policymakers to find out which ones work.

For discussion: Conclusions on generalizability of results from this study 1

- Potential factors affecting generalizability:
  - **Type of subject.** In this study: impact of subject pool on experimental outcomes, in the magnitude of the results.
  - Probably context/game/study-dependent: every scenario tested had an equilibrium that maximized players' payoffs. A different game structure may reveal other impacts.
  - **The experimental framing:** can affect internalized norms of participants and can trigger signals that do (not) matter to the decision-making process of a particular subject.
  - Distribution of players' characteristics and social preferences across subject pools. Some of those influence players' decision significantly, and their effect vary across framings and subject pools.

For discussion: Conclusions on generalizability of results from this study 2

- Potential aspects improving the generalizability of economic experiments:
  - To conduct the experiment with the population that is most concerned by the issue: unique insights.
  - To include the important components of the management issue. Trade-off between capturing the essence of the problem (no over-simplifcation of the game structure) and being able to disentangle the effects.
  - To control for individuals' characteristics among participants and among the ideal population.

- Is strict replication feasible considering that any study is performed in a specific historic context that is always changing?
- If not even replicability can be shown, generalizability is impossible as the finding is so specific to one particular circumstance as to be of no practical use