V ANNUAL CAMPBELL COLLABORATION COLLOQUIUM

Methodological Advances in Meta-analysis Simposium

Towards the Validation of a Scale to Measure the Quality of Primary Studies for Meta-analysis

Salvador Chacón-Moscoso
Susana Sanduvete-Chaves
David Alarcón-Rubio
University of Sevilla, Spain

Introduction

- Systematic reviews pretend results generalization from a group of different studies about some area of interest.

- This procedure has to develop criteria to choose and codify those studies.

- In order to avoid biased or erroneous conclusions, one of the main problems in systematic reviews is to develop clear criteria to combine studies with different degrees of quality.

- We can consider that there is a certain degree of correspondence between used criteria to choose and codify studies in systematic reviews and those design components that are relevant to enhance quality in particular interventions and in their results generalization.

- Design components quality are relevant to increase not only the intervention quality, but also to foster quality in their evaluations and in systematic reviews based on those evaluation results.

General Objective:

- Review and systematize published contents about quality in primary studies.

Specific Objectives:

- Describe main different ways to assess quality systematically. Advantages and disadvantages.
- Review literature about studies quality and present an exploratory system of categories containing some of the most frequent items used to assess quality.
- Describe how published papers about program interventions in USA and Europe (from European Union Countries) present information of some of those previous obtained quality categories.
- Randomized versus not randomized studies.
- Practical questions to take into account when deciding which studies to consider to perform a systematic review.
- Main point to consider for practitioners (social workers, psychologists, social educators,…) in order to increase quality in their interventions.
- Some future developments.

- Study quality is a complex and multidimensional concept that can be defined from different perspectives:
  - **Internal validity**
  - **External validity**
  - **Precision of study report**
  - **Appropriate statistical analysis**
  - **Ethical implications**
  - **Relevance for the intervention area**

Basically there are two perspectives to measure quality systematically:

- **study an individual component of quality as indicator of quality (Type of unit assignment, Attrition, Sample size,..)**

- **Obtain a global quantitative index of quality based on the scores from a selection of weighted quality items applied to each study.**

***Advantages of studying* an individual component of quality as indicator of quality (Type of unit assignment, Attrition, Sample size,..)**

- **Give a direct empirical evidence of one aspect of quality of the study**

For example. Efficacy of young delinquency rehabilitation programs in Europe

Subjects assignment	Number of studies	Medium effect size
Random	8	0,237
Non Random	22	0,451

Redondo, Sánchez-Meca y Garrido (1999)

- **Can be easily applied to any context**

Disadvantages

- **Present just one aspect of quality of the study, but not a global assessment**

***Advantages of* obtaining a global quantitative index of quality.**

- **Introduces a global assessment of quality of the study.**
- **If developed with metric properties it can present adequate indexes of validity and reliability Principal checklists of quality**
- **CONSORT: Consolidated Standards of Reporting Randomized Trials**
  - **First published in 1996 and revised in 2001 (JAMA 1996;276:637-639)**
  - **Checklist of 22 items that should be included in the trial report**
  - **It focus on randomised controlled trials (RCT), analyzing different types of random assignment.**
- **STARD: Standards for Reporting of Diagnostic Accuracy**
  - **First official version published in 2003 (Clin Chem 2003;49:7-18)**
  - **Checklist of 25 items that should be included in the report of a study of diagnostic accuracy**
  - **The objective is to improve the quality of reporting of studies of diagnostic accuracy**
- **TREND: Transparent Reporting of Evaluations with Nonrandomized Designs**
  - **The initial version published in 2004 (Am J. Public Health, 2004;94:361-366)**
  - **Checklist of 22 items relevant for the report of nonrandomized trials**
  - **It is proposed for intervention-evaluation studies using nonrandomized designs.**
- **STRICTA: Standards for Reporting Interventions in Controlled Trials of Acupuncture.**
  - **First STRICTA recommendations published in 2002 (Acupuncture in Med 2002;20(1):22-25)**
  - **Checklist of 6 items that should be included in the report of acupuncture interventions**
  - **The intended outcome is that interventions in RCT of acupuncture will be more adequately reported.**

CONSORT	**TREND**	**STRICTA (Acupuncture)**	**STARD (Diagnostic)**
TITLE AND ABSTRACT	TITLE AND ABSTRACT		TITLE, ABSTRACT, **keywords**
Allocation participants	Allocation participants		
			Identify tipe of diagnistc
	Structured abstract recommended		
	Information on target population		
	Scientifi background		
INTRODUCTION	INTRODUCTION		INTRODUCTION
Research cuestions, study aims			
	Background: theory	JUSTIFICATION OF INTERVENTION	
		NEEDS	
METHODS	METHODS		METHODS
* Participants	* Participants		*Participants
Elegibility criteria for participants	Elegibility criteria for participants		Elegibility criteria for participants
Setting	Setting		Setting
Location where data were collected	Location where data were collected		Location where data were collected
			Describe type of poblation: exclusión criteria
	Method of recruiment		
*Intervention	*Intervention	*INTERVENTION	
		Number of sessions	
		Frecuency of treatments	
Details of different groups (how and when)	Details of different groups (what was given, how, who, where, when, how long, incentives)		Details of different groups (how and when)
		OTHER INTERVENTIONS (CO-INTERVENTIONS)	
		IMPLEMENTERS TRAINING	
		Duration	
		Specific conditions experience	
*Aims	*Aims		
Specific	Specific		
*Outcomes	*Outcomes		
Clearly defined measures	Clearly defined measures		
Quaility measures	Quaility measures: validity		
	Methods used to collect data		
*Sample size	*Sample size		
How was determined	How was determined		
***Randomisation**	*** Assignment method**		
Method	Method		
Method to implement			
	Unit (individual, group, etc.)		
	How minimizate bias due to nonrandomization		

CONSORT	**TREND**	**STRICTA (Acupuncture)**	**STARD (Diagnostic)**
*Blind	*Blind	Blind **of participants**	
	Description		
	If it`s different from de unit of assignment, the analytical method used to account for this		
*Statistical methods	*Statistical methods		
To compare groups, subgroups	To compare groups, subgroups		
	For imputing missing data		
	Software		
RESULTS	RESULTS		RESULTS
*Participant flow	*Participant flow		*Participants
			Demographic characteristics of the stu population
			Atrittion
			* Test results
			Time of intervention
			Model of reference (clinical)
	Results including negative and missing data		Results including missing data
			*Estimates
			Confidence intervals, etc.
			What was doing with missing data
			Variability between subgroups
			Reproductibility, if done
Across time	Across time, allocation, enrollement, assignment, follow-up		
Period of follow-up assigned	Period of follow-up assigned		
Baseline	Baseline (*) each group		
Number of participants each group	Number of participants each group		
Effect size each group	Effect size each group		
Other data apart aims	Other data apart aims		
Adverse events	Adverse events		Adverse events
	Enrollment: Number of participants screened		
	Analysis: number of participants analysed		
DISCUSSION	DISCUSSION		DISCUSSION
Interpretation of results and bias	Interpretation of results and bias		
	Alternative explanations		
	Success		
	Policy implications		
Generalisability (external validity)	Generalizability (external validity)		Applicability
General interpretation	General interpretation		

***Disadvantages of* obtaining a global quantitative index of quality.**

- **There are more than 200 scales of quality. But it is not clear how to measure study quality about primary studies reliably and realistically (different scales can give different quality scores to the same studies).**

- **There are different ways to understand quality (internal validity, external validity, relevance…). There is a wide array of methodological variables related to quality but probably do not assess the same kind of 'quality'.**

- **Feasibility to apply different scales to different contexts.**

- **Different items or different weighted items for a final quantitative score.**

- **Metric weaknesses that implies low validity and reliability indexes in developed scales**

Literature review about quality design.

Sampled papers

Reviewed documents in order to obtain an approach to codify quality design are the following:

- Begg, et.al. (1996); Brown (1991); Emerson et.al. (1990); Greenland (1994); Jüni (1999); McGuire, et al. (1985); Moher (1996);Moher et. al. (1995); Moher (1992); Moher et. al. (1998); Moher et. al. (2001); O´Rourke et. al. (1989); Sánchez, J. & Ato, M. (1989); Tritchler (1999); Weisz et. al. (2000) and Yeaton et. al. (1995)

An exploratory system to codify design quality:

1 Publication year

2 Type of publication.
   1. Journal
   2. Book
   3. Thesis
   4. Congress
   5. Other ones

3 Theoretical orientation
   1. Specified
   2. Inferred
   3 There is no data enough

4 Intervention Field
   1. Sanitary
   2. Educational
   3. Social
   4. Clinical
   5. Organizational
   6. Others

5 Age (Range) referred: Y/N

6 Age (mean)
   Age standard deviation

7 Implementation context:
   1. Urban
   2. Rural
   3. Mixed

System of coding.

8 Units random assignment:

1. None and without control of extraneous variables
2. None but with control of extraneous variables.
3. Yes

9 Methodology or Design

1. Experimental ; randomized
2. Quasiexperimental (two groups without randomized assignment) non-equivalent control groups with pretest and posttest

3 Pre-Experimental (only one group + one measure) / others (questionnaires/observational/naturalistic) .

10 Sample size

1. n <5
2. 5 <n <10
3. n >10

11 Attrition:

1. >30%
2. <30%
3. Without mortality

12 Follow-up period :

1. < 6 months
2. 6-11 months
3. > 12 months

System of Coding.

13 Moments of measurement

1 Post intervention

2. Pre and post intervention

14 Measures in pretest appear in posttest

1. No
2. Some

3 All of them

15 Normalized dependent variables

1. Without (self-reports and post hoc records)
2. Questionnaires or standardized self-reports
3. At least one is objective (psychophysiological measures)

16 Intervention/Study homogeneity

1. Subjects do not receive the treatment in the same contextual conditions
2. Subjects receive treatment in the same contextual conditions

17 Control Techniques

1. Blind (beneficiaries)
2. Blind (implementers)
3. Both
4. Other ones

18 Effect Size and value

19 Level of difficulty to Codify

1. Low
2. Medium
3. High

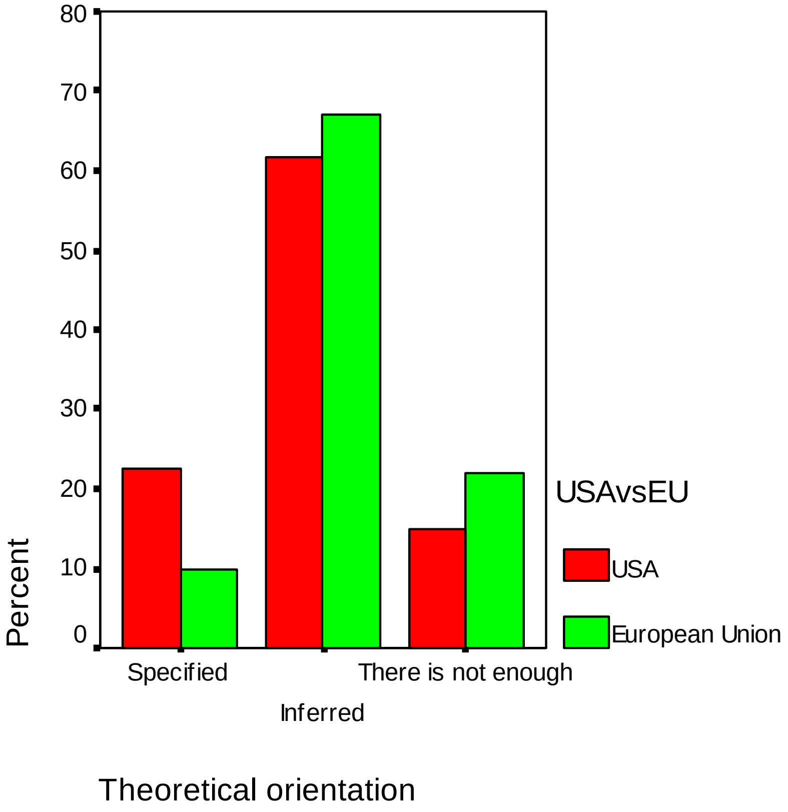
An application of proposed design quality codes to published papers about interventions programs in USA and Europe.

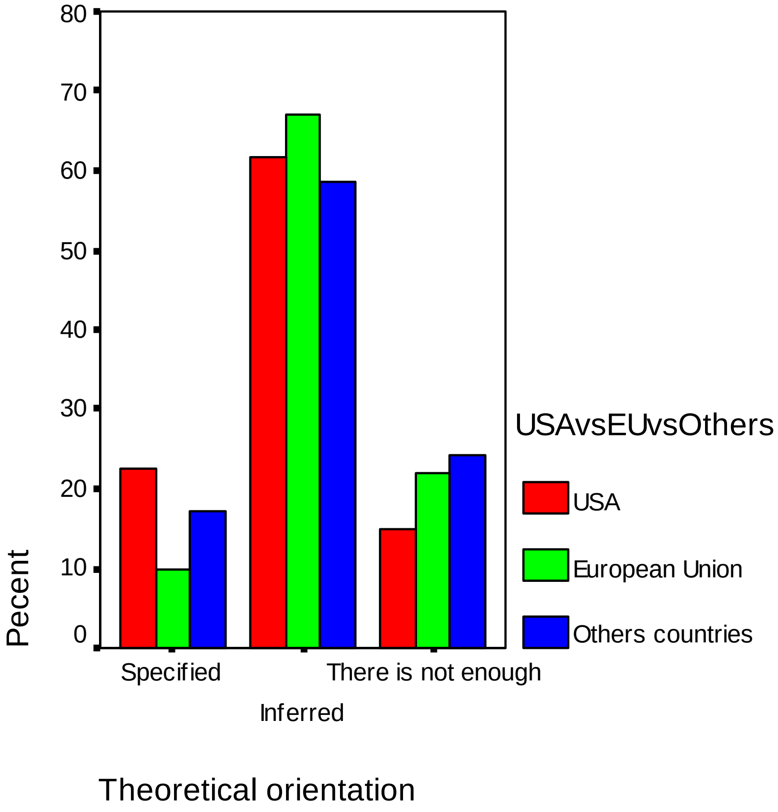
- Procedure:
  - Psycinfo (1887-2004); Eric (1966-2004); Current Contents (1999-2004); and EBSCO Online (1997-2004) databases were used to obtain published interventions.
  - Keywords used to select papers (alone and using all possible combinations):
    - Random; Non-random; Effect size; Quasi-experimental; Experimental; Meta-analysis; Intervention Program; Evaluation; Social; Education; Assessment.

- Sample:
  - 776 papers were used for codification (data availability, human intervention, non-replication within same studies). 194 of those articles weren’t codified because those didn’t describe data enough.
- Instruments:
  - Online-databases available in University of Seville
  - Procite-5 for management database.
  - Spss 11.0 to codify and analyze data.

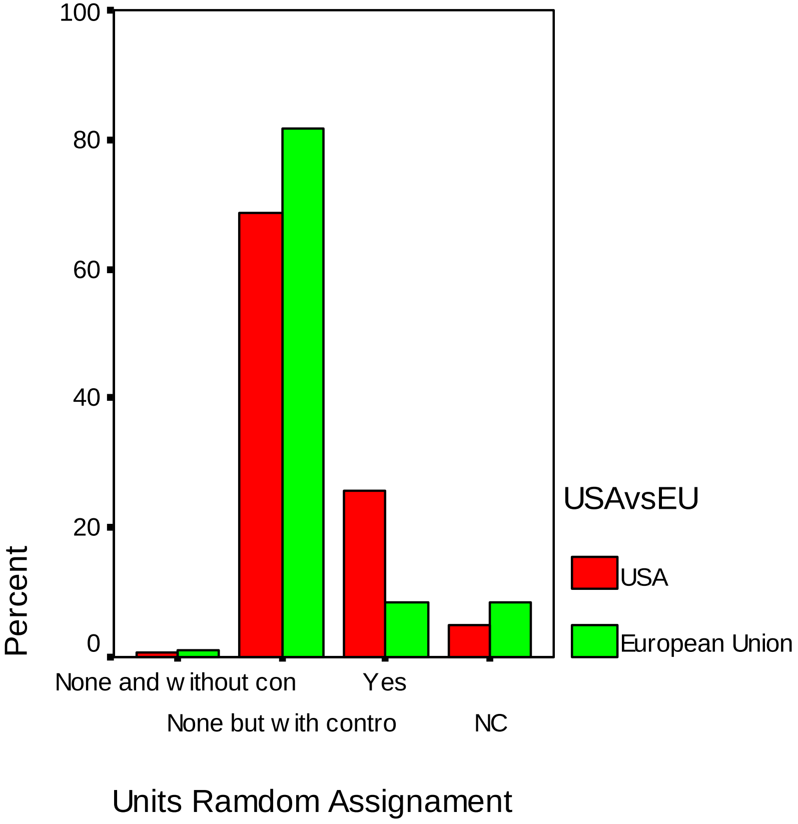
how published papers about program interventions in Europe and United States present information of some of those previous obtained exploratory quality categories.

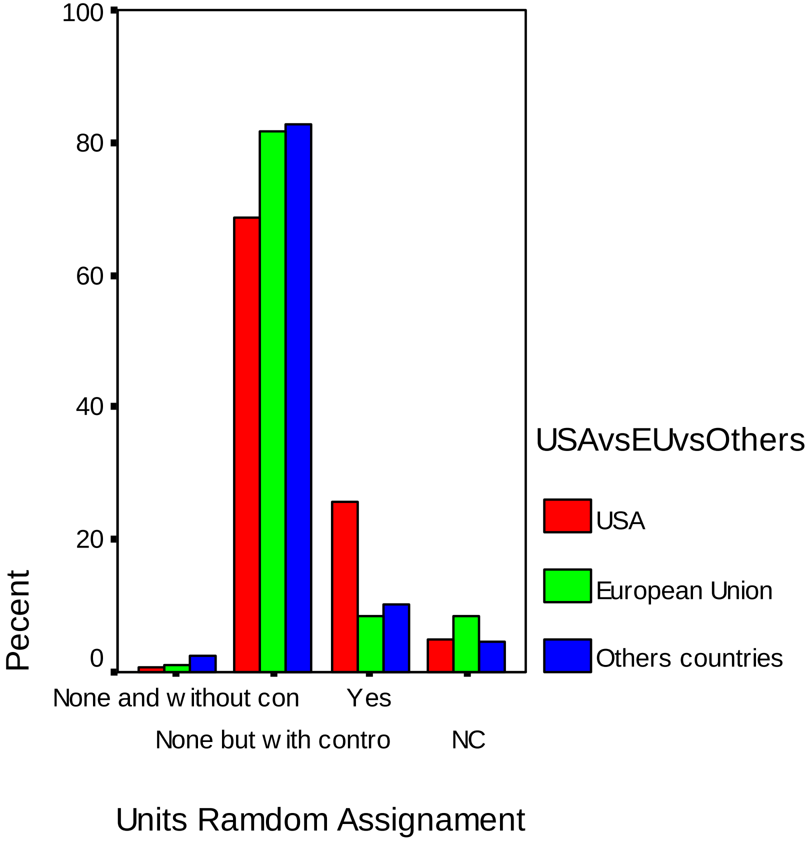
Theoretical orientation. In most cases it can be inferred from initial hypothesis (60%). Nonetheless there is not data enough about theoretical frameworks in an important amount of published researches (23%), Only less than 15% specifies the theoretical orientation clearly. This tendency is similar in USA, Europe and in the rest of the studied continents. Nonetheless Theoretical orientation specification is more frequent in USA (20%) than in EU (9%).



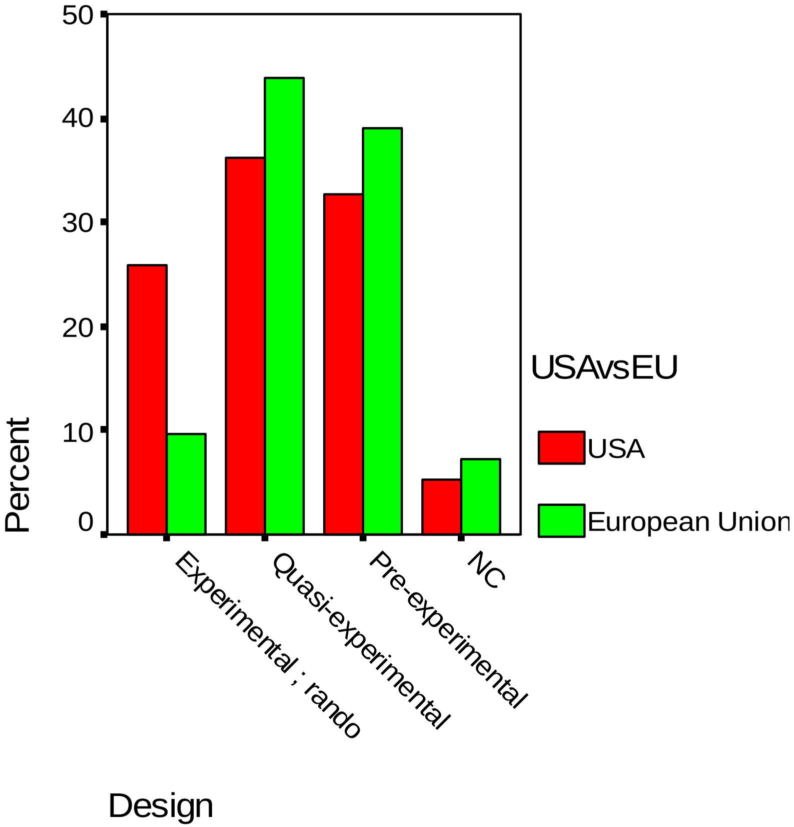


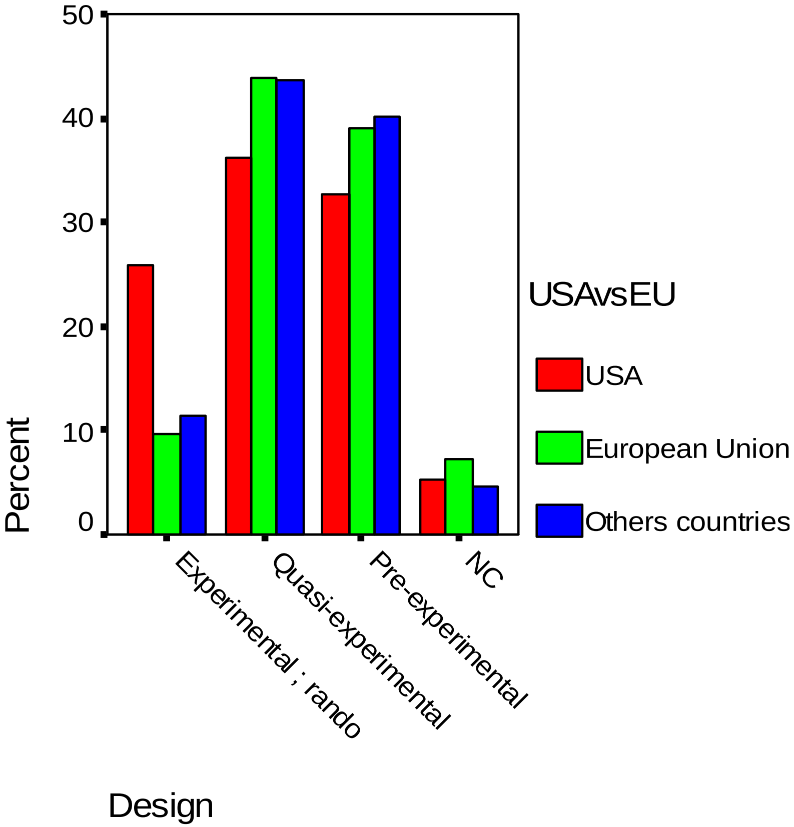
Units random assignment The random assignment was only used in a small percentage of researches (20%). The rest of them (75%) use another kind of control and 5% doesn't use none. It is important to note that random assignment of units is much more frequent in USA (25%) than in Europe and in the rest of other codified countries.



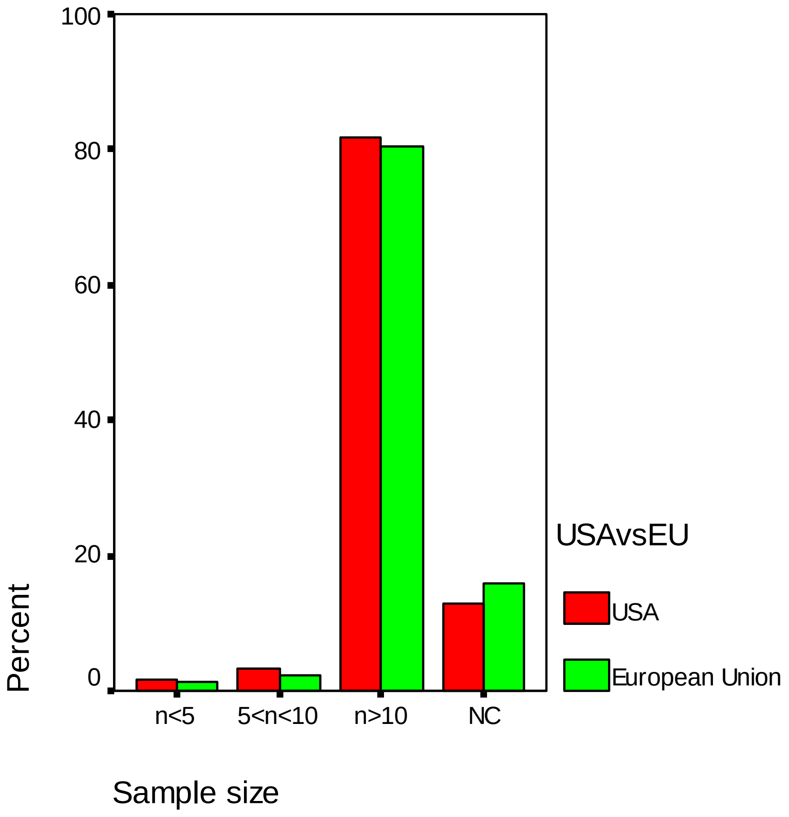


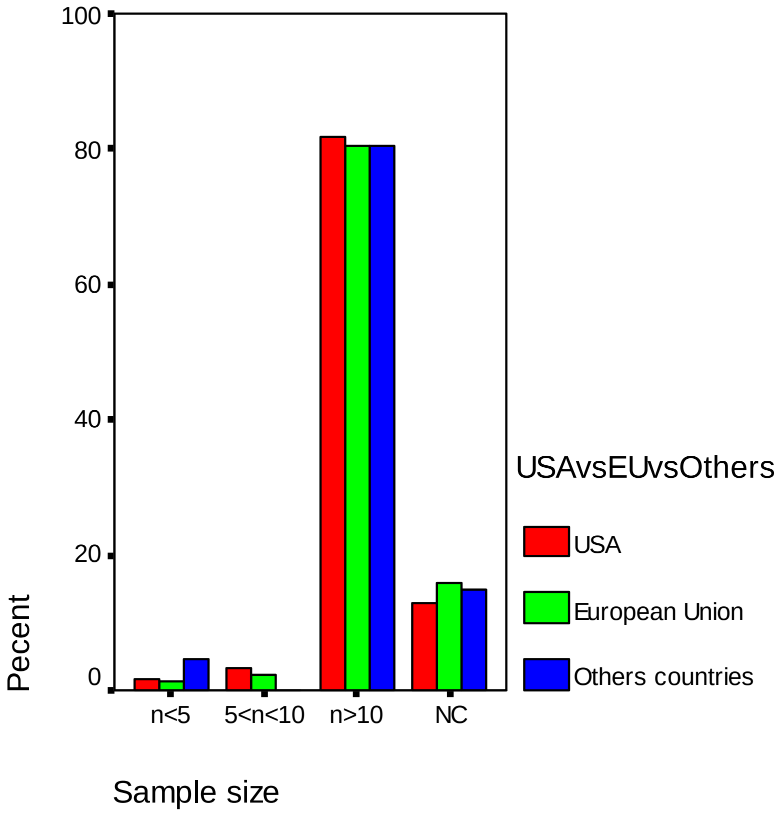
Methodology or Design. Most of programs have a quasi-experimental design (42%), although there are a lot of Pre-experimental design (35%) and less present experimental designs (10%), in this last case they are more frequent in USA (25%).



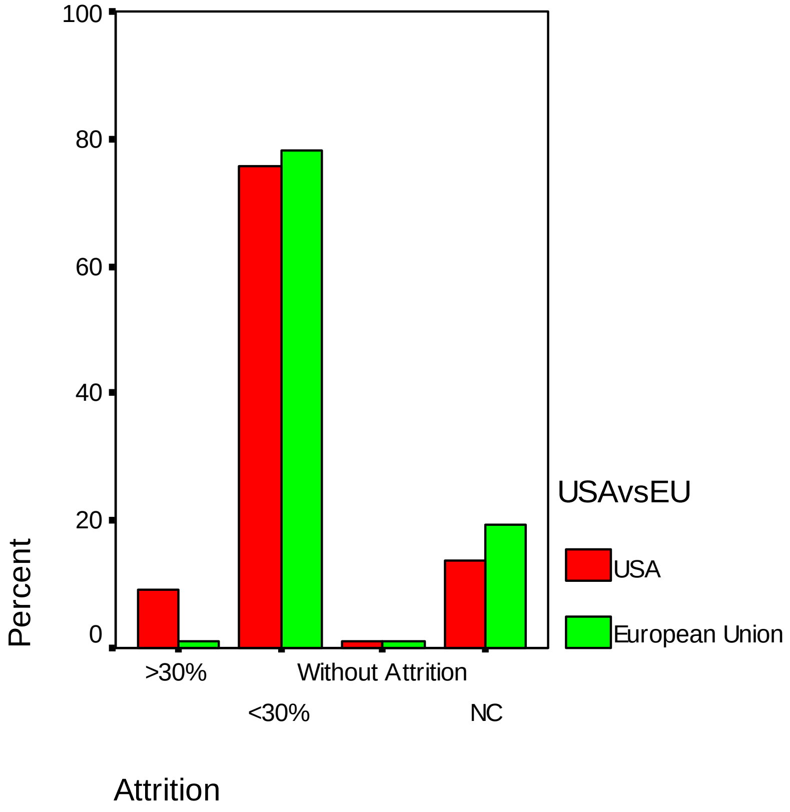


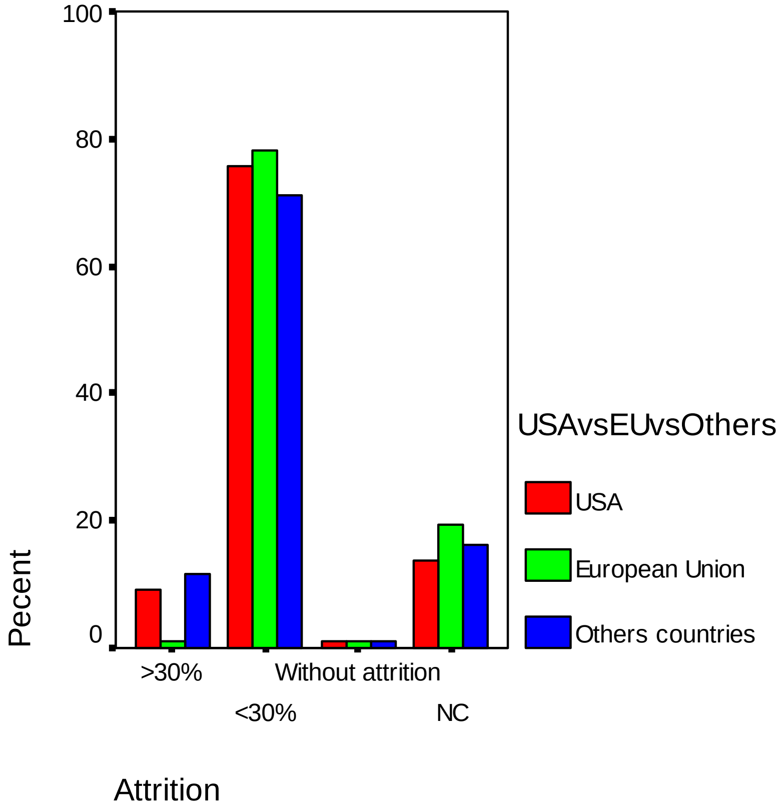
Sample size. Most programs present a sample size bigger than 10 subjects in Europe, USA and in other studied countries.



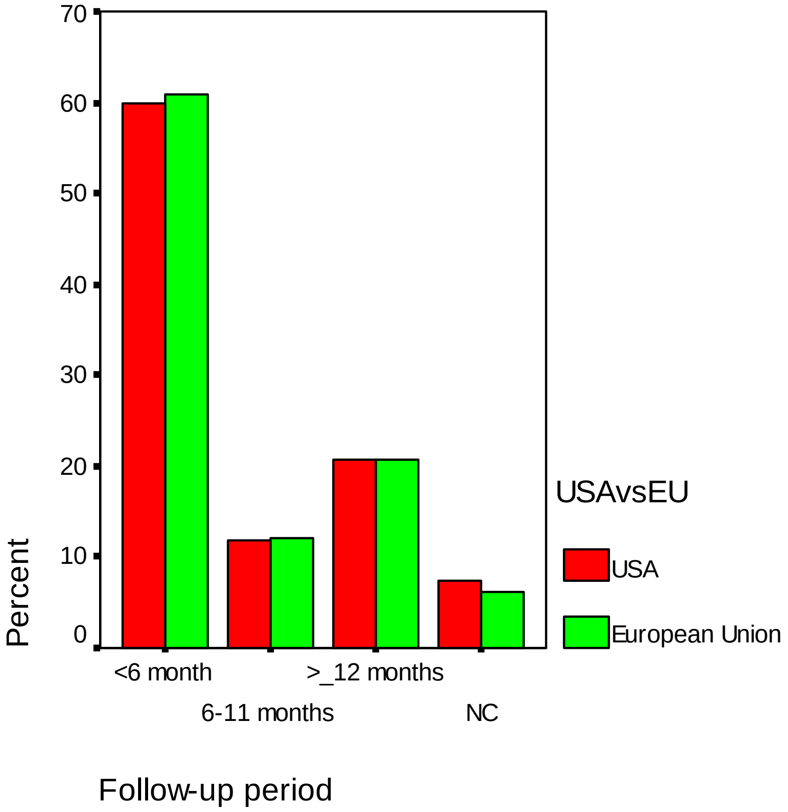


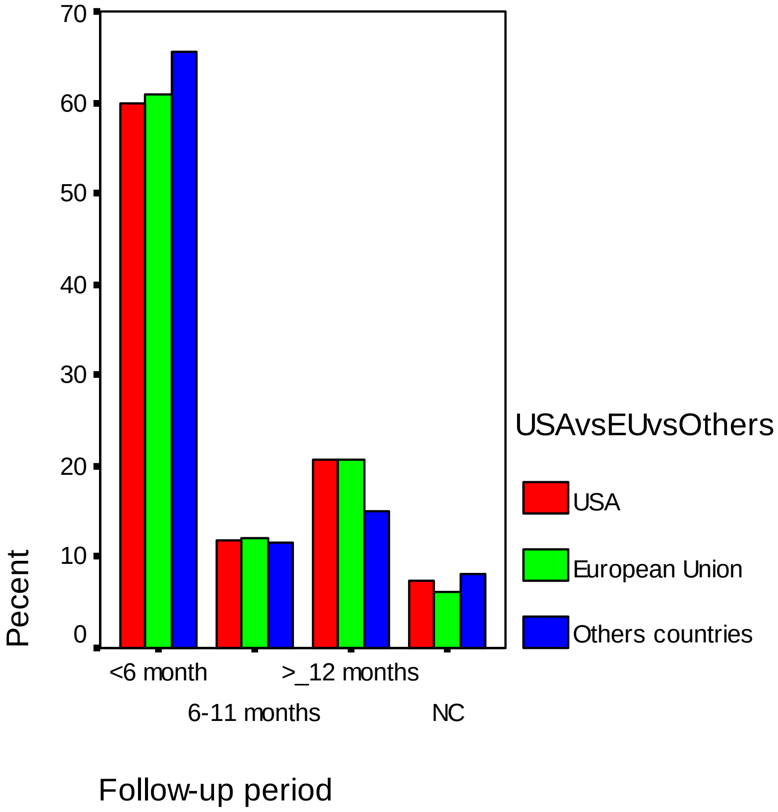
Attrition: In most cases USA, Europe and other studied countries (79%) attrition is smaller than 30%



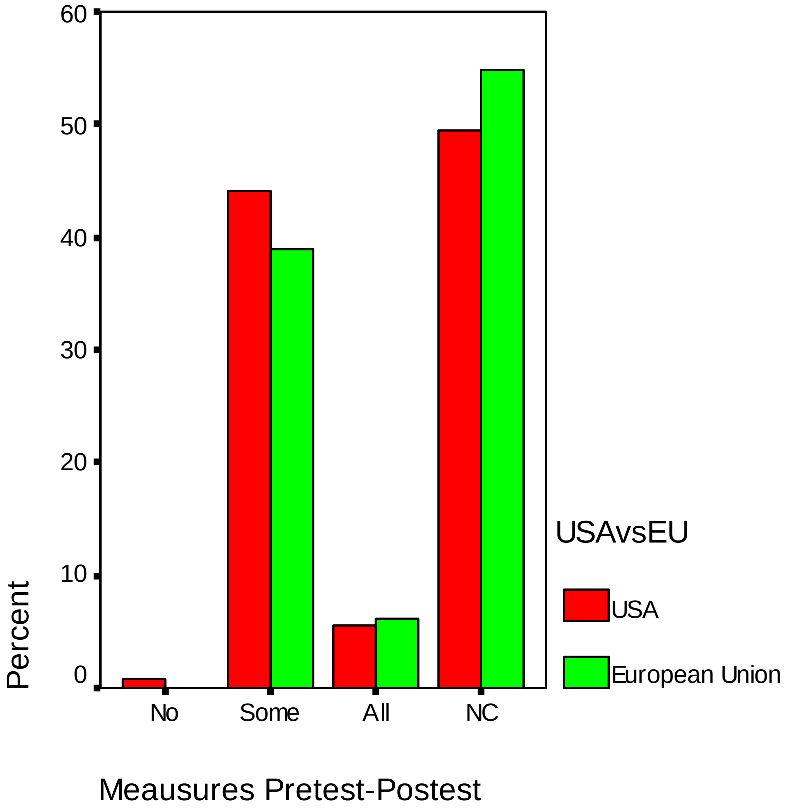


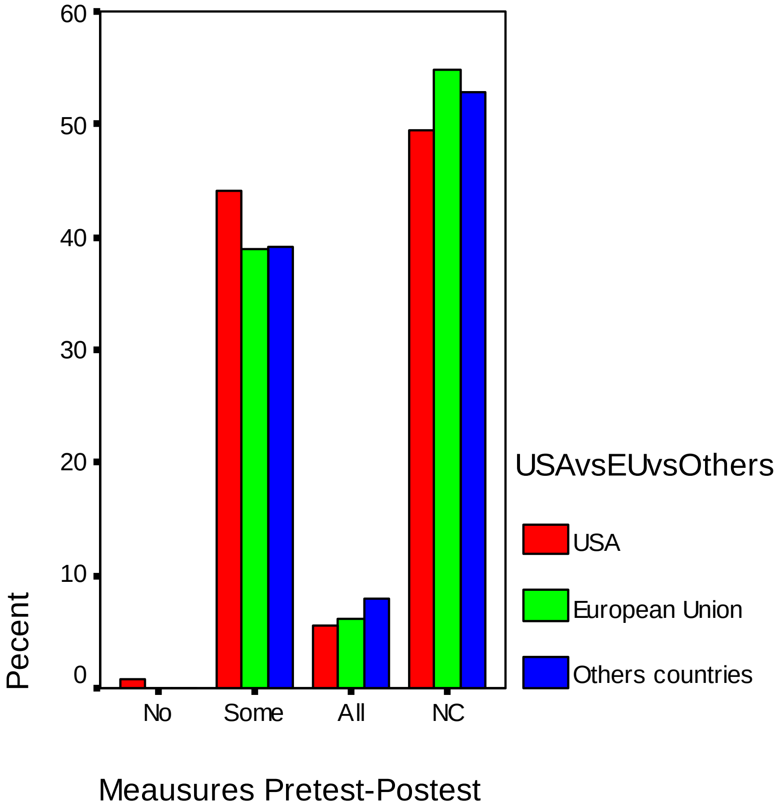
Follow-up. Most studies has done a follow-up period during six months (60%). Only 20% made a year post the intervention measurement. In this case USA and Europe present similar follow-up periods.



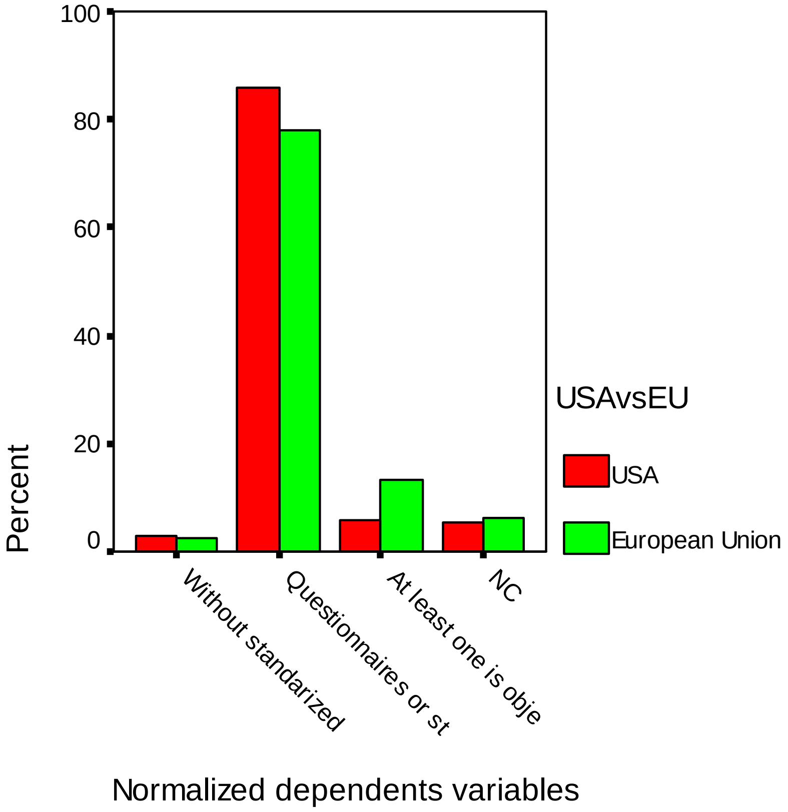


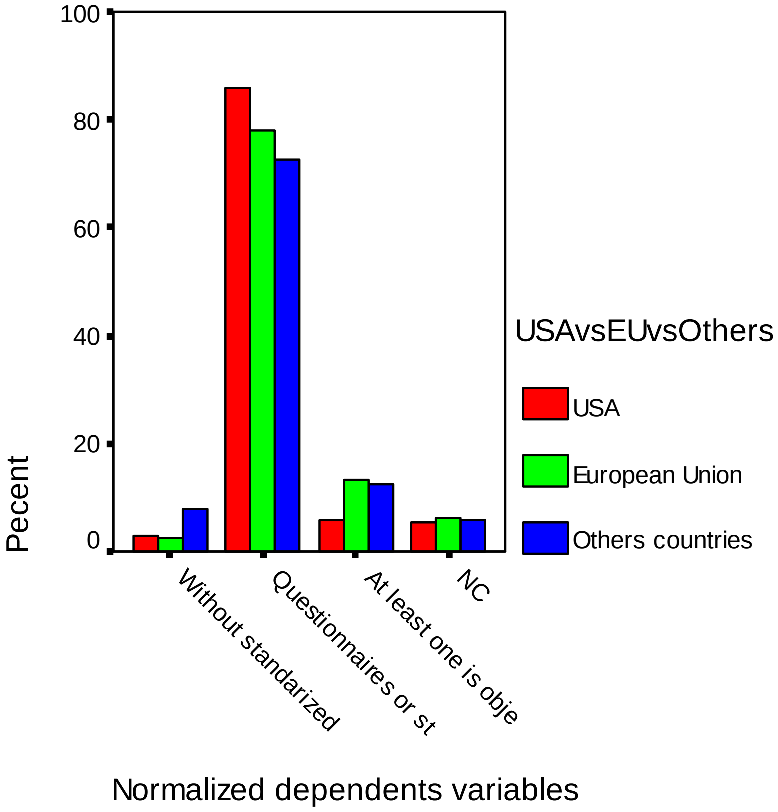
Time of measurement. Only 5% of programs present all the same measures in Pretest and posttest. This tendency is similar in USA, Europe and in the rest of the studied continents. But at least, around 40% present some of them.



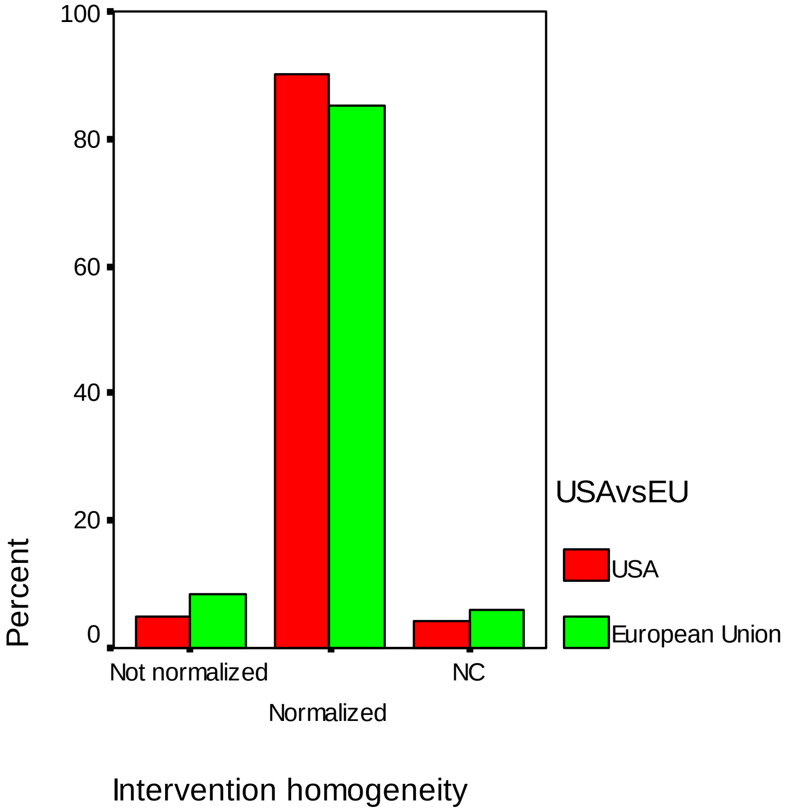


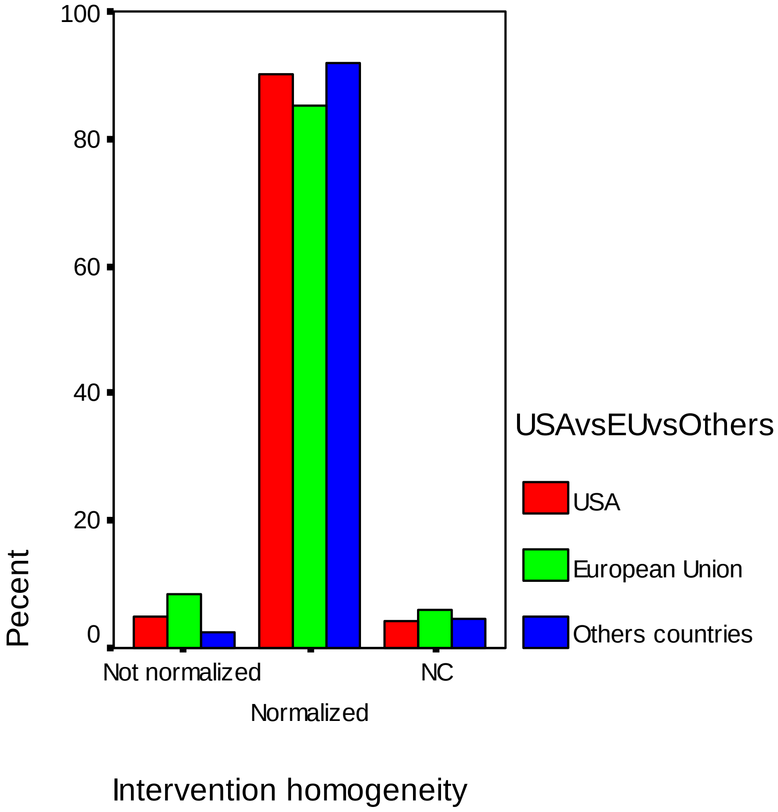
Standardized dependent variables. A high percent of programs used questionnaries or standardized self-reports measures (80%), followed by programs using at least one objective measure (7%), only a few use post hoc instruments (3%).



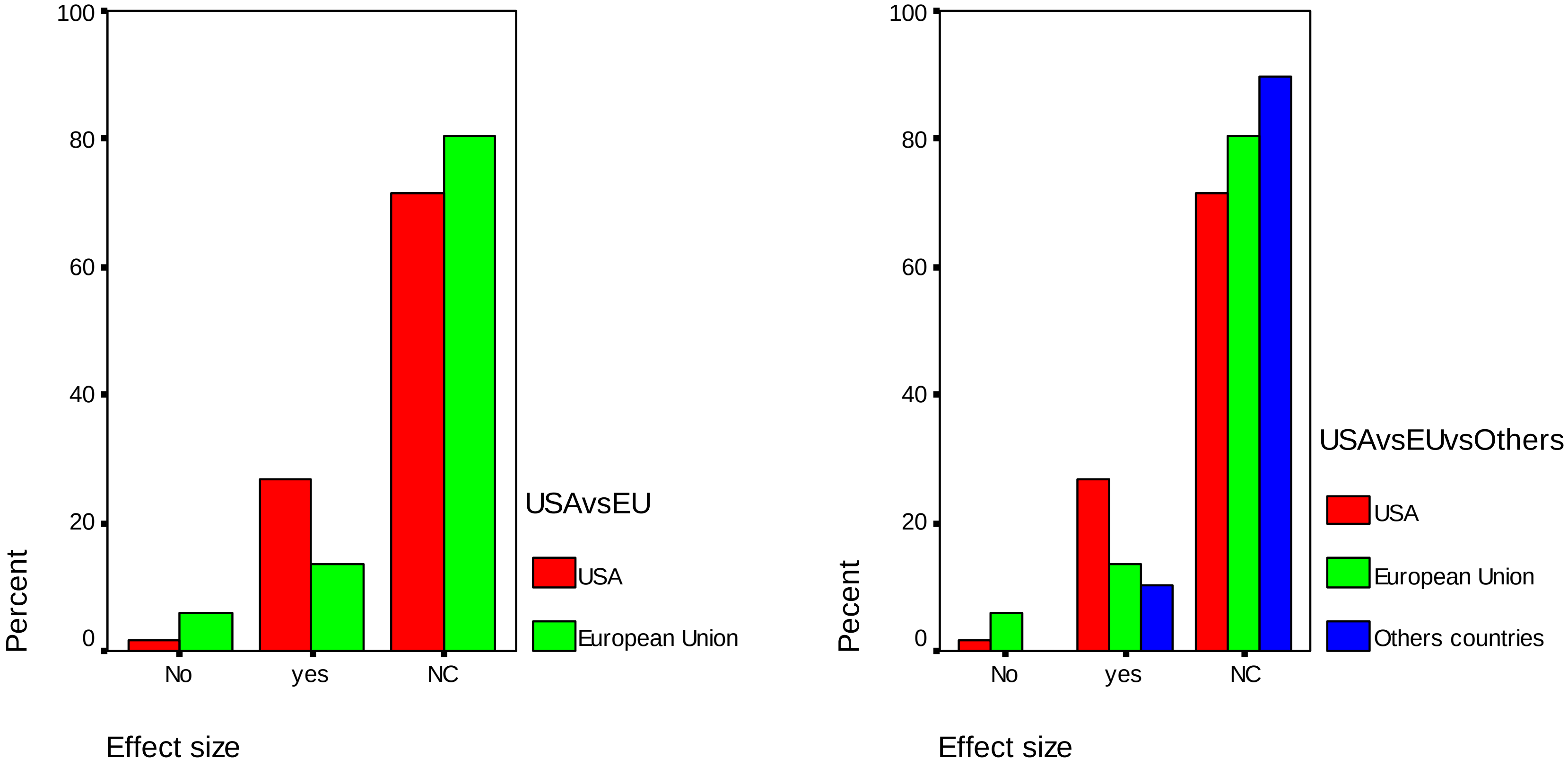


Intervention homogeneity. 85% of revised studies have been done in homogeneous contexts for the sample. This tendency is similar in USA, Europe and in the rest of the studied continents.





Effect Size. It is rarely specified in codified abstracts.



Some of the main characteristics of the research designs used in published papers about randomized and non-randomized interventions in Europe.

- Relatively few experiments exist from Europe compared to the many that have been conducted in the United States.
- Without the data to calculation effect sizes, a study cannot be included in a meta-analysis.
- Most studies used non-experimental or quasi-experimental methods would often result in the elimination of such studies from systematic reviews that use strict inclusion criteria.
- A real paucity of randomized experiments exists in the European context.
- For example, consider two meta-analyses about the same substantive area, the effectiveness of delinquency rehabilitation programs, one of them carried out in Europe (Redondo et al., 2002) and the another one in the United States (Lipsey, 1992). In the Lipsey (1992) meta-analysis, half of the studies were randomized; in the European meta-analysis, only 8.7 percent were randomized.
- In Europe there is a higher degree of centralization and locally implemented intervention programs. This can add to difficulties in carrying out meta-analysis if it results in a restriction of range in the sheer number of studies done, or if it results in the relative exclusion of some kinds of designs compared to others.

Randomized versus not randomized studies.

Random assignment allows unbiased estimates of treatment effects and justifies the theory that leads to tests of significance.

This reasoning justifies a possible hierarchical order of quality design/methodologies (mainly based on the Knowledge of unit assignment criteria , or procedures to avoid error term correlation with parameters to estimate; and because these designs usually do better than others to avoid different kind of biases). An example of a possible hierarchy:

- Randomized controlled trials.
- ‘Natural’ experiments
- Quasi-Experiments (Regression Discontinuity Design, Interrupted time series),
- Matching methods (Propensitive scores)
- Non-experimental data analysis
- Non-equivalent control group designs
- Pre-experimental designs (one group pre-post test)

But it is not such an easy question as randomized assignment must be properly executed and certain assumptions have to be met (e.g., no treatment correlated attrition). And also nonrandomized experiments can approximate results from randomized experiments when for example matching on reliable covariates.

If we have different kinds of design we should separate estimates of intervention effects for these different designs. Further, study separate effect sizes when multiple distinguishable classes exit (for example: randomized designs with or without high level of attrition; or nonrandomized design with nonequivalent control group design or case control design)

It is interesting to study design features separately more than general methodologies. Design differences can be tested and then use regression analysis to control for or take into account methodological features when testing substantive moderators of effect; for example:

Assignment methods
Attrition (total or differential)
Selection (self vs. others)
Pretest differences
Groups conformation
Statistical issues (for example: Cut-point to convert continuous into dichotomous variables)

The objective is to understand both the methodological and substantive factors that may contribute to study results and whether they act similarly or differently across designs

Today it is not possible to list all contingencies. There is the necessity that reviewers explore different possibilities of design coding categories to study how methodological and substantive factors may influences the results of systematic reviews

Following Shadish and Myer (2001), at least some design elements should be taken into account when performing systematic reviews:

In systematic reviews and quantitative syntheses:

-Kind of design

-sample size of treatment group for this effect size

-sample size of treatment comparison for this effect size

When combining randomized and nonrandomized also:

-randomization to the comparison made in the effect size

- Other codes

The objective is to explore reasons that might explain discrepancies in effect sizes among different kinds of designs.

Practical general process proposal to follow when performing a systematic review.

- Register all available studies in the area of interest
- Make a first global classification based on general categories depending on which quality concept has been defined (possible general categories; for example: kind of design, sample size)
- Study possible subcategories (for example: attrition, pretest differences).
- Analyze those groups of studies separately
- Study possible joint analysis and conclusions.

- Following this process we'll have feedback about how methodological and substantive factors may influences the results of systematic reviews **in representative contexts.**

- **We'll increase the knowledge about how different kind of biases are present in different contexts**

- I would never say that I cannot contribute with any information on the effect of an intervention as, in the 'worst' case, an important contribution is to study empirically that there is no 'valid' / reasonable evidence about the efficacy of an intervention. That is an important point to go on and from where to.

- It is very important to have raw data available (at least those data obtained from public funds) and accurate and exhaustive descriptions in study reports

How to improve practice in social intervention programs. Main key points (1)

- **Delimitate theoretical models and previous studies that justifies the intervention program designs (how to delimit an “intervention” successfully).**
- **Assignment procedure of units (subjects) to conditions (causal effects):**
  - **Should be clearly specified (randomly if possible)**
  - **Use similar comparison groups (using matching of units before assignment or cohort groups.**
- **Pretest observations (observations previous to program implementation)**
  - **Enhance using multiple pretest observations (as many as possible, always within boundaries of obtaining valid data) & trying to use high quality measures (for example physiological and standardized ones).**
  - **We must use at least one pretest observation (to test effects of interventions).**
  - **We can use alternative to pretest observations (pretest of independent samples, retrospective measures, proxy pretest of outcomes)**

How to improve practice in social intervention programs. Main key points (2)

- **Post-test observations:**
  - **We will always have a posttest observation, but we should add multiple posttest observations, equal or similar to pretest ones, whenever possible.**
  - **Enhance normalized post-test observations.**
  - **We can combine post-test observations with non-equivalent dependent variables.**
- **Comparison groups.**
  - **More extensive information about sampling features (selection, error, bias, attrition,..) should be detailed.**
  - **Randomly conformed groups should be enhanced; Nevertheless, it is better to use cohort groups or matching than non-equivalent comparison groups.**
  - **Multiple comparison groups should be used.**
  - **In extreme cases we can obtain comparison groups from regression extrapolation, or by using secondary data to make comparisons.**

Some future developments

- **Empirical study of threats to validity.**

  - Develop simulation studies based on causal models and/or theory of quasi-experimentation.

  - Individual studies of specific quality variables.

  - Systematic reviews of available data from different intervention areas.

- **Develop quality scales in specific representative context:**

  - Clarify the kind of quality to assess

  - Precise delimitation of the intervention context

  - Provide metric data about its reliability

  - Explore consistency of quality assessment depending on whether the total score or individual items are used.

- ***Towards the Validation of a Scale to Measure the Quality of Primary Studies for Meta-analysis. Content validity. Why?***

SYSTEM OF CODING

Extrinsic characteristics.

1- Type of publication

1. Journal
2. Book
3. Thesis
4. Congress
5. Other ones

2- Publication year

3- Impact index (only in journals)

4- Data Bases

5- Training of researches

1. Especified
2. No data enough

6- Paper Structure recommended by APA

1. Yes
2. No

Substantives characteristics.

Subjects:

7- Age (Range) referred: Y/N

8- Age (mean)

9- Age standard deviation

10- Cultural origin

1. Only one
2. More than one
3. No data enough

11- Socioeconomic level

1. Low
2. Medium
3. High

SYSTEM OF CODING

Setting/ context:

12- Implementation context

1. Urban
2. Rural
3. Mixed

13- Intervention Field

1. Sanitary
2. Educational
3. Social
4. Clinical
5. Organizational
6. Others

14- Country

Treatment:

15- Theoretical orientation

1. Specified
2. Inferred
3. No data enough

16- Previous Empirical Evidence

1. Specified

2. No data enough

17- Period of treatment

18- Degree of Treatment Intensity (i.e. number of dosages)

19- Units

1. In group
2. Individual

20- Strengths and weakness are discussed (Y/N)

SYSTEM OF CODING

Methodological characteristics.

21- Inclusion and exclusion criteria for units; provided (Y/N)

22- Units random assignment

1. None and without control of extraneous variables
2. None but with control of extraneous variables.
3. Yes

23- Methodology or Design

1. Experimental; randomized
2. Quasi-experimental (two groups without randomized assignment) non-equivalent control groups with pre-test and post-test
3. Pre-Experimental (only one group, one measure)/ others (questionnaires/observational/naturalistic)

24- Sample size

1. n <5
2. 5 <n <10
3. n >10

25- Statistic calculate of sample size (Y/N) (magnitude of sampling error)

26- Attrition

1. >30%
2. <30%

27- Without mortality (N/Y)

28- Attrition between groups

1. Homogeneous
2. Non- homogeneous

29- Exclusions after randomisation (N/Y) (number) (i.e. outliers, non-codified, format errors…)

30- Baseline period

1. < 6 months
2. 6-11 months
3. > 12 months

31- Follow-up period

1. < 6 months
2. 6-11 months
3. > 12 months

SYSTEM OF CODING

32- Moments of measurement (y/n & number)

1. Post intervention
2. Pre and post intervention

33- Measures in pre-test appear in post-test

1. None
2. Some
3. All of them

34- Normalized dependent variables

1. Without (self-reports and post hoc records)
2. Questionnaires or standardized self-reports
3. At least one is objective (psycho-physiological measures)

35- Intervention/Study homogeneity

1. Subjects do not receive the treatment in the same contextual conditions
2. Subjects receive treatment in the same contextual conditions

36- Control Techniques

1. Blind (beneficiaries)
2. Blind (implementers)
3. Both
4. Other ones (necessary to specify)

37- Construct Definition of Outcome

1. Replicable by reader in own setting
2. Vague definition
3. No definition

38- Statistic methods for imputing missing data (Y/N)

39- Specification of confidence intervals in statistic analysis (Y/N)

40- Effect size and value

41- Other data apart aims

1. Positive effects
2. Negative effects.
3. Both
4. None

42- Interpretation of results

1. All
2. Some of them
3. None

43- Interpretation of results bias

1. All
2. Some of them
3. None

- **For further information:**

  E-mails:

- **Salvador Chacón Moscoso: schacon@us.es**
- **Susana Sanduvete Chaves: sussancha@us.es**
- **David Alarcón Rubio: dalarub@dts.upo.es**

- **Webpage:**
- **http://innoevalua.us.es (research group about methodological innovations in program evaluation; University of Sevilla, Spain)**