MEASURING METHODOLOGICAL QUALITY OF PRIMARY STUDIES FOR META-ANALYSIS. PRELIMINARY RESULTS FROM A PROPOSED SCALE

7th Annual International Campbell Collaboration Colloquium
London, May 14-17, 2007

Salvador Chacón Moscoso (University of Seville)
Susana Sanduvete Chaves (University of Seville)
Milagrosa Sánchez Martín (University of Seville)
Julio Sánchez Meca (University of Murcia)

Research supported by the proyect of Ministry of Spanish Sciencies and Education (MEC) "Quality of the design and effect size in training programs transfer evaluation" (ref. SEJ2004-5360/EDUC)

Index of contents

INTRODUCTION$_0$ (framework of analysis)

- We consider that there is a certain degree of correspondence between used criteria to choose and codify studies in meta-analysis 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 meta-analytic studies based on those evaluation results.
- Most researchers agree to obtain a global quantitative index of quality based on the scores, but it is not clear how to measure study quality about primary studies reliably and realistically:
  - different ways to understand quality (internal validity, external validity, relevance…).
  - 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

INTRODUCTION$_{(1)}$

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 this proposal can be easily discussed; 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.

INTRODUCTION

Experimental designs are minority in psychological, educational and social context (Chacón, Sánchez & Sanduvete, 2007; Chacón, Sanduvete & Alarcón, 2005, 2006; Sanduvete, 2006; Shadish, Chacón & Sánchez-Meca, 2005); around:

- 10% of cases used experimental designs.
- 30% of cases used a comparison group.
- Random assignment of units was only used in less than 30% of cases.
- Mask (the control technique characteristically used in experimental designs) hardly is used.

Obvioly, we are not against randomized designs, as when analyzing unbiased effect size, they have demonstrated their superiority (when properly plannified and implemented). But, taking into account previous studies, they do not clearly represent our intervention context neither.

If we consider 'quality' of evidence just as 'unbiased estimation on effect size', this is probably going to provoque an important gap of knowledge between the academic and the real social work! In sum, we would be considering just partial reviews of the available evidence

OBJETIVE

To present, not a global, but a specific scale to measure quality in primary studies –that is supposed to identify methodological features related to quality-; this implies:

1. To analyze its **homogeneity** with respect to scoring and weighting records using different methodologies in the same area and/or the same methodology in different areas.
   (This Implies taking into account, mainly existing theories of validity and measurement)

2. To apply it in different intervention contexts, and demonstrate how it can be **adapted** to the characteristics of intervention studies in the psychological, educational and social fields.
   (this implies taking into account, al least, models of generalizationc and utility)

Important threat to this logic is to believe that we can find a general answer to problems defined contextually.

Synthesis of METHODOLOGY since 2004 – units, instruments and procedures-

- Main stages to develop the methodological scale and main results since 2004:
  - Scale evolution (changes, reasons, consequences).
  - Results and implications during the process.
  - Present and expected future development of the project.
  - Invitation to use and assess feasibility, utility, level of representativeness and reliability of the scale.

	1st phase (2004)	2nd phase (2005)	3rd phase (2006)	4th phase (2007)
SAMPLE	25 available documents abut"measurement of quality in primary studies"	-30 experts in meta-analysis and systematic reviews, quality evaluation and design (university and practitioners)	- Scale with 34 items (result of content validity'05) - Other common available scales (i.e. Sánchez-Meca, research group collaboration, 1998)	- Scale with 33 items
INSTRUMENTS	- Electronic databases - Procite	- 43 items - 3 options - 3 concepts to evaluate: representativeness, utility, feasibility -Internet (sending of results) -Microsoft excel (data analysis)	- Items from content analysis: - 3 characteristics: - extrinsic - substantives - methodological - Other scales	-33 items from new scale
PROCEDURE	Selection of all recorded quality items	- Questionnaires were given to experts - Data were collected - Data were analyzed	- Comparison between other available scales and results of content analysis - New items were added. If it was useful to complete the items, new categories were added	Modifications in terms used and categories looking for: -Homologous comparison referents between designs (quality can sum from 0 to 19) - More concretion and operationalization
RESULTS	- Exploratory questionnaire (43 items) - Exploratory study of published int. Prog.(abstracts)	- Content validity: Osterlind's index - Criteria of inclusion: ≥ 0,5 in at least 2 concepts - 34 items -Exploratory study (abstracts)	-Scale with 33 items - Exploratory studies (abstracts)	- New resulting scale: 38 items - Study (full texts)

Reasons for introducing changes: consequences

	Reasons for introducing changes	Consequences
From 1st (2004) to 2nd phase (2005)	Content validity which gave an exclusion criteria	Basically, **elimination** of less representative, useful and/or feasible items
From 2nd (2005) to 3rd phase (2006)	Complete content validity study with scales that before were not available	Basically, **incorporation** of complete items or partial categories: - More concretion - More utility - More complete
From 3rd (2006) to 4th phase (2007)	- Some designs could punctuate in a higher way than others - Some items were ambiguous	Basically, **terminological modification** and partial **incorporation** of categories to find: - An homologous comparison between designs - More concretion - More utility - More operationalization

RESULTS

Methodological characteristics

ITEM	VALUE	CATEGORY
0. Type of study	0	Theoretical
	1	Observational
	2	Survey
	3	Quasi-experimental
	4	Experimental
1. Control group	0	No
	0.5	Inactive
	1	Active
2. Sample selection criteria	0	Non Specified
	1	Specified
3. Randomization	0	Non randomized when there is not intervention/ Pre-experimental/ Quasi without variables controlled/ Experimental with an incorrect random process
	1	Randomized when there's not intervention/Quasi with variables controlled/Experimental

RESULTS
Methodological characteristics

ITEM	VALUE	CATEGORY
4. Design	0	1-2 measurements when there is not intervention/ Pre-experimental/ Experimental with one pre-post moment of measurement
	0.5	3-29 measurements when there's not intervention/ Quasi-experimental with 2-29 measurements
	0.75	Temporal series
	1	30 or more measurements when there is not intervention/ Discontinuity on regression/ Experimental with 2 or more moments of measurement
5. Sample	0	N<12
	0.5	12≤N≤40
	1	N>40
6. Global attrition	0	≥20%
	0.5	0<N<20%
	1	0%

RESULTS
Methodological characteristics

ITEM	VALUE	CATEGORY
7. Differential attrition	0	≥20%
	0.5	0<N<20%
	1	0%
8. Exclusions after sample assignment	0	≥20%
	0.5	0<N<20%
	1	0%
9. Follow-up	0	None
	0.3	< 6 months
	0.6	[6-11] months
	1	≥12 months
10. Moments of measurement	0	After intervention/ only one measurement if there is not intervention
	1	Before and after intervention/more than one measurement if there is not intervention

RESULTS
Methodological characteristics

ITEM	VALUE	CATEGORY
11. Measurements in every moments	0	More than one measurement doesn't appear in every measure moments
	0.5	One measurement doesn't appear in every measure moments
	1	All the measurements appear in every measure moments
12. Normalized dependent variables	0	Auto-registration non standardized
	0.5	At least one is a questionnaire or auto-registration standardized
	1	At least one is objective or normalized
13. Mask in evaluator	0	No
	1	Yes
14. Mask in participants	0	No
	1	Yes

RESULTS
Methodological characteristics

ITEM	VALUE	CATEGORY
15. Mask in professional of intervention/ internal evaluator	0	No
	1	Yes
16. Homogeneity in process: intensity, duration and professionals	0	Different
	1	Same
17. Construct definition	0	None is defined conceptual and empirically
	0.5	At least one is defined conceptual and/or empirically
	1	All of them are defined conceptual and empirically
18. Statistic methods to infer missing data	0	None
	1	Yes

RESULTS
Methodological characteristics

ITEM	VALUE	CATEGORY
19. Effect size and value	0	Non specified
	1	Specified
20. Index of quality	SUM	From 0 to 19
21. Statistic index calculated	Concrete value	
22. Significant differences between measures	0	No
	1	Yes
23. Variability index	Concrete value	

RESULTS
Substantive characteristics

ITEM	VALUE	CATEGORY
24. Number of participants each group	Concrete Value	
25. Number of groups	Concrete Value	
26. Exclusions after measurements	Concrete Value	
27. Age range	Concrete Value	
28. Average age	Concrete Value	
29. Period of study	Concrete Value	
30. Intensity of the treatment or the measurements when there is not intervention	Concrete Value	
31. Unit of measurement	Concrete Value	
32. Training area	Concrete Value	
33. Intervention field	Concrete Value	

RESULTS
Extrinsic characteristics

ITEM	VALUE	CATEGORY
34. Type of publication	1	Article in journal
	2	Book
	3	Thesis
	4	Congress
	5	Other publications
	6	Non-published studies
THEORETICAL MODEL		
35. Author		
36. Variables used		
37. Evaluation proposal		

LET'S DISCUSS

QUESTIONNAIRE	LAST VERSION	REASONS
22. Units random assignment: 0. None and without techniques to control extraneous variables. 0.5. None but with control of extraneous variables. 1.Yes.	3. Randomization: 0. Non randomized when there is no intervention/Pre-experimental/Quasi without variables controlled/ Experimental with an incorrect random process. 1. Randomized when there's no intervention/ Quasi with variables controlled/ Experimental with a corrects random process	- **Randomization was scored higher** than other kinds of assignment, so experimental design was the only one able to obtain highest score. Quasiexperimental designs with control of extraneous variables are now considered with the same quality than experimental designs. - Before, only was considered assignment; now, we also consider **selection** in studies without assignment.

QUESTIONNAIRE	LAST VERSION	REASONS
32. Occasions of measurement on each variable (specify number): 0. Post intervention only. 1. Pre and post intervention.	10. Moments of measurement: 0. After intervention/only one measurement if there's not intervention. 1. Before and after intervention/ more than one measurement if there's not intervention.	- Also adapted to cases **which there is not intervention**.

DISCUSSION AND CONCLUSIONS

- Referred to the **homogeneity** of scores in different designs:
  - This scale is useful to prove that not necessarily a **relationship** exists **between degree of quality** and **design** features because experimental designs don't always present high degree of quality and, nevertheless, other kind of designs (quasi-experimental, pre-experimental, survey studies or observational ones) may have high degree of quality.
  - Maybe to determine the **efficacy** of a study with experimental designs is the best option, but the real situation makes us think that it's really necessary to take into account the other kinds of design.
- Referred the capacity to **adaptation** to different contexts:
  - Although it might not be applicable in every cases, we think it's more **adaptative** than other previous scales because
  - This scale tries to be **flexible** and representative to the real characteristics of intervention studies in the psychological, educational, and social fields.
- The **implicit objectives** of this work were:
  - Foster the idea of quality methodological scales as an useful tool to enhance homogeneous interventions.
  - Develop from an inductive view intervention models, that practitioners use, but without practically any systematic record.

FUTURE DEVELOPMENTS OF THE PROJECT

- Making an **empirical check**, comparing results in quality index of primary studies previously measured with other scales. In case of finding differences, it would give interesting plausible different alternatives hispothesis.
- We already presented an application of this scale in **training programs**. We are going to apply it in other contexts to see how it works.

FUTURE DEVELOPMENTS OF THE PROJECT

We invite you to use and assess feasibility, utility, level of representativeness and reliability of this scale.

http://innoevalua.us.es

(SCALE AVAILABLE ON LINE OR BY E-MAIL; previous request to authors)

THANK YOU FOR YOUR ATTENTION

REFERENCES

- Chacón, S., Sánchez, J. y Sanduvete, S. (2007, February). Measuring the quality of primary studies in meta-analysis. Paper presented at *X Congress of Methodology in Social and Health Sciences*. Granada.
- Chacón, S., Sanduvete, S. & Alarcón, D. (February, 2005). Towards the Validation of a Scale to Measure the Quality of Primary Studies for Meta-analysis. Paper presented at the *V Annual Campbell Collaboration Colloquium*. California.
- Chacón, S., Sanduvete, S. & Alarcón, D. (2006, september). Content validity of a scale to assess the quality of primary studies in Meta-analysis. Paper presented at *IX Congress of Methodology in Social and Health Sciences*. Granada.
- Chacón, S., Sánchez, J., Sanduvete, S. & Alarcón, D. (in process). Validation of a scale to measure the quality of primary studies for meta-analysis. *Evaluation and program planning*.
- Osterlind, S. J. (1998). *Constructing tests items*. Boston: Kluwer Academic Publishers.
- Sánchez-Meca, J., Rosa, A. I. & Olivares, J. (1998). Cognitive-behavioral techniques in clinic and healthy disorders. Meta-analysis of Spanish literature. *Psicothema, 11*(3), 641-654.
- Sanduvete, S. (2006). Methodological advances to improve interventions with elderly people. Not published dissertation.
- Shadish, W. R., Chacón, S. & Sánchez-Meca, J. (2005). Evidence-based decision making: enhancing systematic reviews of program evaluation results in Europe. *Evaluation, 11*(1), 95-109.

APENDIX

RESULTS 1
Content validity
(Chacón, Sanduvete y Alarcón, 2005)

EXTRINSIC CHARACTERISTICS (N = 10)	R	U	F
1. Type of publication	0.3	**0.6**	**0.7**
2. Year of publication	0	0.2	**0.8**
3. Impact index (only in journals)	-0.2	0.1	0.3
4. Data Bases	-0.2	0.3	**0.6**
5. Training of researches	0.2	0.4	0.1
6. Paper Structure recommended by APA	0.1	0.1	0.1

RESULTS 1

SUBSTANTIVE CHARACTERISTICS (N = 30)	R	U	F
SAMPLE			
7. Range of age	**0.6**	**0.5**	**0.6**
8. Mean of age	**0.8**	**0.8**	**0.7**
9. Standard deviation of age	0.4	0.1	0.4
10. Cultural origin	0.1	0.2	0.3
11. Socio-economic level	-0.1	0.1	-0.3
CONTEXT			
12. Implementation context	-0.2	0.1	0
13. Intervention field	**0.5**	0.4	**0.9**
14. Country	0.4	0.4	**0.7**
TREATMENT			
15. Theoretical orientation	0.3	**0.8**	0
16. Previous empirical evidence	0.1	0.3	0.1
17. Period of treatment	**0.8**	**0.9**	**06**
18. Degree of treatment intensity	**0.8**	**0.9**	**0.8**
19. Units (in group or individual)	**1**	**0.9**	**0.9**
20. Strengths and weaknesses are discussed	0.4	-0.1	0

RESULTS 1

METHODOLOGIC CHARACTERISTICS (N = 30)	R	U	F
21. Inclusion and exclusion criteria for units are provided	**0.6**	**0.9**	**0.5**
22. Units random assignment to groups	**0.9**	**1**	**0.6**
23. Type of methodology/ design	**0.9**	**0.9**	**0.6**
24. Sample size	**0.8**	**0.9**	**1**
25. Statistic used to calculate the sample size	0.4	**0.5**	0.3
26. Attrition	**0.7**	**0.9**	0.1
27. Without attrition	**0.6**	**0.5**	0.4
28. Attrition between groups	**0.7**	**0.9**	0.1
29. Exclusions after randomization	**0.6**	**0.6**	0.2
30. Baseline period	0.1	0.2	0
31. Follow-up period	**0.6**	**0.7**	0.3

RESULTS 1

METHODOLOGICAL CHARACTERISTICS (II) (N = 30)	R	U	F
32. Moments of measurement	**0.9**	**0.9**	**1**
33. Measures in pretest appear in postest	**0.8**	**0.9**	0.4
34. Normalized dependent variables	**0.6**	**0.6**	0.4
35. Homogeneity of the intervention	**0.6**	0.4	-0.1
36. Control techniques	**0.7**	**0.9**	0.2
37. Construct definition of outcome	**0.9**	**0.7**	-0.1
38. Statistic methods for inputting missing data	**0.6**	**0.6**	0.2
39. Specification of confidence intervals in statistical analysis	0.1	0.2	**0.5**
40. Effect size and value	**0.7**	**0.8**	**0.6**
41. Other data apart aims	0.1	0.2	0.4
42. Interpretation of results	0.1	0.1	0.2
43. Interpretation of results bias	0.4	0.2	0.1

RESULTS 2

Results joined to items frequently used in meta-analysis (Sánchez-Meca)

Methodological characteristics

ITEM	ORIGIN
0. Type of study	General
1. Control group	Frequently used in meta-analysis
2. Sample selection criteria	Content validity (R, U & F ≥ 0.5)
3. Randomization	Frequently used in meta-analysis and content validity (R, U & F ≥ 0.5)
4. Design	Frequently used in meta-analysis and content validity study (R, U & F ≥ 0.5)
5. Sample	Frequently used in meta-analysis and content validity study (R, U & F ≥ 0.5)

RESULTS 2.
Methodological characteristics

ITEM	ORIGIN
6. Global attrition	Frequently used in meta-analysis and content validity study (R & U ≥ 0.5)
7. Differential attrition	Frequently used in meta-analysis and content validity study (R & U ≥ 0.5)
8. Exclusions after sample assignment	Content validity study (R & U ≥ 0.5)
9. Follow-up	Frequently used in meta-analysis and content validity study (R, U & F ≥ 0.5)
10. Moments of measurement	Frequently used in meta-analysis and content validity study (R & U ≥ 0.5)
11. Measurements in every moments	Frequently used in meta-analysis and content validity study (R & U ≥ 0.5)
12. Normalized dependent variables	Frequently used in meta-analysis and content validity study (R & U ≥ 0.5)
13. Mask in evaluator	Meta-analysis and cont. validity (R & U)

RESULTS 2.
Methodological characteristics

ITEM	ORIGIN
14. Mask in participants	Meta-analysis and cont. validity (R & U)
15. Mask in professional of intervention	Meta-analysis and cont. validity (R & U)
16. Homogeneity in intervention (or process of measurement)	Frequently used in meta-analysis and content validity study (not considered good)
17. Construct definition	Content validity study (R & U ≥ 0.5)
18. Statistic methods to infer missing data	Content validity study (R & U ≥ 0.5)
19. Effect size and value	Content validity study (R, U & F ≥ 0.5)
20. Index of quality	Sum
21. Statistic index calculated	Possible modulator variable
22. Significant differences	Possible modulator variable
23. Variability index	Possible modulator variable

RESULTS
Substantive characteristics

ITEM	ORIGIN
24. Number of participants each group	Possible modulator variable
25. Number of groups	Possible modulator variable
26. Exclusions after measurements	Possible modulator variable
27. Age range	Content validity study (R, U & F ≥ 0.5)
28. Average age	Content validity study (R, U & F ≥ 0.5)
29. Period of study	Content validity study (R, U & F ≥ 0.5)
30. Intensity of the treatment	Content validity study (R, U & F ≥ 0.5)
31. Unit of intervention or measurement	Content validity study (R, U & F ≥ 0.5)
32. Training area	Possible modulator variable
33. Intervention field	Content validity study (R & F ≥ 0.5)

RESULTS
Extrinsic characteristics

ITEM	ORIGIN
34. Type of publication	Content validity study (U & F ≥ 0.5)
THEORETICAL MODEL	
35. Author	To acquire a general view
36. Variables used	To acquire a general view
37. Evaluation proposal	To acquire a general view