A SCALE TO MEASURE THE QUALITY OF PRIMARY STUDIES FOR META-ANALYSIS

A CONTENT VALIDITY STUDY

The Sixth International Campbell Collaboration Colloquium
California, February 22-24, 2005

Salvador Chacón Moscoso
Julio Sánchez Meca
Susana Sanduvete Chaves
David Alarcón Rubio

INTRODUCTION (I)

- Meta-analysis and systematic reviews pretend integrating the findings 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 is to develop clear criteria to decide how 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 and those design components that are relevant to enhance quality in particular interventions and in their findings integration.

INTRODUCTION (II)(State of the art I)

- 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**

INTRODUCTION (II)(State of the art II)

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.**

INTRODUCTION (II)(State of the art III)

***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**

INTRODUCTION (II)(State of the art IV)

***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**

An example: Jadad's scale (1996) (0-5 range points in a global score):

- **Randomization (maximum 2 points)**
  - **does the study describe a randomized procedure?**
  - **Is the randomized procedure adequate?**

- **Blinding/ Masking (maximum 2 points)**
  - **does the study describe a double masking procedure?**
  - **Is the double masking procedure appropriate?**

- **Attrition (Maximum 1 point)**
  - **is subject attrition well described, as well as the possible reasons?**

INTRODUCTION (II) (State of the art V)

***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**

INTRODUCTION (III)

(Some features of intervention programs in primary studies)

- Most of intervention programs in different areas lack methodological rigor in some aspects (Chacón, Sanduvete & Alarcón, submited).
  - 17.7% specifies **theoretical orientation**.
  - 15.7% assigns people to different groups **randomly**
  - 14.7% uses a **experimental design**.
  - 15.3% presents a **follow-up period** longer than a year.
  - 31.4% presents **measurements before and after** the intervention.
  - 5.3% presents **every** measures before and after the intervention.
  - 6.4% presents at least a **objective measurement**.
  - 17% reports **effect size**.

INTRODUCTION (IV)

Why is it interesting to use a scale to measure the quality of primary studies?

INTRODUCTION (IV)

For example: Considerations about randomized versus non 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
- High quality 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.

INTRODUCTION (IV)

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)

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

Today, or maybe never, will be 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

INTRODUCTION (IV)

- Then in this context: Why is it interesting to use a scale to measure the quality of primary studies?
  - It is the most used method (Jüni, Altman & Egger, 2001) .
  - A global quantitative data reflects the quality of the study.
  - Develop explicit criteria to decide how to choose, codify and combine studies with different degrees of quality.
  - Reliability and validity indexes could be studied.

INTRODUCTION (IV)

- If there are so many scales, why is interesting to develop another one (Most researches in social research consider they are not useful as it is not possible to have a global assessment of quality for specific areas of research.)?
  - To try to systematize useful indicators.
  - To study systematically and empirically if quality scales are really useful or not.

OBJECTIVE

- To study the content validity of a quality scale that includes most frequent items used to assess quality.
- Focus attention on methodological features.

METHOD First phase (develop the scale):

- Sample: review available articles about measuring quality in primary studies (Sánchez-Meca & Ato (1990); O'Rourke & Detsky (1989); Weisz et al. (2000); Tritchler (1999); Jüni et al. (2001); Sánchez-Meca (1997); Sutton et al. (2000); Moher et al. (2001); Begg et al. (1996); Moher et al. (1998); Des Jarlais et al. (2004); Bossuyt et al. (2003); Bossuyt et al. (2003); Olivares et al. (2000); Education Group for Guidelines on Evaluation (1999); Campbell et al. (2004); Bosch et al. (2003); Altman et al. (2001); Brown (1991); Jüni et al. (1999); McGuire et al. (1985); Emerson et al. (1990); Moher et al. (1996); McPherson et al. (2002); Moher et al. (2001); Greenland (1994)…).

- Instrument: Database to research documents related to quality in primary studies.

- Procedure: to develop the scale we collected usually cited quality items from literature; we obtained a draft list containing 43 items grouped in three dimensions:

  a) Extrinsic characteristics of the studies (6 items)
  b) Substantive characteristics (14 items)
      b1) Sample (5 items)
      b2) Intervention context (3 items)
      b3) Treatment (6 items)
  c) Methodological Characteristics (23 items)

METHOD Usually cited quality items:

- **Extrinsic Characteristics:**
- 1- Type of publication (1.journal, 2.Book, 3.Thesis, 4.congress, 5. other ones)
- 2- Year of publication
- 3- Citation Impact factor for the journal in which an article appeared
- 4- Is the raw data from the study available?
- 5- Training of treatment implementers (1. specified; 2. No data enough)
- 6- APA format

- **Substantive Characteristics –Sample-**
- 7- Did the study report participant 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).

- **Substantive Characteristics -Setting-**
- 12- Implementation context (1.Urban; 2. Rural; 3.Mixed)
- 13- Intervention Field (1.Inpatient clinical; 2.Educational; 3.Social; 4.Outpatient clinical; 5.Organizational; 6.Others)
- 14- The authors report the country in which study was conducted**Substantive Characteristics –**

METHOD Usually cited quality items:

Substantive Characteristics –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 (quantitative time)
- 18- Degree of Treatment Intensity (i.e. number of dosages)
- 19- Units (1. In group; 2. Individual)
- 20- Strengths and weakness of treatment are discussed (Y/N)

- **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 <15; 2. 15 <n <30; 3.n >30)
- 25- Did the authors say they did a power analysis to calculate sample size (Y/N)
- 26- Attrition (1. >30%; 2. <30%)

METHOD Usually cited quality items:

- **Methodological Characteristics**
- 27- No attrition occurred (N/Y)
- 28- Attrition between groups (1. Homogeneous; 2. Non- homogeneous)
- 29- Exclusions after randomization (N/Y) (specify number)
- 30- How long were units studies before treatment implementation (1.< 6 months; 2. 6-12 months; 3.> 12 months)
- 31- Follow-up period (1. < 6 months; 2. 6-12 months; 3. > 12 months)
- 32- Occasions of measurement on each variable (specify number; 1. Post intervention only; 2. Pre and post intervention)
- 33- Measures in pre-test appear in post-test (1.None; 2.Some; 3.All of them)
- 34- Standardized dependent variables: 1.Without (self-reports and post hoc records); 2. Standardized questionnaires or standardized self-reports.
- 35- Intervention context 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- Statistical methods for imputing missing data (Y/N; Specify)
- 39- Specification of confidence intervals in statistic analysis (Y/N)
- 40- Effect size value
- 41- Effectiveness of treatment (1.Positive effects; 2. Negative effects; 3. Both; 4.None
- 42- Interpretation of results (1.All; 2.Some of them; 3.None)
- 43- Discussion of bias and limitations (1.All; 2.Some of them; 3.None)

METHOD Second phase (expert judge):

- Sample: 30 'professionals assessed items': 13 experts in meta-analysis and systematic reviews and 17 applied psychologists (social, education, developmental, clinical).

- Instruments: We developed a Content validity Questionnaire. Professionals assessed each of the 43 items from 1 (minimum level) to 5 (maximum level) with respect to representativeness, utility and feasibility.

  a) Representativeness: How much the specific item represents the quality subdomain where it was assigned?
  b) Utility: How much the specific item is useful to assess the quality of the study with respect to the quality domain where it is assigned.
  c) Feasibility: How feasible is that item to code.
  d) Also, experts included any comments they considered in each item, and had the possibility to suggest any other item that they thought it is important to take into account.

  - We used Microsoft Excel software for data analysis.

- Procedure: Distribution of the questionnaires and collection of data was done by e-mail and in person during the fifth *Annual Campbell Collaboration Colloquium,* Lisbon, February 2005 and IX Conference of the Spanish Association of Methodology, Granada, September, 2005.

Data analysis:

Study of content validity with congruence index (Osterlind,1998)

$$I_{ik} = \frac{(N-1)\sum_{j=1}^{n} X_{ijk} + N\sum_{j=1}^{n} X_{ijk} - \sum_{j=1}^{n} X_{ijk}}{2(N-1)n}$$

- N = number of domains
- X_{ijk} = value that each judge give to each item
- n = number of judges

I_{ik} = 0.5 minimum level of congruence

RESULTS

EXTRINSIC CHARACTERISTICS (N = 30)	R	U	F
1. Type of publication	0.4	**0.6**	**0.7**
2. Year of publication	-0.1	0.2	**0.9**
3. Impact index	-0.1	0.1	0.3
4. Database where it were found	-0.2	0.4	0.4
5. Training of researches	0.1	**0.5**	0
6. Paper structure recommended by APA	-0.1	0	0.1

RESULTS (II)

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 (III)

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 (IV)

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 aids	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

DISCUSSION

- The number of suitable items will vary depending on the **cut-point** to use (from 129 values, using 0.5 as cut-point, 60 were suitable; if the cut-point were 0.7, there would be only 37 suitable) and the assessed concepts (representativeness, utility and feasibility).

FUTURE DEVELOPMENTS

- **Comparison** of meta-analysis results with quantitative indexes of quality of primary studies included in already performed meta-analytical studies obtained from existing scales and the one presented in this work, including different combinations:
- Every items with positive values in representativeness, utility and feasibility.
- Other combinations (only representativeness and utility).

FUTURE DEVELOPMENTS

- **Methodological features seems to have ‘highest scores’** focusing attention on methodological quality features of primary studies as ‘more generalize able’ than extrinsic and substantives ones.

focusing attention on methodological quality

Scale to codify methodological characteristics of primary studies (I)

1. Control Group: 0-inactive; 1: active
2. Units Assignment criteria; (S,C&C'02; p.323): 0- non-specified; 1-specified.
3. Design: 0-pre-experimental; 0'5-quasi-experimental; 0'75- Interrupted time series (pre obs n ≥30 & post obs n ≥30) and/or Regression Discontinuity Design; 1- Experimental Randomized.
4. Sample size: 0 - n<12; 0'5 - n=[12-39]; 1- n≥40.
5. Global Attrition: 0- ≥20%; 0'5 - %=]0-20[; 1- 0%6.
6. Differential attrition: 0- ≥20%; 0'5 - %=]0-20[; 1- 0%.
7. Follow-up period: 0 - < 6 meses; 0'5 – meses = [6-11]; 1- ≥12 meses.
8. Occasions of measurement on each variable: 0- Post intervention only; 1- Pre and post intervention.
9. Measures in pre-test that do not appear in post-test: : 0- >1; 0'5=1; 1=none.

focusing attention on methodological quality

Scale to codify methodological characteristics of primary studies (II)

10. Standardized dependent variables: 0- only non-standardized self-reports; 0'5 at least one standardized measure; 1- Standardized questionnaires or standardized self-reports or meausres.
11. Blind (evaluator): 0- non-specified; 1-specified.
12. Blind (implementers): 0- non-specified; 1-specified.
13. Blind (implementers); 0- non-specified; 1-specified.
14. Intervention context homogeneity: 0- Subjects do not receive the same intensity of treatment, during the same time period and by the same professional; 1 - Subjects receive the same intensity of treatment, during the same time period and by the same professional.
15. Construct Definition of Outcome/s: 0 –non-specified; 0'5 – specified (without empirical definitions); 1 – empirical definition/s.
16. Missing data analysis: 0-no-specified ('completers analysis'); 1-specified ('intention-to-treat analysis').

Global methodology quality score: adding up scores from item 1 to 16.

Focusing attention on methodological quality of primary studies & its relation With already performed meta-analysis results

- For example, effect sizes from primary studies included in a ‘panic disorders’ meta-analysis:
- (standardized mean difference between treatment and control groups)
- ‘d’ temporary measures:
- In pretest
- In post-test
- During follow-up periods

PRACTICAL POINT OF VIEW

Practical general process proposal to follow in order to codify studies when performing a meta-analysis and/or 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.

For further information:

E-mails:

- **Salvador Chacón Moscoso: schacon@us.es**
- **Julio Sánchez Meca: jsmeca@um.es**

- **Webpages:**

- **http://www.um.es/fcapsi/metaanalysis (Meta-analysis unit from University of Murcia, Spain)**

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