Psicothema

Psicothema 2014, Vol. 26, No. 1, 91-96
doi: 10.7334/psicothema2013.144

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Methodological convergence of program evaluation designs

Salvador Chacón-Moscoso[1], M. Teresa Anguera[2], Susana Sanduvete-Chaves[1] and Milagrosa Sánchez-Martín[1]

[1] Universidad de Sevilla, and [2] Universidad de Barcelona

Abstract

Background: Nowadays, the confronting dichotomous view between experimental/quasi-experimental and non-experimental/ethnographic studies still exists but, despite the extensive use of non-experimental/ethnographic studies, the most systematic work on methodological quality has been developed based on experimental and quasi-experimental studies. This hinders evaluators and planners' practice of empirical program evaluation, a sphere in which the distinction between types of study is changing continually and is less clear. **Method:** Based on the classical validity framework of experimental/quasi-experimental studies, we carry out a review of the literature in order to analyze the convergence of design elements in methodological quality in primary studies in systematic reviews and ethnographic research. **Results:** We specify the relevant design elements that should be taken into account in order to improve validity and generalization in program evaluation practice in different methodologies from a practical methodological and complementary view. **Conclusions:** We recommend ways to improve design elements so as to enhance validity and generalization in program evaluation practice.

Keywords: Validity, generalization, structural design dimensions, evaluation research, methodological complementarity.

Resumen

Convergencia metodológica de los diseños de evaluación de programas. **Antecedentes:** por una parte, actualmente todavía existe la visión dicotómica en que se presentan confrontados los estudios experimentales/cuasi-experimentales y no-experimentales/etnográficos; y por otra parte, a pesar del extendido uso de los estudios no-experimentales/etnográficos, el trabajo más sistemático sobre calidad metodológica se ha llevado a cabo en los estudios experimentales y cuasi-experimentales. Esto dificulta la práctica de quienes evalúan y planifican los programas a nivel empírico, un área donde la distinción entre tipos de estudio está en cambio constante y es menos clara. **Método:** tomando como referencia el marco clásico de validez en estudios experimentales/cuasi-experimentales, realizamos una revisión de la literatura con el fin de analizar la convergencia de los elementos de diseño en calidad metodológica de los estudios primarios en revisiones sistemáticas e investigación etnográfica. **Resultados:** explicitamos los elementos de diseño relevantes que habrían de tenerse en cuenta para mejorar la validez y generalización en evaluación de programas en las diferentes metodologías desde una aproximación práctica de complementariedad metodológica. **Conclusiones:** proponemos recomendaciones para mejorar los elementos de diseño y así potenciar la validez y la generalización en la práctica de evaluación de programas.

Palabras clave: validez, generalización, dimensiones estructurales del diseño, investigación en evaluación, complementariedad metodológica.

The present paper introduces a summary of advances obtained in methodological complementarity in program evaluation designs by our research group over the last 25 years. These advances have been presented regularly in conferences of European Association of Psychological Assessment (EAPA), European Association of Methodology (EAM) and Asociación Española de Metodología de las Ciencias del Comportamiento (AEMCCO) since 1990.

Different methodological approaches to program evaluation designs

A classic analysis of the difference between experiments, quasi-experiments and observational studies might refer only to the degree of control the practitioner has over the intervention (Chacón, Sanduvete, Portell, & Anguera, 2013). Thus, a randomized experiment (R-E; also called a randomized controlled trial, RCT) is a study in which an intervention is deliberately introduced to observe its effects, and whose units are assigned randomly to conditions (Shadish, Cook, & Campbell, 2002). A quasi-experiment (Q-E) does the same, except that the units are not randomly assigned to conditions. Finally, an observational design, also called non-experimental/ethnographic study (N-E), is the systematic recording and quantification of behavior as it occurs, without manipulating it, in natural or quasi-natural settings, and where the collection, optimization, and analysis of the data thus obtained is underpinned by observational methodology (Anguera, 2008; Chacón et al., 2013).

Following this criterion, a N-E might be called a low-level intervention study (observers do not have any control over the situation and they simply observe behaviors that appear according to the subjects'/users' wish). A Q-E could be regarded as a medium-level intervention study (practitioners have a

Received: May 20, 2013 • Accepted: September 6, 2013
Corresponding author: Salvador Chacón-Moscoso
Facultad de Psicología
Universidad de Sevilla
41018 Sevilla (Spain)
e-mail: schacon@us.es

certain degree of control over the situation, without being able to assign subjects randomly to conditions, and they can provoke subjects' behaviors by manipulating variables). Finally, a R-E might be referred to as a high-level intervention study (practitioners have a high degree of control over the situation, and provoke behaviors). Of course, this distinction between low-, medium-, and high-level intervention studies is merely one of convenience, as some Q-E can involve high control over assignment to the intervention, such as in the case of regression discontinuity designs (Shadish et al., 2002). Indeed, the distinction is more a matter of degree rather than something which is absolute (Anguera, 2008).

For several decades now attention has been focused almost exclusively on strong interventions, i.e. those in which there is some control over the situation to be evaluated and where program users are given instructions to ensure that the actions are implemented according to the practitioners' plan (Cook, Scriven, Coryn, & Evergreen, 2010). However, there has been a progressive rise in the number of programs that are implemented without instructions and which take place in the natural and/or habitual context of program users, taking advantage of their spontaneous and/or usual everyday activities. Indeed, the literature on methodological quality in primary studies is basically focused on what are known as randomized controlled trials (RCT/R-E) (e.g., Auweiler, Müller, Stock, & Gerber, 2012). In the present paper it is argued that the different levels of intervention (high, medium and low) form part of a continuum. While the two ends of this hypothetical continuum (non-experimental/ethnographic and experimental methodologies) are apparent, it becomes more difficult once we enter the domain that separates them to establish clear criteria for distinguishing between different designs. This problem is heightened by the fact that in the sphere of real intervention the procedures used in the same evaluation program may vary and change (Kundin, 2010), hence the use of the expression 'design mutability' (Anguera, 2001; Anguera & Chacón, 1999). For example, a program to exercise the memory of elderly people in a seniors' day-care can be considered a Q-E at the first moment because new activities are included in the users' life; nevertheless, after two years, this program becomes part of the daily life of these elderly people, so the same activities could be then considered a N-E.

In this sense, scientific validity is a property of knowledge claims, not methods. No method guarantees validity. As such, we can use the same validity logic to judge knowledge claims whether they come from RCT/R-E (Shadish, 1995). Of course, these three methods often focus on different objects of evaluation (that is, different focal questions of interest), with R-E and Q-E being oriented to assess causal inferences while N-E is mainly focused on descriptions.

The aims of the paper are as follows: (1) to specify systematically the relevant design elements that should be taken into account in order to improve validity and generalization, thereby providing professionals with guidelines for choosing a given variant during the continuous decision-making process of evaluation; (2) to specify systematically the degree of correspondence/complementarity between these elements in the structural dimensions of an evaluation design, from the point of view of different methodologies and intervention contexts; and (3) to recommend ways of improving design elements so as to enhance validity and generalization in program evaluation practice.

Conceptual structure for analyzing the evaluation process from the perspective of the validity framework

Thus, rather than describing the differential characteristics of different kinds of methodology that are difficult to apply in a real intervention, we introduce basic structural design dimensions that will be used to distinguish between N-E, Q-E and RCT/R-E. We will also show how these dimensions might relate to the different types of validity, doing so on the basis of the broad components of $UTOST_i$: users/units (U), treatment (T), data/instruments, i.e. outcomes (O), setting (S) and time (T_i). Such a description suggests that even if they provide different solutions, every evaluation design should consider these aspects *a priori* in order to increase the rigor of a specific evaluation and the likelihood that the results will be generalizable (Chacón et al., 2013).

Derived from previous research (Campbell, 1957; Campbell & Stanley, 1963; Cook & Campbell, 1979; Cronbach, 1982; Shadish et al., 2002), Figure 1 shows a conceptualization of the four types of validity.

Without considering the causal relationship between 't' and 'o', this coding system can be also applied to non-causal studies, such as N-E, in which elements referred to units, treatments, outcome, setting and time are found, without necessarily having to suggest cause-effect relationships (Shadish, 1995).

Structural design dimensions in program evaluation according to the conceptual structure of the validity framework

The generalization of evaluative results refers to the possibility of drawing conclusions that can be applied to particular intervention contexts (Chacón et al., 2013). This analysis must take into account the five aspects of validity mentioned above: units, treatment (or program actions) observations/outcomes, setting and time. Furthermore, two types of generalization can be performed for each of these aspects of validity (Chacón & Shadish, 2008; Shadish et al., 2002): firstly, there is the question of which population constructs ($UTOST_i$) are associated with specific cases of '$utost_i$', those used

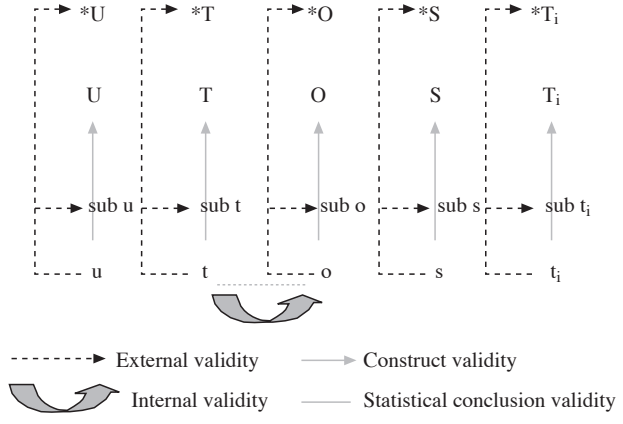

Figure 1. *Conceptualization of validity. $utost_i$ = units, treatment, outcome, setting and time in a particular sample; sub-$utost_i$ = the same elements in different subgroups within the study sample; UTOSTi = the same elements referred to the defined population; *U*T*O*S*Ti = the same elements in a differential population. Adapted from "Validity in program evaluation," by S. Chacón and W. R. Shadish, 2008, in Evaluación de programas sociales y sanitarios. Un abordaje metodológico, p. 74. Copyright 2008 by Síntesis*

in the evaluation (construct validity), and secondly, the question of the degree to which the results obtained in a particular evaluation may be extrapolated to different populations (*U*T*O*S*T_i) or to sub-samples of the same study (sub-utost$_i$) (external validity).

Within this framework for analyzing generalization the practitioner who is designing an evaluation, be it experimental, quasi-experimental or non-experimental/ethnographic, must from the outset consider a series of basic structural design dimensions that are linked to each one of these aspects of 'UTOSTi', in order to increase the rigor of the specific evaluation and the likelihood that the results will be generalizable.

As regards *units (users)*: (a) the *selection criteria* refer to the reasons why some individuals are eligible to participate in the study while others are explicitly excluded (e.g., Lord & Kuo, 2012); (b) the *assignment criteria* are the procedures used to include the units in specific comparison groups/conditions (Shadish et al., 2002); and (c) the *comparison groups/conditions* refer to each opportunity there is to apply or not apply the treatment, generally individually or on a group basis (Shadish et al., 2002).

With respect to the *treatment* or *program actions* (Anguera, 2008): (a) the *level of intervention* indicates the degree to which the treatment or program actions alter the everyday routines of participants; and (b) *changes in the level of intervention* refer to modifications over time in the degree to which the participants' everyday routines are altered.

Regarding the *results* and/or the *instruments* with which these are obtained (*outcomes*): (a) the *type of data* refers to the scale on which the gathered information is found, usually classified as: nominal, ordinal, interval or ratio (Matisoff & Barksdale, 2012); (b) *data quality,* the *justification of instruments* and the *types of instruments* are defined according to the degree of standardization of the instruments used (Anguera, Chacón, Holgado, & Pérez, 2008; Shadish et al., 2002); and (c) the *changes in instruments* indicates whether, over time, the dependent variable was measured in different ways (Anguera, 2001).

In relation to the *setting (implementation context)*: (a) the *aspects related to feasibility* refer to the requirements which must be met in order to implement the program (Muñiz, 1997), and (b) the *modulator contextual variables* are those characteristics of a physical, social or any other nature that may influence the implementation of the program and the results obtained (Kegler, Rigler, & Honeycutt, 2011; Pereira & Coelho, 2013).

Finally, as regards *time* (Shadish, Chacón, & Sánchez-Meca, 2005): (a) the *number of measures (≤1, ≥2)* distinguishes between measurements taken on just one occasion versus more than one, and (b) *measurement points* indicates whether all the data were gathered after the intervention, or whether some were also collected before or during its implementation.

Based on the literature on methodological quality in primary studies in systematic reviews, as well as on data quality in ethnographic research, in Table 1, we define the minimum design elements that must be considered in relation to the levels of intervention (high, medium or low), taking into account the similarities and differences between them, applying the same validity framework from the perspective of design mutability. It can be seen that, on the basis of the same design elements (Dziak, Nahum-Shani, & Collins, 2012), the various methodological restrictions lessen as the level of intervention decreases.

Conclusions and implications for the practice of evaluation

Program evaluation should be regarded as a unitary and integrated process because it is to the benefit of users that a program is evaluated with the most suitable methodology, rather than being bound to certain procedural modalities, and because, in many programs, it can be useful to use different methodologies in complementary fashion or to alter the design so as to shift from one methodology to another in response to the changing reality of users or, on occasions, of the context across the period of implementation. We acknowledge that there are practical difficulties in achieving this kind of integration, especially in terms of what it may mean to abandon an orthodox methodological position. The guiding principle behind this proposed integration is fundamentally pragmatic. When it comes to evaluating an intervention program and considering the extent to which its results (whether positive or not) may be extrapolated, the key question does not concern the distinction between one kind of methodology and another, or which analytic technique is currently most favored for evaluation purposes. Rather, the question is a direct and common-sense one, although this does not make it easy to answer. We need to ask ourselves, what is the most *suitable* procedure, not only for the purposes of evaluation but also in terms of being able to extrapolate the results obtained to other settings, which will vary in the extent to which they resemble the study context.

In light of the above there are a number of general points that can be made regarding the practice of evaluation: (a) base evaluations on a framework of validity rather than on the use of certain methodologies; (b) allow for flexibility ('mutability') of design dimensions so as to adapt them to the intervention context; (c) when the conditions are right for causal explanation, promote the use of experimental methodology, as this favors the unbiased estimation of effect sizes (which in turn will enable subsequent systematic reviews and meta-analyses); and (d) high-quality information, even if it is only descriptive (e.g., observation or survey), is better than inferential information of poor quality obtained via experimental methodology.

In summary, and with respect to the structural dimensions of program evaluation designs, it can be stated that, regardless of the methodology that is ultimately chosen, all professionals should bear the following aspects in mind in order to improve the methodological quality of their intervention: (a) the sample characteristics, such as the *selection criteria*, should be described in detail; (b) the *group/condition assignment criteria* must be clearly set out; ideally a randomization procedure will be used so as to obtain an unbiased estimate of effect size, but if this is not possible efforts should be made to create comparable groups/conditions, applying different control techniques such as pre-assignment matching or the use of cohort groups; (c) the *levels of intervention* will vary according to the objectives set, while the *feasibility* of their application will depend on *contextual variables*; (d) as regards the *number of measures,* it is advisable to take as many as possible, both before and after the intervention, while seeking to ensure that the *data* obtained are of the highest possible *quality* and are gathered using standardized *instruments*, wherever possible, or with the maximum guarantees of rigor, and with the possibility of recording non-equivalent dependent variables; and (e) with respect to *measurement points* it is useful, in addition to having post-test measures, to take at least one pre-test measure and others during the intervention; an alternative to pre-test measures would be measures of independent samples, retrospective measures, or measures that approximate the effect variable.

Table 1
Structural design dimensions in program evaluation with respect to the different levels of intervention

	UNITS –users- (U)
Selection criteria	*R-E/Q-E: Randomization/Known and unknown criterion*.
	N-E: Known and unknown criterion. Attempts have been made to solve the problem of the lack of randomization. Conflict may arise in relation to ethical principles (Anguera, Chacón, & Sanduvete, 2008).
Assignment criteria	*R-E/Q-E: Randomization/Known and unknown criterion*. If the assignment rule is known (e.g., regression discontinuity design), this provides more elements for analyzing what proportion of the observed variability in the results may be due to the use of this rule.
	N-E: Known and unknown criterion. Usually, the users come from groups in which certain needs have been identified.
Comparison groups/ conditions	*R-E/Q-E: Groups/persons/conditions*. The study of group assignment rules is related to the need to create groups that are similar to one another, such that, potentially, they only differ in terms of whether or not they have received the program.
	N-E: Units/Plurality of units (persons and behaviors). The design will be either idiographic or nomothetic depending on the program users and the units they form, as well as on the number of response levels the practitioners are interested in. This poses important methodological questions (Sánchez-Algarra & Anguera, 2013) in relation to whether users are considered individually or as a group (or a representative sample of a certain population), and whether one or several behaviors are of interest.
	TREATMENT -program actions- (T)
Level of intervention	*R-E/Q-E: High/Medium*.
	N-E: Low. The potential object of evaluation will be determined by three requirements: it must be perceivable, it must form part of an individual's everyday life, and it must exist in an interactive relationship with the environment.
Changes in the level of intervention	*R-E/Q-E:* What was initially a change in everyday routines may become a normal activity within the range of activities performed by program users (Anguera, 2001).
	N-E: Programs may experience changes to the degree or level of intervention, which will depend on how flexible the design is and on the rate of its implementation (Chacón, Anguera, & Sánchez-Meca, 2008).
	OUTCOMES -results/instruments- (O)
Type of data	*R-E/Q-E: Scale, ordinal*.
	N-E: Nominal (categorical), ordinal. The data derived from category systems and field formats are categorical; although they are usually referred to a single dimension they can also consider a multidimensional system. Ordinal data are also obtained in some less common situations and the scaling methods are used (Sanduvete et al., 2009).
Data quality	*R-E/Q-E: Assumed as high; related to decreasing standardization of the instruments*. In the case of standardized instruments the procedure for their application and subsequent interpretation is already established. Hence it is usually assumed that there is no need to analyze data quality in itself, as the instrument used is supposed to be reliable and valid (Anguera, Chacón, Holgado, et al., 2008). However, with semi-standardized instruments, which are usually ad hoc, more attention is paid to address a number of issues regarding the quality of the instrument, principally its reliability, validity and measurement error corrections (Chacón, Pérez, & Holgado, 2000).
	N-E: High control over data quality (Blanco, Sastre, & Escolano, 2010). Inter- and intra-observer agreement.
Justification of instruments	*R-E/Q-E: Standardized/Semistandardized*. Standardized evaluation instruments are used in those cases where there is a precise and systematic procedure for collecting data.
	N-E: Non-standardized instruments. The inherent characteristics of a natural and/or habitual context, together with the diffuse nature of many aspects of the program, are what make standardized or semi-standardized instruments unsuitable in most such cases.
Types of instruments	*R-E/Q-E: Standardized tests and psychological measures/Semistandardized questionnaires*. An instrument is said to be standardized when both its use (instructions to subjects, elements of the instrument, the order of test items, etc.) and the criteria for scoring it and its type of reference norm are predetermined (Visser, Ruiter, van der Meulen, Ruijssenaars, & Timmerman, 2012; Zamarrón, Tárraga, & Fernández-Ballesteros, 2008). A semi-standardized instrument (Afonso & Bueno, 2010) is one that has all the above-mentioned features except those related to norms or potential general uses of the instrument.
	N-E: Non-standardized instruments. On occasions, archive material is the only way of accessing vital information, whereas some situations oblige the practitioner to develop ad hoc observation instruments: category system and field format.
Changes in instruments	*R-E/Q-E:* Minimal/higher, with important repercussions in terms of program implementation (Chacón, Anguera, et al., 2008).
	N-E: It is often the case that the design will have changed at some point as a function of aspects related to the design schedule (total duration planned, influence of contextual factors, etc.). This design mutability has important implications in terms of the actual implementation of a program (Chacón, Anguera, et al., 2008). In many cases it will be necessary to modify the instrument used, although not always the level of intervention.
	SETTING -implementation context- (S)
Aspects related to feasibility	*R-E/Q-E: Severe/intermediate restrictions over the use of the program*. The higher the level of intervention the greater will be the potential restrictions regarding program implementation. In general, the way in which the program is implemented is of fundamental importance in relation to drawing causal inferences (Wecker, 2013). In Q-E, the following designs may be used, in descending order of preference: switching replications design, switching the treatment and control group, repeated treatment (ABAB), reversed treatment and removed treatment designs, and designs that use a 'dose' of exposure to treatment (Chacón, Shadish, & Cook, 2008).
	N-E: Minimal restrictions over the use of the program. There are not serious problems of applicability, due essentially to the 'weak' nature of the intervention.
Modulator contextual variables	*R-E/Q-E:* They may be controlled by means of techniques such as masking, blocking, matching or stratification, and their possible influence may be studied a posteriori, comparing the results from different groups (Shadish et al., 2002).
	N-E: The treatment or program actions do need to be described in detail (to avoid too much discretion on the part of the evaluator, which could undermine the 'naturalness' of the planned intervention), and the proposed scope of the program must also be clearly set out, i.e. in relation to the geographical area in which it will be applied, the schedule and the characteristics of users (Anguera, 2008).
	TIME (Ti)
Number of measures (≤1, ≥2) and measurement points	*R-E/Q-E: Before, during and after the program*. As a general rule the more measurements that are taken and the more measurement points used, the greater the possibility of analyzing different aspects of the program under evaluation. Obviously, this assertion assumes that the measures used have, as far as possible, evidence of validity (Chacón & Shadish, 2008), there being the possibility of recording non-equivalent dependent variables. In addition to pre- and post-test measures it is advisable to collect information during program implementation, as this will enable improvements to be made as required and not only after completion of the program. This is known as formative -as opposed to summative or final- evaluation (Anguera, Chacón, & Sánchez, 2008). An alternative to pre-test measures would be measures of independent samples, retrospective measures, or measures that approximate the effect variable.
	N-E: Any intervention program is structured according to a system of inter-related factors that act in one way or another as a function of time (Hernández-Mendo & Anguera, 2001). It is less common for interest to lie in conducting a program evaluation at a single point in time as such a snapshot does not capture the dynamic nature of the process, although an evaluation of this kind may, however, be useful as information gathered at certain points of an ongoing intervention.

Note: R-E = randomized experiment (high-level intervention); Q-E = quasi-experiment (medium-level intervention); N-E = non-experimental/ethnographic study (low-level intervention)

Acknowledgements

This study forms part of the results obtained in research projects PSI2011-29587 funded by the Spanish Ministry of Science and Innovation, and DEP2012-32124 funded by the Spanish Ministry of Economy and Competitiveness. Also, it forms part of Research Group from Generalitat de Catalunya (2009-SGR-829).

References

Afonso, R., & Bueno, B. (2010). Reminiscencia con distintos tipos de recuerdos autobiográficos: efectos sobre la reducción de la sintomatología depresiva en la vejez [Reminiscence with different types of autobiographical memories: Effects on the reduction of depressive symptomatology in old age]. *Psicothema, 22*(2), 213-220.

Anguera, M.T. (2001). Hacia una evaluación de la actividad cotidiana y su contexto: ¿Presente o futuro para la metodología? Discurso de ingreso como académica numeraria electa [Towards an evaluation of the daily activity and its context: Is it present or future for methodology? Talk to join as long-standing elected academician]. Reial Acadèmia de Doctors, Barcelona (1999). In A. Bazán & A. Arce (Eds.), *Estrategias de evaluación y medición del comportamiento en psicología* (pp. 11-86). México: Instituto Tecnológico de Sonora y Universidad Autónoma de Yucatán.

Anguera, M.T. (2008). Diseños evaluativos de baja intervención [Low-level evaluative designs]. In M.T. Anguera, S. Chacón & A. Blanco (Eds.), *Evaluación de programas sociales y sanitarios: un abordaje metodológico* (pp. 153-184). Madrid: Síntesis.

Anguera, M.T., & Chacón, S. (1999). Dimensiones estructurales de diseño para la evaluación de programas [Structural dimensions of design for program evaluation]**.** *Apuntes de Psicología, 17*(3), 175-192.

Anguera, M.T., Chacón, S., Holgado, F.P., & Pérez, J.A. (2008). Instrumentos en evaluación de programas [Instruments in program evaluation]. In M.T. Anguera, S. Chacón & A. Blanco (Eds.), *Evaluación de programas sociales y sanitarios: un abordaje metodológico* (pp. 127-152). Madrid: Síntesis.

Anguera, M.T., Chacón, S., & Sánchez, M. (2008). Bases metodológicas en evaluación de programas [Methodological groundings in program evaluation]. In M.T. Anguera, S. Chacón & A. Blanco (Eds.), *Evaluación de programas sociales y sanitarios: un abordaje metodológico* (pp. 37-68). Madrid: Síntesis.

Anguera, M.T., Chacón, S., & Sanduvete, S. (2008). Cuestiones éticas en evaluación de programas [Ethical issues in program evaluation]. In M.T. Anguera, S. Chacón & A. Blanco (Eds.), *Evaluación de programas sociales y sanitarios: un abordaje metodológico* (pp. 291-318). Madrid: Síntesis.

Auweiler, P.W.P., Müller, D., Stock, S., & Gerber, A. (2012). Cost effectiveness of Rituximab for Non-Hodgkin's Lymphoma: A Systematic Review. *Pharmacoeconomics, 30*(7), 537-549.

Blanco, A., Sastre, S., & Escolano, E. (2010). Desarrollo ejecutivo temprano y Teoría de la Generalizabilidad: bebés típicos y prematuros [Executive function in early childhood and Generalizability Theory: Typical babies and preterm babies]. *Psicothema, 22*(2), 221-226.

Campbell, D.T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*(4), 297-312.

Campbell, D.T., & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: RandMcnally.

Chacón, S., Anguera, M.T., & Sánchez-Meca, J. (2008). Generalización de resultados en evaluación de programas [Generalization of results in program evaluation]. In M.T. Anguera, S. Chacón & A. Blanco (Coords.), *Evaluación de programas sociales y sanitarios. Un abordaje metodológico* (pp. 241-258). Madrid: Síntesis.

Chacón, S., Pérez, J.A., & Holgado, F.P. (2000). Validez en evaluación de programas: una comparación de técnicas de análisis basadas en modelos estructurales [Validity in program evaluation: A comparative approach based on structural models]. *Psicothema, 12*(2), 122-126.

Chacón, S., Sanduvete, S., Portell, M., & Anguera, M.T. (2013). Reporting a program evaluation: Needs, program plan, intervention, and decisions. *International Journal of Clinical and Health Psychology, 13*(1), 58-66.

Chacón, S., & Shadish, W.R. (2008). Validez en evaluación de programas [Validity in program evaluation]. In M.T. Anguera, S. Chacón & A. Blanco (Coords.), *Evaluación de programas sociales y sanitarios. Un abordaje metodológico* (pp. 69-102). Madrid: Síntesis.

Chacón, S., Shadish, W.R., & Cook, T.D. (2008). Diseños evaluativos de intervención media [Evaluative designs of medium intervention]. In M.T. Anguera, S. Chacón & A. Blanco (Coords.), *Evaluación de programas sociales y sanitarios. Un abordaje metodológico* (pp. 185-218). Madrid: Síntesis.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Cook, T.D., Scriven, M., Coryn, C.L.S., & Evergreen, S.D.H. (2010). Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven. *American Journal of Evaluation, 31*(1), 105-117.

Cronbach, L. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Dziak, J.J., Nahum-Shani, I.N., & Collins, L.M. (2012). Multilevel factorial experiments for developing behavioral interventions: Power, sample size, and resource considerations. *Psychological Methods, 17*(2), 153-175.

Hernández-Mendo, A., & Anguera, M.T. (2001). Análisis psicosocial de los programas de actividad física: evaluación de la temporalidad [Psychosocial analysis of physical activity programs: Evaluation of temporality]. *Psicothema, 13*(2), 263-270.

Kegler, M.C., Rigler, J., & Honeycutt, S. (2011). The role of community context in planning and implementing community-based health promotion projects. *Evaluation and Program Planning, 34*(3), 246-253.

Kundin, D.M. (2010). A conceptual framework for how evaluators make everyday practice decisions. *American Journal of Evaluation, 31*(3), 347-362.

Lord, D., & Kuo, P-F. (2012). Examining the effects of site selection criteria for evaluating the effectiveness of traffic safety countermeasures. *Accident Analysis and Prevention, 47*(July), 52-63.

Matisoff, M., & Barksdale, L. (2012). Mathematical & statistical analysis of bloodstain pattern evidence (part II). *Forensic Examiner, 21*(Summer), 22-32.

Muñiz, J. (1997). Aspectos éticos y deontológicos de la evaluación psicológica [Ethical and deontological issues in psychological evaluation]. In A. Cordero (Coord.), *La evaluación psicológica en el año 2000* (pp. 307-345). Madrid: TEA Ediciones.

Pereira, M.C., & Coelho, F. (2013). Work hours and well being: An investigation of moderator effects. *Social Indicators Research, 111*, 235-253.

Sánchez-Algarra, P., & Anguera, M.T. (2013). Qualitative/quantitative integration in the inductive observational study of interactive behaviour: Impact of recording and coding among predominating perspectives. *Quality and Quantity, 47*(2), 1237-1257.

Sanduvete, S., Barbero, M.I., Chacón, S., Pérez, J.A., Holgado, F.P., Sánchez, M., & Lozano, J.A. (2009). Métodos de escalamiento aplicados a la priorización de necesidades de formación en organizaciones [Scaling methods applied to set priorities in training programs in organizations]. *Psicothema, 21*(4), 509-514.

Shadish, W.R. (1995). The logic of generalization: Five common principles to experiments and ethnographies. *American Journal of Community Psychology, 23*(3), 419-428.

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), 95-110.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

Visser, L., Ruiter, S.A.J., van der Meulen, B.F., Ruijssenaars, W.A.J.J.M., & Timmerman, M.E. (2012). A review of standardized developmental assessment instruments for young children and their applicability for children with special needs. *Journal of Cognitive Education and Psychology, 11*(2), 102-127.

Wecker, C. (2013). How to support prescriptive statements by empirical research: Some missing parts. *Educational Psychology Review, 25*, 1-18.

Zamarrón, M.D., Tárraga, L., & Fernández-Ballesteros, R. (2008). Plasticidad cognitiva en personas con la enfermedad de Alzheimer que reciben programas de estimulación cognitiva. *Psicothema, 20*(3), 432-437.