



International Journal of Clinical and Health Psychology

www.elsevier.es/ijchp



THEORETICAL ARTICLE

Reporting a program evaluation: Needs, program plan, intervention, and decisions

Salvador Chacón Moscoso^{a,*}, Susana Sanduvete Chaves^a, Mariona Portell Vidal^b,
M. Teresa Anguera Argilaga^c

^aUniversidad de Sevilla, Spain

^bUniversidad Autónoma de Barcelona, Spain

^cUniversidad de Barcelona, Spain

Received September 29, 2012; accepted October 17, 2012

KEYWORDS

Report;
Program evaluation;
Needs;
Design;
Theoretical study

Abstract The approach to intervention programs varies depending on the methodological perspective adopted. This means that health professionals lack clear guidelines regarding how best to proceed, and it hinders the accumulation of knowledge. The aim of this paper is to set out the essential and common aspects that should be included in any program evaluation report, thereby providing a useful guide for the professional regardless of the procedural approach used. Furthermore, the paper seeks to integrate the different methodologies and illustrate their complementarity, this being a key aspect in terms of real intervention contexts, which are constantly changing. The aspects to be included are presented in relation to the main stages of the evaluation process: needs, objectives and design (prior to the intervention), implementation (during the intervention), and outcomes (after the intervention). For each of these stages the paper describes the elements on which decisions should be based, highlighting the role of empirical evidence gathered through the application of instruments to defined samples and according to a given procedure.

© 2012 Asociación Española de Psicología Conductual. Published by Elsevier España, S.L.

All rights reserved.

PALABRAS CLAVE

Informe;
Evaluación
de programas;
Necesidades;

Resumen Se encuentran diferentes posturas respecto a programas de intervención en función de la perspectiva metodológica adoptada, por lo que el profesional de la salud no dispone de unas directrices claras de actuación, dificultándose la acumulación del conocimiento. El objetivo propuesto es concretar los aspectos básicos/mínimos y comunes a explicitar en el informe de evaluación de cualquier programa, útil para el profesional, independiente-

*Corresponding author at: Facultad de Psicología, Universidad de Sevilla, Campus Ramón y Cajal, C/ Camilo José Cela, s/n, 41018 Sevilla, Spain.

E-mail address: schacon@us.es (S. Chacón Moscoso).

Diseño;
Estudio teórico

mente de la opción procedimental que se elija, fomentando de este modo la integración y la complementariedad entre metodologías, como respuesta a las circunstancias reales del contexto de intervención en cambio continuo. Estos aspectos se encuadran en las principales fases de evaluación: necesidades, objetivos y diseño (antes de la intervención), implementación (durante esta) y resultados (después de la intervención). En cada una de ellas, se explicita en qué elementos basar la toma de decisiones a partir de evidencias empíricas registradas mediante instrumentos en unas muestras siguiendo un procedimiento determinado.

© 2012 Asociación Española de Psicología Conductual. Publicado por Elsevier España, S.L.
Todos los derechos reservados.

The approach to intervention programs varies depending on the methodological perspective adopted (Anguera, 2003; Wallraven, 2011), and this means that the health professional lacks clear guidelines regarding which designs and implementation and evaluation procedures to use in what are constantly changing intervention contexts. This situation hampers the integrated accumulation of knowledge. The aim of this paper is to set out the essential and common aspects that should be included in any program evaluation report, thereby providing a useful guide for professionals regardless of the procedural approach used (*i.e.*, qualitative or quantitative, experimental, quasi-experimental, or observational).

The premise of the paper is that the design and evaluation facets are in constant interaction with one another, and the conclusions drawn are set out below in operational terms. It is argued that from the initial needs assessment through to the final evaluation of outcomes there is a continuum of decision making that must be based on empirical evidence gathered by means of scientific methodology (Anguera & Chacón, 2008), regardless of the specific methodological or procedural approach that is chosen at any given stage of the process. By setting out the decision-making criteria for the whole intervention process, the aim is to provide professionals with a useful resource based on common principles, thereby fostering the integration of scientific knowledge in the health context.

At present, the wide variety of interventions and the different ways in which they are communicated (*e.g.*, Bornas et al., 2010; Gallego, Gerardus, Kooij, & Mees, 2011; Griffin, Guerin, Sharry, & Drumm, 2010) prevent any systematic evaluation and the extrapolation of results (Chacón & Shadish, 2008). Although other authors have previously described the elements to be included in a program evaluation (Cornelius, Perrio, Shakir, & Smith, 2009; Li, Moja, Romero, Sayre, & Grimshaw, 2009; Moher, Liberati, Tetzlaff, & Altman, 2009; Schulz, Altman, & Moher, 2010), it is considered that this approach has a number of advantages in comparison to the extant literature: a) It is integrative and can be adapted to all potential methodologies, rather than being exclusively applicable to evaluations based on experimental and/or quasi-experimental designs; b) it focuses on common methodological aspects throughout the process, thereby enabling it to be extrapolated and generalized to a range

of topics in the clinical and health context, instead of being limited to a specific question or setting; c) the use of a common methodological framework applied to different contexts facilitates comparison between interventions and, therefore, fosters the systematic accumulation of knowledge; d) not only can it be used to analyze those interventions which are already underway or completed, but it may also be useful for the planning, design, implementation, and evaluation of future programs; e) it is easy and quick to apply, as should be clear from the small number of elements it contains in comparison with other approaches; and (f) the language used is not overly specific, thereby enabling it to be applied by professionals without specialist methodological training.

The starting point for this approach is that the aim of program evaluation is to pass judgment on the value of a given program or some of its elements (Anguera & Chacón, 2008). With this in mind, the evaluation process can usefully be divided into three stages, namely the periods prior to, during, and after the intervention (Anguera, Chacón, & Sánchez, 2008). The first of these stages can be further subdivided into the needs assessment, evaluation of objectives, and evaluation of the design, while the second refers to evaluation of the implementation and the third to evaluation of outcomes. Table 1 outlines in chronological order the aspects to be considered in a program evaluation report, with illustrative examples being provided in each section.

These stages constitute a formative and summative evaluation in continuous interaction. As this approach means that the program is subject to continuous evaluation, said program becomes self-regulating and is able to provide useful information to all those involved across the whole intervention process (Chacón, López, & Sanduvete, 2004).

The decisions made throughout the evaluation process (from needs through to outcomes) must be based on empirical evidence gathered by means of appropriate instruments in specific samples and according to a given procedure. Therefore, before moving on to a more detailed description of the basic aspects included in each stage of the evaluation process, the following Method section of this paper presents a general conceptualization of sample, instruments, and procedure as elements common to all of the stages.

Table 1 Outline of a program evaluation report.*Needs assessment*

- Conceptualization: problem to be addressed. Type of need: normative, perceived, expressed, comparative
- Sample: drawn from the target population
- Instruments: e.g., surveys, interviews, discussion groups, life stories
- Procedure: e.g., development of instruments, application, data registry
- Assessment: pertinence, criteria for prioritization of needs, coverage, and operative inclusion/exclusion criteria

Evaluation of objectives

- Conceptualization: general objectives (goals). Specific objectives (identifiable, measurable/operative, and expressed in relation to a timeframe): linked to the general objectives and to specific needs
- Sample: existing theoretical models, previous research
- Instruments: e.g., secondary data, coding systems
- Procedure: e.g., systematic reviews, simulation studies
- Assessment: usability, sufficiency

Evaluation of the design

- Conceptualization: work plan in which the activities, the necessary resources, the timeframe, and the data collection points are linked to specific objectives in accordance with needs (types of design; design components and validity)
- Sample: existing similar programs
- Instruments: e.g., secondary data, coding systems
- Procedure: e.g., systematic reviews, simulation studies
- Assessment: internal coherence, feasibility, acceptability, viability

Evaluation of the implementation

- Conceptualization: execution of what has been planned
- Sample: users from whom the relevant variables will be recorded
- Instruments: e.g., interviews, questionnaires, observational records, discussion groups, standardized data
- Procedure: e.g., development of instruments, application, data analysis
- Assessment: quality or fit, resolving discrepancies, evaluability

Evaluation of outcomes

- Conceptualization: degree to which objectives have been met
- Sample: participants from whom data are collected
- Instruments: similar to those used in the needs assessment stage; e.g., surveys, interviews, discussion groups
- Procedure: e.g., development of instruments, application, data analysis
- Assessment: efficacy, effectiveness, efficiency, continuity, progress, suitability, utility, probity, equity

Method

Sample

In general terms the sample comprises those users from whom data will be gathered in order to draw conclusions in the evaluation (Anguera et al., 2008b). Thus, in terms of needs assessment, evaluation of the implementation, and evaluation of outcomes the sample would be formed by those persons who are studied in order to determine, respectively, the gaps which the program seeks to fill or address, the extent to which the intervention is being carried out according to plan, and the effects of the program. A clear description of the characteristics of users (the sample) is essential for enabling the possible generalization of outcomes (Chacón, Anguera, & Sánchez-Meca, 2008a). Note, however, that the sample may not always refer to people, but could comprise study units, such that the sample in the evaluation of objectives and evaluation of the design stages would be the theoretical models and similar programs that are studied and from which the researchers draw possible objectives and design

elements which are applicable to the program being planned.

Instruments

Instruments are methodological resources that enable researchers to gather the empirical information needed to carry out the evaluation (Anguera, Chacón, Holgado, & Pérez, 2008). Several data collection techniques are available, ranging from the gathering of secondary data, which requires a minimum interaction with program users (e.g., review of archive data or observation, preferably based on recordings made after obtaining consent from users and/or managers), to those which imply their active involvement (e.g., interviews and discussion groups). Between these two extremes there are other approaches that imply a moderate level of interaction with users and which consist in eliciting a response from them (e.g., administration of scales, tests, questionnaires, and other standardized systems for recording data). This criterion should be considered in combination with the level of standardization shown by the instrument (Anguera, Chacón, Holgado et al., 2008).

The instruments most widely used for gathering information at the needs assessment stage are surveys of potential users, interviews with those involved and with experts in the field, discussion groups, and life stories. When it comes to evaluating objectives and the design, the emphasis is on consulting secondary data and coding systems. At the implementation stage the instruments most commonly used are interviews and questionnaires administered to users, observational records, discussion groups, and standardized data registers such as indicators, index cards, checklists, or self-reports. Finally, at the outcomes stage almost all of the abovementioned instruments may be employed, with special interest being attached to the use of the same or similar instruments to those chosen at the needs assessment stage so as to facilitate the comparison of pre- and post-intervention data (e.g., Casares et al., 2011).

Procedure

The procedure is the way of carrying out the actions required for data recording and analysis, for example, how the measurement instrument is drawn up, how it is applied, the points at which measurements are taken, and how the data are analyzed in accordance with the objectives, the nature of the data, and the size of the samples (both users and data recording points). The steps to be followed should be clearly described so as to enable the procedure to be replicated. Although the procedure will vary considerably depending on the methodology chosen (Anguera, 2003; Wallraven, 2011), this protocol adopts a position of methodological complementarity rather than one of contrast (Chacón, Anguera, & Sánchez-Meca,).

Stages of evaluation

Having defined the method in terms of the aspects common to all the stages of evaluation (sample, instruments, and procedure), it will be now provided a more detailed description of the other basic and specific aspects to be considered in relation to the conceptualization (stated aim) and assessment (criteria for decision-making to value the evidence gathered) of each of the stages.

Needs assessment

- **Conceptualization.** The problem to be addressed. The first stage of program evaluation, that is, needs assessment, seeks to analyze the key features of the problem at hand. This is a systematic procedure involving the identification of needs, the setting of priorities, and decision making regarding the kind of program required to meet these needs (Chacón, Lara, & Pérez, 2002; Sanduvete et al., 2009). In general, a need is considered to be a discrepancy between the current status of a group of people and what would be desirable (Altschuld & Kumar, 2010). The needs assessment consists of an analysis of these discrepancies and a subsequent prioritization of future actions, such that the first needs to be addressed are those of greatest concern or which are regarded as most fundamental or urgent.

- **Type of need.** Depending on the source of information, different types of need can be distinguished (Anguera, Chacón, & Sánchez, 2008):

- a) **Normative:** this is a need defined by an expert or professional in a specific context and is based on standard criteria or norms. The specialist establishes a “desirable” level and compares it with the current situation; a need is said to be present when an individual or group do not reach this desired level. A normative definition of need is not absolute, and were other dimensions to be considered they would not necessarily yield the same normative need. The point in time is also important: for example, discrepancies would inevitably arise when comparing hospital building based on the number of beds per million inhabitants now and twenty years ago.
- b) **Experienced, felt, or perceived:** in this case, need corresponds to a subjective sense of something lacking, and such information will be obtained by eliciting some kind of response from the subject (for example, by means of a questionnaire item). When evaluating a specific need within a service the individuals involved will be asked whether they believe the need is relevant to them, in other words, whether they perceive it as a need. A felt or perceived need is, on its own, an inadequate measure of actual need because it fluctuates and may be modulated by the individual’s current situation, which can involve both aspects that are lacking and those which are surplus to requirements.
- c) **Expressed or requested:** this refers to the explicit request for a given service. It differs from a perceived need in that here the person expresses the need under his/her own initiative, without having to be asked or given the opportunity to do so.
- d) **Comparative:** this kind of need is established according to differences between services in one area and those in another. Note, however, that such differences do not necessarily imply a need, as it may be the case that a given service is not needed in one area. Furthermore, it can be difficult to establish whether all the needs in the reference area are actually met. This kind of need has frequently been used to draw up lists of characteristics of patients who require special care (Milsom, Jones, Kearney-Mitchell, & Tickle, 2009).

- **Assessment.** Based on the evidence gathered from the sample (recruited from the target population) using defined instruments and procedures, assessments can then be drawn regarding four key aspects: a) The pertinence of the program: is it really necessary, are there really needs to be met (Anguera & Chacón, 2008; Fernández-Ballesteros, 1995; World Health Organization [WHO], 1981); b) The criteria for prioritization of needs: this implies dealing with the most urgent needs first, these being identified according to a set of previously defined criteria, for example, the consequences of not addressing these needs or the number of people affected; c) The coverage of the program: the identification of users and potential users (Anguera & Chacón, 2008); and d) The operative criteria for program inclusion/exclusion: it is necessary to

clearly set out the reasons why some potential beneficiaries will be included in the program while others will not.

Evaluation of objectives

- **Conceptualization.** The objectives cover all those aspects which the program seeks to address in terms of the identified need, in other words, the outcomes which one seeks to achieve. The general objectives or goals should be set out in terms of a series of specific objectives that are clearly defined, identifiable, measurable/operative, and related to a timeframe (Anguera, Chacón, & Sánchez, 2008).
- **Assessment.** Two basic aspects should be assessed in this stage, usually in relation to existing theoretical models and similar programs: a) The usability of the program (Anguera & Chacón, 2008), that is, the extent to which the pre-established program, set within a theoretical framework, can be used to meet the identified needs. The aim here is to explain why certain specific objectives are set and not others. This requires a theoretical justification based on a literature search (e.g., a systematic review) in order to identify the objectives of similar programs and to study the potential problems which may arise, examining their causes and possible ways of resolving them. It is also advisable to carry out simulation studies so as to estimate possible outcomes prior to program implementation. The second aspect b) relates to the intervention context and involves taking decisions about the program's sufficiency: do its objectives address all the needs which have been prioritized (Anguera & Chacón, 2008; Fernández-Ballesteros, 1995; WHO, 1981), and are these objectives likely to be sufficient in order to meet, fully or otherwise, these needs.

Evaluation of the design

- **Conceptualization.** The design is the work plan that will subsequently be implemented. It must be consistent with the proposed objectives and should be set out in such a way that it could be put into practice by a professional other than the person who has drawn it up. It is helpful if the design has a degree of flexibility, as when it is implemented certain problems may arise that require a quick response and modification of the original plan. The content of the design is usually based on existing similar programs that will have been identified through the literature search (e.g., systematic reviews). As noted above, it is also advisable to carry out simulation studies before implementing the program so as to detect possible ways of improving the design. At all events the design should clearly describe: a) The activities that will be carried out in order to meet each of the stated objectives in accordance with the identified needs; b) the resources required (human, financial, material, etc.); c) a specific timeframe, which may be based on a PERT (Project Evaluation Review Technique) or CPM (Critical Path Method), and which should clearly set out when the different activities will take place in relation to one another (their diachronic and synchronic relations) (Anguera, Chacón, & Sánchez, 2008).; d) the data recording points during the evaluation; and e) any other

aspect of the intervention which may be of interest. Different types of design are possible within this broad framework, but attention must be paid to the minimum features that are required in order to strengthen the validity of the evaluation.

- **Types of design.** This paper takes an integrative approach to the types of design used in program evaluation, classifying them according to the degree of intervention or control they exert over the evaluation context (Chacón & Shadish, 2001). This criterion yields a continuum, at one end of which lie experimental designs, those with the maximum level of intervention, while at the end of minimum intervention are situated open-ended designs mainly based on qualitative methodology. Along this continuum it is possible to place three, structurally distinct families of design: experimental, quasi-experimental, and observational* listed here in descending order regarding the degree of manipulation and control they imply. Experimental designs (high level of intervention) are those in which a program is deliberately implemented in order to study its effects, and where the sample of users is randomly assigned to different conditions (Peterson & Kim, 2011; Shadish, Cook, & Campbell, 2002). Quasi-experimental designs (moderate level of intervention) differ from experimental ones mainly in that the sample is not randomly assigned to different conditions (Shadish et al., 2002). Finally, observational designs (low level of intervention) quantify the occurrence of behavior without there being any kind of intervention (Anguera, 2008; Anguera & Izquierdo, 2006).

From a purely methodological point of view, and when the object of study is causal, it is considered better to use experimental designs because, under optimal conditions of application, they allow the researcher to obtain an unbiased estimate of the effect size associated with the intervention (Chacón & López, 2008).

As regards the measurement points (Chacón & López, 2008), and within the context of experimental designs, a further distinction is made between cross-sectional designs (where the variables are measured at a specific point in time), longitudinal designs (where the variables for one group are measured at more than one point in time), and mixed (in which at least one variable is measured cross-sectionally and another longitudinally). Depending on the number of independent variables that are measured (Chacón & López, 2008), designs may be termed single-factor (when there is only one independent variable) or factorial (more than one independent variable). In the former a distinction is made between the use of just two values (two conditions), where only the difference between them can be studied, and the presence of more than two values (multiple conditions), where it is possible to analyze in greater depth the relationship between the independent variable and the dependent variable(s).

*The term "observational design" refers to general action plans whose purpose is the systematic recording and quantification of behavior as it occurs in natural or quasi-natural settings, and where the collection, optimization, and analysis of the data thus obtained is underpinned by observational methodology (Anguera, 1979, 1996, 2003).

Of particular relevance in these designs is the use of techniques for controlling extraneous variables, the aim being to neutralize the possible effects of factors unrelated to the object of evaluation and which might confound the measurement of its effect (Chacón & López, 2008).

It should be noted that it is not always possible to assign participants randomly to different groups, perhaps because these groups already exist naturally, or due to a lack of resources, or for ethical reasons, etc. (Anguera, Chacón, & Sanduvete, 2008). In situations such as these, researchers can turn to quasi-experimental designs, the main types of which are described below.

Pre-experimental designs (Chacón, Shadish, & Cook, 2008) are not, strictly speaking, quasi-experimental designs, but they are the origin of the latter. In a one-group posttest-only design the outcomes of a group of users are measured after the intervention; there is no control group and measures are not taken prior to the intervention. By adding a pre-intervention measure one arrives at the one-group pretest-posttest design, whereas including the measurement of an additional control group after the intervention would yield a posttest-only design with nonequivalent groups. It is difficult to infer cause-effect relationships from these designs, as the results obtained could be due to a large number of potential threats to validity.

If it is added to the one-group posttest-only design not only a control group but also a pre-intervention measure, a truly quasi-experimental design is obtained (Chacón et al., 2008; Shadish et al., 2002): the pretest/posttest design with non-equivalent control group. Although this approach avoids numerous threats to validity it may still be affected by others such as maturation. The designs described below address this problem.

The basic cohort control group design consists in comparing measures taken from special groups known as cohorts, which may be defined as groups of users (of a program or of a formal or informal organization) that are present over time within the different levels of a given organization. This regular turnover of users usually produces similar groups, although obviously they are available at different points in time. Certain cohorts may sometimes be the subject of a specific intervention that, for whatever reason, was not applied to other past or future cohorts. The logic behind this design is the same as that which guides the pretest/posttest design with non-equivalent control group, the added advantage being that the cohort groups are quasi-comparable.

The regression discontinuity design consists in making explicit the inclusion criterion for one group or another, this being known as the cut-off point. This specific value of the dependent variable serves as a reference, such that all those individuals above the cut-off will take part in the program. This design is regarded as the best among quasi-experimental approaches, as the fact that the criterion for group assignment is known means that it shares a characteristic with experimental designs. In this regard, it has been shown that under optimum conditions of application this design is able to offer unbiased estimates of a program's effects that are equivalent to those which can be obtained with a randomized experiment (Trochim, 2002). In order to implement this design a number of

conditions must be fulfilled, and its use is therefore restricted to situations when: a) there are linear relationships between the variables; and b) it can be assumed that the regression line would extend beyond the cut-off point were the intervention not to be implemented. Additional problems are that it is difficult to generalize from one situation to another, due to the specificity of each design of this kind, and also that it can be difficult to establish an appropriate cut-off point.

The interrupted time series design (Shadish et al., 2002) involves taking successive measures over time, whether from the same users observed at different points or from different users who are considered similar. It is also necessary to know the exact point at which the program is implemented so as to be able to study its impact or effect on the data series (Anguera, 1995). The effects of the program can be appreciated in the differences between the series obtained before and after its implementation with respect to: a) a change of level: discontinuity at the point of program implementation, even though the slope of both series is the same; or b) a change in slope. The observed effects may be maintained or disappear across a series, while on other occasions their appearance may lag behind the introduction of the program. By taking multiple measurements over time, both before and after a program is implemented, this type of design improves upon those approaches which rely on a single measure taken before and/or after an intervention.

The third broad category of design corresponds to observational designs, which involve a low level of intervention (Anguera, 2008). These designs are a response to the highly complex nature of social reality, in which the individuals or groups that take part in programs do not form a compact reality, the dynamics of the processes followed are not uniform, there are serious doubts regarding supposed causal relationships, and it is sometimes extremely difficult to collect data that meet all the required standards of rigor (Anguera, 2008).

Traditionally, attention was focused exclusively on those interventions in which there was some control over the situation to be evaluated, with program users being given instructions to ensure that the program was implemented according to the researchers' plan. However, an increasing number of programs are now implemented without clearly specified instructions and in the natural and/or habitual context of program users, taking advantage of their spontaneous and/or everyday activities. Evaluation designs that involve a low level of intervention are well suited to these situations, and can respond to both the restrictions and the opportunities they imply. The priority in these designs shifts from the control of potential confounders (a key feature of experimental and quasi-experimental research) towards the minimization of bias due to reactivity and contextual representativeness. In this case, causal inference can be based on the analysis of behavioral sequences.

Observational designs are basically classified according to three criteria (Anguera, 2008): a) Depending on the number of units considered in the program they may be idiographic (with a single user or group considered as the unit of study, or centered on a single response level) or nomothetic (more than one user or multiple levels of response); b) depending

Table 2 Basic types of observational design.

		Number of data collection points	
		Point (1)	Follow-up (>1)
Number of units	Ideographic (1) Nomothetic (>1)	II. III. Synchronic designs	I. Diachronic designs IV. Mixed/lag-log designs

on when the data are recorded they may be point (a specific point in time is analyzed) or follow-up (more than one time point); and c) depending on the different dimensions or aspects to which the low-level intervention is applied they may be unidimensional (one dimension) or multidimensional (several dimensions). Table 2 summarizes the basic types of design that can be generated by three of the possible combinations of criteria a) and b); quadrant II does not give rise to any design as it lacks the potential to produce data. Each of these designs will become more complex if, in accordance with criterion c), they are multidimensional.

Diachronic designs (quadrant I) can be sub-divided, depending on the type of data recorded, into: a) Extensive (the frequency of an event's occurrence is recorded) and b) Intensive (in addition to frequency the order of appearance and, optionally, the duration of an event is recorded). Depending on the number of measurement points the extensive approach includes panel designs (two time points), trend designs (between three and 49 time points) and time series (50 or more time points). Intensive designs may be sequential (prospective when one analyzes which behaviors follow the event under study, and retrospective when past behavior is studied) or based on polar coordinates (both prospective and retrospective analysis).

Synchronic designs (quadrant III) can be sub-divided, depending on the type of relationship between the events under study, into symmetrical (when one studies association without directionality in the relationship) and asymmetrical (when causality is studied).

Finally, there are 24 types of mixed or lag-log designs (quadrant IV), resulting from the combination of three criteria: a) The type of follow-up (extensive or intensive); b) the combination of the number of response levels (1 or more) and the number of users (1 or more) studied; and c) the type of relationship between units (independent, when there is no relationship between response levels and/or users; dependent, when there is an asymmetrical relationship; interdependent, when the relationship is bidirectional; and hybrid, when the relationship does not match any of the above).

- Components of the design and validity of the evaluation. In order to enhance the validity of an evaluation and, therefore, to enable the outcomes obtained to be generalized, it is necessary to consider at least the following design components: users/units (U), intervention/treatment (T), outcomes (O), setting (S), and time (Ti), referred to by the acronym UTOSTi. On the basis of these components the different kinds of validity could be defined as follows (Chacón & Shadish, 2008; Shadish et al., 2002): a) statistical conclusion validity refers to the validity of inferences about the correlation

between 't' (treatment) and 'o' (outcome) in the sample; b) internal validity refers to the validity of inferences about the extent to which the observed covariation between t (the presumed treatment) and o (the presumed outcome) reflects a causal relationship; c) construct validity refers to the validity of inferences about the higher-order constructs that represent particular samples (generalization of the samples to the population of reference); and (d) external validity refers to the validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, times, and measurement variables (generalization of the samples to populations that are not the reference one, or between different sub-samples).

- Assessment. Here the following aspects need to be assessed (Anguera & Chacón, 2008): a) The degree of internal coherence: the degree of fit between the identified needs and the objectives set out in order to address them, the activities designed to meet these objectives, and the resources available to enable these activities to be put into practice within an adequate timeframe; b) feasibility: the extent to which program implementation is possible in relation to the proposed actions (Fernández-Ballesteros, 1995); c) acceptability: the extent to which each group involved may fulfill its assigned role, without interfering with other groups; and d) viability: the minimum requirements that must be fulfilled in order for the program to be implemented.

Evaluation of the implementation

- Conceptualization. The implementation consists in putting into practice the previously designed program (Anguera, Chacón, & Sánchez, 2008).
- Assessment. Here the following aspects must be assessed: a) Quality or fit: the extent to which the nature and use of each of the program actions is consistent with technical and expert criteria (Anguera & Chacón, 2008). The analysis should examine the degree of fit between what was designed in the previous stage and what has actually been implemented. Any possible discrepancies and their causes will be looked for in each of the program components: activities carried out, the timeframe, the human and material resources used, program users, and measurement points, among others; b) The solution to possible discrepancies: any lack of fit will put to the test the flexibility of the original work plan, as ideally it will be able to be modified and adapted to the new situation; and previously to the evaluation of outcomes; c) The degree of evaluability of the program, that is, the extent to which it can be evaluated according to the various criteria of

relevance (Anguera & Chacón, 2008; Fernández-Ballesteros, 1995).

In this context it is worth distinguishing between monitoring/follow-up and formative evaluation, where the former involves solely the recording of data during the process, whereas the latter includes the possibility of redesigning the program in light of the data obtained in this process.

Evaluation of outcomes

- **Conceptualization.** The outcomes are derived from the analysis of the program's effects (Anguera, Chacón, & Sánchez, 2008). This involves determining empirically the extent to which the stated objectives have been achieved, as well as identifying any unforeseen outcomes and examining the cost-benefit relationships involved.
- **Assessment.** The following components should be assessed here (Anguera & Chacón, 2008): a) **Efficacy:** were the proposed objectives met (World Health Organization, WHO, 1981); b) **Effectiveness:** in addition to the proposed objectives, were any unexpected effects observed, whether positive, negative or neutral (Fernández-Ballesteros, 1995; World Health Organization, WHO, 1981); c) **efficiency:** this refers not only to the cost-benefit relationship shown by the obtained outcomes (World Health Organization, WHO, 1981) but also covers the complementary aspects of cost-effectiveness and cost-utility (Neumann & Weinstein, 2010), the latter being relevant to the issue of user satisfaction. In order to determine whether the same objectives could have been met at a lower cost, or better outcomes achieved for the same cost, it is necessary to compare similar programs; d) **Continuity:** the extent to which the initial formulation of proposed objectives and planned actions was maintained throughout; e) **Progress:** the degree of program implementation that is evidenced by the meeting of objectives (Fernández-Ballesteros, 1995; WHO, 1981); f) **Suitability:** the extent to which the evaluation was based on technically appropriate information (Fernández-Ballesteros, 1995); g) **Utility:** the extent to which the outcomes obtained can be directly and automatically used in subsequent decision making (Fernández-Ballesteros, 1995); h) **Probity:** the degree to which the evaluation has been carried out according to ethical and legal standards, with due attention being paid to the wellbeing of users and others involved (Fernández-Ballesteros, 1995); and i) **Equity:** this primarily refers to the extent to which the program planning and implementation have applied the same standards to all those individuals that make up the group of users or potential users, such that they are awarded equal opportunities (support, afford, etc.).

Discussion and conclusions

Perhaps the first point which should be made is that the proposed guidelines for reporting a program evaluation are based on essential and common aspects, the aim being to enable them to be applied across the wide range of settings that are encountered in the real world. For reasons of

space the different elements have not been described in detail.

In order to structure the presentation a distinction has been made between the different types of evaluation design. In practice, however, it is argued for the need to integrate designs involving a high, medium, and low level of intervention, there being three main reasons for this (Chacón, Anguera, & Sánchez-Meca, 2008): a) because program evaluation should be regarded as a unitary and integrated process; b) because users call for certain program actions that will need to be evaluated with the most appropriate methodology, which goes beyond the distinction between different kinds of procedure; and c) because in many programs it will be helpful to combine different methodologies, or even to introduce modifications that imply a shift from one method to another, this being done in response to the changing reality of users, and sometimes of the context, across the period of program implementation.

It is argued, therefore, that in terms of the methodological and substantive interests that converge within a program evaluation, it is the latter which should take precedence, notwithstanding the ethical requirement to avoid anything which might be detrimental to the physical and mental wellbeing of the individuals involved (Anguera, Chacón, & Sanduvete, 2008).

Given the limited amount of resources available it is essential for any program evaluation to include an economic analysis of the cost of program actions, of the efficacy achieved throughout the process and upon its completion, of the partial and total efficiency of the program actions, of the levels of satisfaction among users, and of a number of other parameters that enable the interrelationship between various elements of the process to be adjusted (Anguera & Blanco, 2008).

Finally, it should be noted that there is little point carrying out a program evaluation unless those involved in the program are encouraged to participate. Indeed, such participation is the key to a) obtaining information that will be useful for each of the groups concerned, b) increasing the likelihood that the outcomes of the evaluation will actually be used, c) making it easier to implement what has been planned, and (d) strengthening the validity of the evaluation (Chacón & Shadish, 2008).

Funding

This study forms part of the results obtained in research project PSI2011-29587, funded by Spain's Ministry of Science and Innovation.

The authors gratefully acknowledge the support of the *Generalitat de Catalunya* Research Group [Grup de Recerca e Innovació en Dissenys (GRID). *Tecnologia i aplicació multimedia i digital als dissenys observacionals*], Grant 2009 SGR 829.

The authors also gratefully acknowledge the support of the Spanish government project *Observación de la interacción en deporte y actividad física: Avances técnicos y metodológicos en registros automatizados cualitativos-cuantitativos* (Secretaría de Estado de Investigación, Desarrollo e Innovación del Ministerio de Economía y Competitividad) during the period 2012-2015 [Grant DEP2012-32124].

References

- Altschuld, J. W., & Kumar, D. D. (2010). *Needs assessment. An overview*. Thousand Oaks: Sage.
- Anguera, M. T. (1979). Observational typology. *Quality & Quantity. European-American Journal of Methodology*, 13, 449-484.
- Anguera, M. T. (1995). Diseños. In R. Fernández Ballesteros (Ed.), *Evaluación de programas. Una guía práctica en ámbitos sociales, educativos y de salud* (pp. 149-172). Madrid: Síntesis.
- Anguera, M. T. (1996). Introduction. *European Journal of Psychological Assessment* (monograph on Observation in Assessment), 12, 87-88.
- Anguera, M. T. (2003). Observational methods (General). In R. Fernández-Ballesteros (Ed.), *Encyclopedia of behavioral assessment*, vol. 2 (pp. 632-637). London: Sage.
- Anguera, M. T. (2008). Diseños evaluativos de baja intervención. 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., & Blanco, A. (2008). Análisis económico en evaluación de programas. In M. T. Anguera, S. Chacón, & A. Blanco (Eds.), *Evaluación de programas sociales y sanitarios: un abordaje metodológico* (pp. 259-290). Madrid: Síntesis.
- Anguera, M. T., & Chacón, S. (2008). Aproximación conceptual en evaluación de programas. In M. T. Anguera, S. Chacón, & A. Blanco (Eds.), *Evaluación de programas sociales y sanitarios: un abordaje metodológico* (pp. 17-36). Madrid: Síntesis.
- Anguera, M. T., Chacón, S., Holgado, F. P., & Pérez, J. A. (2008). Instrumentos en evaluación de programas. 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. 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. In M. T. Anguera, S. Chacón, & A. Blanco (Eds.), *Evaluación de programas sociales y sanitarios: un abordaje metodológico* (pp. 291-318). Madrid: Síntesis.
- Anguera, M. T., & Izquierdo, C. (2006). Methodological approaches in human communication. From complexity of situation to data analysis. In G. Riva, M. T. Anguera, B. K. Wiederhold, & F. Mantovani (Eds.), *From communication to presence. Cognition, emotions and culture towards the ultimate communicative experience* (pp. 203-222). Amsterdam: IOS Press.
- Bornas, X., Noguera, M., Tortella-Feliu, M., Llabrés, J., Montoya, P., Sitges, C., & Tur, I. (2010). Exposure induced changes in EEG phase synchrony and entropy: A snake phobia case report. *International Journal of Clinical and Health Psychology*, 10, 167-179.
- Casares, M. J., Díaz, E., García, P., Sáiz, P., Bobes, M. T., Fonseca, E., Carreño, E., Marina, P., Bascarán, M. T., Cacciola, J., Alterman, A., and Bobes, J. (2011). Sixth version of the Addiction Severity Index: Assessing sensitivity to therapeutic change and retention predictors. *International Journal of Clinical and Health Psychology*, 11, 495-508.
- Chacón, S., Anguera, M. T., & Sánchez-Meca, J. (2008). Generalización de resultados en evaluación de programas. In M. T. Anguera, S. Chacón, and A. Blanco (Eds.), *Evaluación de programas sociales y sanitarios: un abordaje metodológico* (pp. 241-258). Madrid: Síntesis.
- Chacón, S., Lara, A., & Pérez, J. A. (2002). Needs assessment. In R. Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment*, vol. 2 (pp. 615-619). London: Sage.
- Chacón, S., & López, J. (2008). Diseños evaluativos de intervención alta. In M. T. Anguera, S. Chacón, & A. Blanco (Eds.), *Evaluación de programas sociales y sanitarios: un abordaje metodológico* (pp. 219-240). Madrid: Síntesis.
- Chacón, S., López, J. M., & Sanduvete, S. (2004). *Evaluación de acciones formativas en Diputación de Sevilla. Una guía práctica*. Sevilla: Diputación de Sevilla.
- Chacón, S., & Shadish, W. R. (2001). Observational studies and quasi-experimental designs: Similarities, differences, and generalizations. *Metodología de las Ciencias del Comportamiento*, 3, 283-290.
- Chacón, S., & Shadish, W. R. (2008). Validez en evaluación de programas. In M. T. Anguera, S. Chacón, & A. Blanco (Eds.), *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. In M. T. Anguera, S. Chacón, & A. Blanco (Eds.), *Evaluación de programas sociales y sanitarios: un abordaje metodológico* (pp. 185-218). Madrid: Síntesis.
- Cornelius, V. R., Perrio, M. J., Shakir, S. A. W., & Smith, L. A. (2009). Systematic reviews of adverse effects of drug interventions: A survey of their conduct and reporting quality. *Pharmacoepidemiology and Drug Safety*, 18, 1223-1231.
- Fernández-Ballesteros, R. (1995). El proceso de evaluación de programas. In R. Fernández-Ballesteros (Ed.), *Evaluación de programas. Una guía práctica en ámbitos sociales, educativos y de salud* (pp. 75-113). Madrid: Síntesis.
- Gallego, M., Gerardus, P., Kooij, M., & Mees, H. (2011). The effects of a Dutch version of an Internet-based treatment program for fear of public speaking: A controlled study. *International Journal of Clinical and Health Psychology*, 11, 459-472.
- Griffin, C., Guerin, S., Sharry, J., & Drumm, M. (2010). A multicentre controlled study of an early intervention parenting programme for young children with behavioural and developmental difficulties. *International Journal of Clinical and Health Psychology*, 10, 279-294.
- Li, L. C., Moja, L., Romero, A., Sayre, E. C., & Grimshaw, J. M. (2009). Nonrandomized quality improvement intervention trials might overstate the strength of causal inference of their findings. *Journal of Clinical Epidemiology*, 62, 959-966.
- Milom, K. M., Jones, M. C., Kearney-Mitchell, P., & Tickle, M. (2009). A comparative needs assessment of the dental health of adults attending dental access centres and general dental practices in Halton & St Helens and Warrington PCTs 2007. *British Dental Journal*, 206, 257-261.
- Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic review and meta-analyses: The PRISMA statement. *BMJ*, 339, 332-336.
- Neumann, P. J., & Weinstein, M. C. (2010). Legislating against use of cost-effectiveness information. *The New England Journal of Medicine*, 363, 1495-1497.
- Peterson, C., & Kim, C. S. (2011). Psychological interventions and coronary heart disease. *International Journal of Clinical and Health Psychology*, 11, 563-575.
- 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. *Psicothema*, 21, 509-514.
- Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *Annals of Internal Medicine*, 152, 726-732.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.
- Trochim, W. M. K. (2002). *The research methods knowledge base*. New York: Atomic Dog Publ.
- Wallraven, C. (2011). *Experimental design. From user studies to psychophysics*. Boca Raton: CRC Press.
- World Health Organization, WHO (1981). *Evaluación de los programas de salud. Normas fundamentales. Serie Salud para todos, 6*. Geneva: WHO.