Quality & Quantity (2005) 39:733–762
DOI 10.1007/s11135-005-3150-6

Evaluation in Design-Oriented Research

PIET VERSCHUREN* and ROB HARTOG
Department of Methodology, Nijmegen School of Management, Radboud University, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands. E-mail: p.verschuren@fm.ru.nl

Abstract. Design has been recognized for a long time both as art and as science. In the sixties of the previous century design-oriented research began to draw the attention of scientific researchers and methodologists, not only in technical engineering but also in the social sciences. However, a rather limited methodology for design-oriented research has been developed, especially as to the social sciences. In this article we introduce evaluation methodology and research methodology as a systematic input in the process of designing. A designing cycle is formulated with six stages, and for each of these stages operations, guidelines and criteria for evaluation are defined. All this may be used for a considerable improvement of the process and product of designing.

1. Introduction

Up till the last decade most of research methodology in the social sciences is primarily concerned with theory-oriented research, as at that stage most of these disciplines aimed at knowledge just for knowledge (l'art pour l'art). As a consequence of a push from society, from then on scientific researchers and methodologists pay increasing attention to practice-oriented research. In this challenge they are mainly focussed on improvements of existing reality. More specifically, they aim at the solution of what may be called *improvement* problems. However, the last few decades gradually came into being another type of practice-oriented research that is aimed at the creation of a new artefact. Here the researcher aims at a solution of a so called *construction* problem. In design literature these improvement and construction problems are labelled "normal", respectively, "radical" (Vincenti, 1990) or "inventive" (Dasgupta, 1996) problems.

In this article, we will focus on research aiming at solving construction or inventive problems, which we will call *design-oriented* research. This type of research exists already for a long time in the technical disciplines.

*Author for correspondence: Piet Verschuren, Department of Methodology, Nijmegen School of Management, Radboud University, P.O. Box 9104, 6500 HE Nijmegen, The Netherlands. Tel: 31243611469, 31243581324; Fax: 31243612351; E-mail: p.verschuren@fm.ru.nl

But in the social sciences it is rather new, so that it lacks in large part the support of design methodology. Moreover, in our view existing design methodology does not provide sufficient explicit rules on evaluation as an integral part of a designing process. Literature on designing (Alexander, 1979; Brown, 1988; Simon, 1996) indicates that designers should be well aware that designing involves more "perspiration than inspiration". That is, the designer must be very critical as to the utility and satisfaction of the future users and the other stakeholders. So the artefact to be designed, once realized, should satisfy a set of design criteria. From this it follows that evaluation should play an important role in the process of designing. And this in its turn means that the designer may benefit from existing research methodology in general (Yin, 1984; Creswell, 1994; Denzin and Lincoln, 1994; Yin, 1994; Babbie, 1998), and from evaluation methodology in particular (Mohr, 1995; Patton, 1997; Rossi et al. 1998, Pawson 1998; and Tiley, 1997). For that reason in this paper both evaluation and empirical research will have a central role in the *process* of designing. We will match design-oriented research on the one hand, and existing know how on evaluation research and research methodology in general on the other.

We firstly unravel the process of designing in six stages, the so-called designing cycle, as a counterpart of the intervention or policy cycle in business and policy administration (Section 2). Next we give a short overview of different types of evaluation that are relevant for designing (Section 3). Then criteria for evaluation of processes and products of designing are formulated, ready for use as touch stones in each instance of evaluation (Section 4). Finally with the aid of these tools we elaborate on evaluation within each of the six stages of the designing cycle (Section 5), followed by conclusions.

2. The Designing Cycle

In order to create in Section 5, a detailed overview of the role that evaluation may play during and after the designing process, we firstly make a systematic and generic inventory of the designing process. We distinguish six stages of what we will call the designing cycle.

(1). *First hunch:* The very first stage of a designing process is the appearance of a first *hunch* and *initiative* for constructing a new material or immaterial artefact. The main result of this stage should be a small set of goals [G] to be realised with the artefact to be designed. For instance the goal of an aircraft designer may be the construction of a new type of aircraft aimed at the transportation of flowers from Africa to the northern hemisphere. Or a manager may want to have designed a helpdesk system that supports all employees of an organisation with respect to the use of office applications such as a spreadsheet program and a word processing program.

(2). *Requirements and assumptions:* The next step entails a specification of the *requirements* [R] to be fulfilled within the frame that is defined by the goal(s) [G]. We distinguish three main types of requirements, of which two are to be divided in sub types. The first are f*unctional requirements* [R_f]. They indicate the functions that the artefact should fulfil or enable to perform once it is realised, given the goal(s) [G]. In our example functional requirements derived from the goal [G] of the aircraft designer are that it allows rapid charging, and that internal climate conditions can be controlled within a wide range of temperature and humidity. Functional requirements of the helpdesk in the other example are that it can retrieve and deliver adequate information to employees who got stuck with one of the office programs. The rest of the requirements to be fulfilled in the process of designing regard the interface between the artefact to be designed and the "world outside". A first category of requirements of this type is the set of users requirements [R_u], to be fulfilled on behalf of the future users of the artefact. In the first example, these are demands of the pilot, of the rest of the crew, of the pursers and of the employees in charge of the maintenance of the aircraft. The user's requirements of the helpdesk may be that the help comes in time, and that it is offered in different forms such as audiovisuals and text.

A third and last category of requirements are the *contextual requirements* [R_c]. These are prerequisites set by the political, economical, juridical and or social environment. In our aircraft example there may be all kinds of constraints set by the governments involved as to the environment, pollution and noise caused by the aircraft. The design of the helpdesk system will have to take into account laws that are intended to protect employees and for instance to prevent repetitive strain injury or complaints related to intensive display screen use.

However, the designer not only has to design the artefact in such a way that it fulfils the desires of the future users and the demands coming from the context. He or she should also specify what qualities the users and the context should have in order to make a fruitful use possible. We will denote these as *assumptions* [A]. They are (to be) made by the designer, and thus must be checked with respect to their credibility and feasibility as part of the designing process. Just as was the case with the design requirements, the assumptions may regard the future users [A_u], the context [A_c] and the functions to be fulfilled [A_f]. For instance, as to the users of the aircraft, the designer will have to make assumptions about the minimum length of the legs and arms of the pilot. As to the context, the runway should have a certain length, and there are constraints as to the existence of obstacles in the form of buildings at the end of it. Without cooperation of local and national authorities the realization of these assumptions [A_c] may be difficult or even impossible. The helpdesk system designers will also have to make assumptions, such as assumptions about the language or languages that will be supported.

(3). *Structural specifications:* In this stage, we derive from the design requirements [R] and assumptions [A] the *structure* of the artefact-to-be, i.e., the characteristics, aspects and parts that the material or immaterial artefact must have in order to satisfy the whole set of requirements [R] and assumptions [A] from stage 2. These will be called the *structural specifications* [S]. In the aircraft example, a few structural specifications [S] may be derived from the functional requirements [R_f] already specified. These are: a relatively light engine, a large cargo door, and a climate control system with a wide operating range. Of course, all three should be further specified. The end product of the first three stages is a document on paper or in some electronic form, describing a first draft of the design in full detail.

In fact this third stage is the most complex one of the designing cycle, as here the design has to get its first form. For that reason we will make a few further distinctions. The first regards a distinction between systems and sub-systems. Systems of an aircraft are the fuselage, the engine, the electrical system, the hydraulic system, the wings and the landing gear. Sub systems of the electrical system are the battery, the wiring, the alternator and so on. However, sometimes it is easier to focus on a hierarchy of *processes* rather than of structures. For instance the task of the helpdesk in our second example is clearly a process that can be decomposed into sub-processes such as accepting a helpdesk call, interpreting it, matching the call with a set of possible responses et cetera. This last sub process can be decomposed into a full text search, an indexed search, browsing a frequently asked questions list, et cetera. The building of a smart hierarchy of different processes and sub-processes may result in a more efficient or effective helpdesk.

Another distinction is the one between a *general* (or rough) and a *specific* (or detailed) design. The general design is an overall architecture of the object to be designed. It will contain only the first and second level of the hierarchy of subsystems of this object. Decision-making at this preliminary stage is of strategic nature in the sense that these strategic decisions have consequences for the remainder of the designing process, for the result as a whole, and cannot be reversed without reversing many other decisions as well. Decision making during the specific design stage is of a tactical/operational nature in the sense that decisions in the detailed design stage are only related to subsystems at a lower level in the hierarchy, have limited consequences for other subsystems, and require limited or no knowledge of the system as a whole.

(4). *Prototype:* The next step is *realisation* (immaterial artefact) or *materialisation* (material artefact) of the design into a prototype (''). This prototype embodies the complete design and is useful for empirical evaluation. The designer should make clear whether all the structural specifications [S] are preserved in the prototype once it is realised. If there are differences, he or she has to find out whether these are logic or functional. Both in the case of the flower transport aircraft as in the case of the helpdesk system, we can realise such a prototype. However, with the current state of technology the helpdesk system will mainly be a business process performed by helpdesk employees. Although supported by software, we should be aware that the helpdesk prototype is not just a "thing" like the aircraft prototype.
A special form of a prototype is a, mostly incomplete, mock up. It is used as a means to make discussions about the functional [R_f] and the users requirements [R_u] less abstract (Stapleton, 1997). For instance, in the early stages of the design of a decision support system (DSS) a series of mock-ups may show the users what they may expect from the DSS. As this form of a prototype in principle is not an object for evaluation, in this article we will refer to the first complete and full sized form of a prototype unless mentioned otherwise.

(5). *Implementation:* In this stage, the designer has to put into practice the prototype, preferably in a real life context, as a first check whether it will work appropriately in the next stage. This means that a context must be realised that is compliant with the assumptions. As the designers had to assume at least certain competencies for different classes of users, implementation almost always implies quite a bit of training of those users who will take part in the tests of the prototype. For instance, the employees of the helpdesk may be trained in and or tested for communication skills, and the clients of the helpdesk who take part in the tests may be trained in how they are supposed to access the helpdesk.

(6). *Evaluation:* The last step of the designing cycle is to check whether the short and long term effects of utilisation of the prototype fit the design goal(s) [G] and satisfy the expectations of the designer and notably of the various stakeholders. This appears to be mainly the ex post summative type of evaluation (Section 3).

After this overview a few remarks on the designing cycle have to be made. First of all it should be noted that very often evaluation in stage 6 points out that the artefact not yet fully comes up to the goals [G] and the expectations or requirements of the stakeholders. This may be an occasion to start a second run of the designing cycle. If on the basis of evaluation it is clear in what stage in the first run deficiencies occurred, then the second run may start at this very stage. If not the designer most probably has to start a second run in the first or the second stage of the cycle.

A second remark is that, although the two final stages of the designing cycle explicitly aim at evaluation, we will explain how important it is that evaluation takes place during the whole process of designing. This is extensively elaborated in Section 5.

And last but not least, although the stages of a designing process are presented in a linear order, it should be realized that the designing process should be highly *iterative*. This means that the designer continuously goes back and forth between the several stages (at least mentally), looking what repercussions a decision in one stage has for earlier as well as for later stages. Often a number of iterations both within one stage and across stages is necessary to obtain a final or definitive design or artefact that is well balanced and comes up to all our expectations. Especially, if the design stages are very time consuming and/or expensive, it is worthwhile to prevent disappointments in later stages by means of early control measures, both in terms of feed forward, feed back and ex ante evaluation (see Section 3). Finally, in industrial designing the transition to large scale production involves a number of extra stages (Asimow, 1962). These do not fall within the scope of this article.

3. Types of Evaluation

There is quite a bit of literature on design methodology, although this is more the case in the domains of technical engineering (Asimow, 1962; Cross, 2000), architecture (Alexander, 1964) and the building of information systems (Dasgupta, 1991; Gamma et al., 1994) than in the social sciences. However, this literature is surprisingly implicit on the subject of evaluation. What we need is an explicit role of evaluation in the process of designing, as well as a conception of evaluation that goes far beyond common sense. For that reason in this and the following section, we will link the designing process to existing evaluation methodology and research methodology.

In this article, by evaluation we mean: "*to compare separate parts of a designing process with selected touchstones or criteria (in the broadest sense of the word), and to draw a conclusion in the sense of satisfactory or unsatisfactory*". Within the context of designing we make a distinction between the following three rough stages of designing: (a) The plan (on paper) of the design, i.e., the product of the first three stages of the designing cycle, (b) the realisation or carrying out of this plan, which roughly regards the stages 4 and 5, and (c) the effects that the use or the presence of the artefact has, i.e., stage 6. This threefold grouping of stages coincides with a well known distinction in evaluation methodology, i.e., *plan, process and product* evaluation, respectively. So the first three stages 1–3 the methodology of plan evaluation should be used, in the stages 4 and 5 we may make use of process evaluation and in the last stage 6 product evaluation is at stake.

Plan evaluation implies an assessment of the quality of the design on paper. If we call the combined set of requirements [R], the assumptions [A] and the specifications [S] the *means* to achieve the goal [G], then a plan evaluation involves mainly a separate test of the adequacy of (1) the goal, (2) the means and (3) the relationship between the goal and the means. More details about plan evaluation follow in Section 4, where criteria for evaluation are formulated.

For the second group of stages a focus on process evaluation implies that the issues and objects of consideration are the constructive activities and the means that are used in realizing the plan that was the result of stage (3).

Product evaluation, finally, involves finding out what are the results of the designing process, what the value of these results is, and what are the short and long term effects of the artefact once it came into being.

Although, above the three types of evaluation are linked with three rough stages in the designing cycle, we may also use process and product evaluation to each separate stages 1–6. That is, besides an evaluation from the designing activity as a whole, each separate stage asks for carrying out activities (process) and should end up with a result (product). So, each of the six stages should also be evaluated on its own merits by means of process and product evaluation. Especially, process evaluation of this type is very important.

Plan, process and product evaluation differ highly as to the aim of the evaluation and the way the evaluation has to be carried out. The aim of a plan evaluation is a logical, ethical and empirical check of (the quality and appropriateness of) all separate design requirements [R], design assumptions [A], structural specifications [S], and the design goal(s) [G]. It should also be evaluated whether they form a coherent and balanced whole. One reason for the latter is that the whole in principle is, or at least should be, more than the sum of its parts.

In general the aim of a process evaluation is to *improve* the process, and via this the product, of designing. Very often process evaluation is also essential in order to prevent defects that will be hard to detect, let alone to repair, in the very last stage of the designing process. For instance, the number of tests that should be done on the final version of the software of the helpdesk system in order to ensure that there are no errors, is too large and too time consuming from a practical point of view.

The aims of a product evaluation differ from those of a process evaluation. One possible aim of product evaluation is making a decision whether to *stop* or *continue* the designing process. The decision to stop the designing process may either be based on lack of progress, or because the design goal [G] is fully achieved, so that the artefact is not needed any longer. A second and mostly very important aim of a product evaluation may be *legitimating* the activity for the stakeholders, the efforts it takes and the money it costs, or *motivating* the stakeholders for delivering or continuing their passive or active support.

The three types of evaluation, plan, process and product evaluation not only differ with respect to their aim. They also are to be conducted in a quite different way. First of all process and especially product evaluation both are mainly empirical, i.e., based on sensory observation. In contrast a plan evaluation besides an empirical also is a logical and mental test, and thus in principle refers to desk research besides an empirical research. Second, in a process evaluation we in principle need a diachronic or longitudinal approach, as we want insight in the designing *process*. More specifically this regards the process of development and implementation of the prototype. For doing this most often *qualitative* methods of research are more apt than quantitative ones. The reason for this is that processes mostly are so complex that we need tenths or even hundreds of variables to grasp it in its full extent (Verschuren et al., 1997).

Types of qualitative data gathering are open or participant observation, qualitative content analysis of written and audio-visual documents and open or in-depth interviews with relevant actors. As a qualitative overall research *strategy* the case study design may be used. For more information on qualitative research methods the reader is referred to (Yin, 1984; Creswell, 1994; Denzin and Lincoln, 1994; Verschuren, 2003).

In contrast, in a product evaluation we in principle need measurements at least at two different points in time; once just before the artefact to be designed is realised, used or put into practice, and once after it is realised or used for a while. By comparing the results of these two measurements we know whether something has changed or not, and if so, how much and in what direction it changed. This effectiveness research most often is of a reductionistic type (Verschuren, 2001), based on quantitative measurement. In general two different research strategies may be used here, i.e., the correlational approach (large scale survey) and the (quasi) experimental research strategy. Applied to the designing cycle, product evaluation may mean that the researcher/designer tries to figure out whether the artefact helps in achieving the goal [G], both in the short and the long run. This involves a check whether this goal came closer compared to the situation before the artefact became operational. Besides this overall evaluation, product evaluation can also be applied to each separate stage of the designing process, as will be pointed out in Section 5.

Closely related to the distinction of plan, process and product evaluation is the distinction between *summative* and *formative* evaluation. Roughly the difference may be indicated as follows: "If the client tastes the soup this is summative, and if the cook tastes the soup this is formative evaluation". In fact the definition in italics of evaluation above fits summative evaluation. For formative evaluation we have to add to this definition *" … in order to make an improvement, i.e., to come closer to* [G], [R_f], [R_u], [R_c], [A_f], [A_u], [A_c] and [S]". Here the main objective of the designer/evaluator is to find out how the activity or the product of this activity, i.e., a design or an artefact, can be improved so that it fits better the set of requirements, assumptions and specifications. To achieve this we in principle need to know how the designing process was executed. For that reason, formative evaluation often involves what is called a process evaluation. In short, formative evaluation is to be characterised as a learning activity on the basis of both process evaluation and product evaluation.

Although summative evaluation comes close to product evaluation, and formative evaluation to the combination of plan and process evaluation, they better are not regarded as the same. First of all there are different forms of product evaluation, depending on the touchstone or criterion for testing that is being used. Only some of them may be qualified as summative. Four generic criteria may be used in product evaluation: (a) Has something changed? (b) In the right direction, i.e. the direction of the design goal [G]? (c) As a consequence of the designing activities or of the resulting artefact? (d) With minimal efforts, costs and or negative consequences or side effects? Type (b) indicates *goal achievement*, (c) represents the *effectiveness* criterion, and (d) implies a current definition of *efficiency* of the designing process. In general only (c) and (d) are qualified as summative evaluation. Thus we cannot take summative evaluation and product evaluation as identical. Second, sometimes we directly know on the basis of the type of deficiencies of the product, how the designing process can be improved, without ex post evaluation of this process. Thus process evaluation and formative evaluation neither can be regarded as identical, although they are closely related.

According to the above, often used alternative concepts for summative evaluation of type (c) are *effect measurement* and *effectiveness assessment*. Here the designer/evaluator has to find out whether the use or simply the presence of the artefact has the effect that was aimed at, i.e., the achievement of the goal [G]. This boils down to proving a causal relationship between the presence or use of the artefact on the one hand, and the changes that are observed afterwards on the other. Implicitly here we use the following definition of causality: A phenomenon X is said to cause another phenomenon Y if a change in Y, the effect, would not have occurred without an earlier change in X, the cause. So in the case of designing, an effectiveness assessment entails the observation of a change in the object or process to be changed, in the direction of the set goal(s) [G], and that would not have occurred without the designing activity. If only the first two requirements are realised this is an instance of *goal achievement.* The latter only indicates effectiveness of the designer under the condition that there is a causal relationship between the goal achievement and the designing activity.

A third distinction that is important in the context of designing artefacts is the one between *ex ante* and *ex post* evaluation. Ex ante evaluation of an activity (process) or its result (product) is evaluation before this activity has started, respectively, before the aim of this activity is realised or put into practice. Ex ante process evaluation is usually meant to assure the correctness of the designing process and to incorporate guarantees that the resulting design will not be a failure. (Asimow, 1962) formulates it in terms of confidence; ex ante evaluation is a process of finding evidence that increases or decreases our belief that a particular concept can be realised physically. Ex post evaluation is evaluation after the construction has been finished or the result of a stage has been realised or brought to practice. In an *ex ante* evaluation we preview the guidelines and constraints of later stages of the designing process, in order to take these into consideration in advance. The function of ex post evaluation of designing is mostly to give feed back to the actor about his or her performances or to decide on continuation or discontinuation of the designing process or of a line of thought in the designing process. In contrast ex ante evaluation is rather oriented at feed forward, that is to set constraints on future actions in the process of designing in order to assure its outcome. As such ex ante evaluation is at the heart of an iterative designing strategy. In fact ex ante evaluation is an important means to reduce the number of iterations needed.

Some authors argue that in our case of designing ex ante evaluation is the most important of the two, and often even is the only realistic option for the designer because ex post evaluation may come too late (see Alexander, 1964). For instance, ex post evaluation of the design of a skyscraper, assessing the potential danger of earth quakes is not a realistic option. Generally, if changes turn out to be disastrous it is too late to change the design. Another instance of little or no use of an ex post evaluation, occurs if the situation that required the design does not exist any more after the completion of the artefact. This may be the case in situations that change very quickly. This makes ex post evaluation a purely academic activity. However, all this does not imply that ex post evaluation should not be carried out. It just means that ex post evaluation is not relevant for that specific design.

Besides, the reader should keep in mind that the concept of 'ex post' is relative. Ex post evaluation may be relevant for further improving the artefact after the first run of the designing cycle is finished. That is, it can be used as an input for a second run aimed at a further development of the artefact or for adapting it to changing conditions. This situation may happen quite often, as most (construction) problems are so complex that they cannot be solved in one single run. In that case it has a *formative* function as opposed to summative evaluation. And, last but not least, ex post evaluation is important for improving design methodology. That is, especially with respect to our faults made during the process of designing as a whole we can learn how to do better next time.

A fourth distinction is the one between *goal based* and *goal free* evaluation. In a goal based evaluation we judge a design or parts thereof as to the extent that it contributes to achieving the design goal(s) [G]. So, as a generic example, effectiveness assessment is goal based by definition. In contrast, in a goal free evaluation the evaluator observes whether the design satisfies general professional or practical criteria or standards not directly linked with the design goal [G].

For most people goal based evaluation is the normal case. However, there are at least two handicaps in goal-based evaluation in the context of designing that may make it a difficult if not impossible job. The first is as already said that in general the design goal [G] seldom is quite clear at the beginning of the designing process. And if it is clear it often is defined only at a conceptual level instead of in operational terms. The reason for this is that there is very easily disagreement between the stakeholders as to the design goal(s). Keeping them at an abstract level or vague is one of the most used strategies to achieve consensus. Secondly there may be several goals without an indication of their order of priority. For obvious reasons in a goal-based evaluation we need operationally defined, stable goals that have a predetermined order of priority.

Finally a reason why goals in general and designing goals in particular are often not operationally defined is that people in general are less "goal rational" than they seem to be at first sight. Accordingly, the design goals [G] very often are not clear at the start and the designer further develops them during the designing process. Unfortunately for the evaluator this is rather common practice. (Verschuren and Zsolnai, 1998). He or she may try to solve this by carrying out a goal-free evaluation. This may mean that instead of using design goals as a standard, the designer/evaluator has to use other general criteria, such as whether future users will accept the artefact to be designed, or whether the artefact comes up to general professional standards or to expectances of the public.

4. Criteria for Evaluation

According to our definition of evaluation we have to compare *facts*, i.e., processes, plans and products of designing, with a touchstone or a set of *criteria*. In this section, we will develop criteria that may be used by a designer. We will elaborate on criteria for (a) plan evaluation, (b) process evaluation and (c) product evaluation respectively, both on the level of de designing process as a whole and to each separate stage. A plan evaluation mostly is not needed on the level of each separate stage.

Plan evaluation: As already said, a plan evaluation entails an overall evaluation of the design on *paper*, i.e., a first draft, which covers the first three stages of the designing cycle. The reader should make sure that these three stages have a hierarchical order, in so far as they constitute a goal – sub goal structure. That is, the design requirements [R] and design assumptions [A] represent sub goals to achieve the design goal [G]. The requirements and assumptions in their turn will be fulfilled with the aid of the structural specifications [S] of the design. So the entities that are at stake in a plan evaluation are the goals [G], the requirements [R], the assumptions [A] and the structural specifications [S]. In a plan evaluation these have to be considered as to: (a) *their own* separate value and (b) the way they are*related* to each other.

As to (a), criteria from which the evaluator/designer might choose in order to judge the design goals [G] are clearness, consensus of the stakeholders, feasibility, affordability, opportunity, ethical acceptability, and in case there is more than one design goal [G], whether they are rank ordered as to priority. Especially clearness is very important, as a most popular means to achieve consensus of the stakeholders is to keep the goals vague. This vagueness will severely hinder a product evaluation of the artefact later on, as already pointed out above.

Clearness also is a very important criterion for evaluating at face value the requirements [R] and the assumptions [A]. For purposes of goal based product evaluation to be carried out later on in the designing process they mostly are neither *sufficiently detailed* nor *operational*. So they have to be (a) unravelled into several constituting parts and aspects, and (b) made operational. In the context of designing operationalisation means two things: (1) To make clear at what modality (nominal variable) or score (metric variable) on a criterion the designer can be satisfied, i.e., putting *concrete* and *exact* norms. (2) To make clear what the designer has to *do* in order to (better) come up to the norms in case of formative evaluation. An example of unravelling is the following one. Imagine one of the users' requirements for the aircraft example is that it has favourable control characteristics. Then "control characteristics" has to be broken down into aspects such as "behaviour in turbulent conditions", "sensibility for crosswind during take-off and landing", and "response time to the control wheel". These in their turn may ask for further unravelling. For instance, "behaviour in turbulent conditions" may regard different categories of users. It may include the effects on the control wheel and rudder pedals (pilot), the shaking of the fuselage (passengers) or the allowed maximum speed in specified conditions (all).

As to type 1 operationalisation, an example can be taken from instructional design, more specifically from course design. If the goal [G] in a course "decision theory" is that students become motivated to achieve generic insights in the phenomenon of decision-making, this by far is not an operationally defined goal. An operational specification is for instance that 90% or more of the students who attend at least 80% of the lectures should have at least an 80% score on a measurement scale for measuring motivation. The reader should make clear that here again a list of additional conditions must be specified as to the type of students, the type of teachers, all kinds of learning aids et cetera, in order to make the requirement fully operational. Once the requirements [R_f], [R_u] and [R_c] have been unravelled and operationalised, i.e., translated in observable or measurable terms, they are labelled (operational) *design criteria* [C]. Depending of what type they are, they are denoted [C_f], [C_u] and [C_c]. They are the result of unravelling and an operational specification of more roughly defined functional, users and contextual requirements, respectively.

As to type 2 operationalisation the following is relevant. Defining a mode or a score is not enough if it comes to *formative* evaluation. If for instance the aircraft turns out to consume more kerosene than was specified in advance, this fact alone does not provide any clue as to what can be done in order to come closer to the required criterion in the next test. However, a design requirement regarding the composition of the exhaust fumes may point to incomplete combustion or combustion at certain temperatures, and thus give information about the efficiency of the combustion. Also if the students of the course "decision theory" receive a low score for motivation this in itself does not tell the designer how to improve the learning material. For a formative evaluation we also need information as to which parts of the learning material are appreciated most and which least by the students.

Besides, this operational clearness, for an evaluation of the requirements $R_{f,}$ roughly the same criteria may be used as for the goal(s) [G]. As to R_u and R_c by far the most important criterion for evaluating is validity, i.e., the question whether these correspond to reality. Next may also be used criteria such as clearness, feasibility, affordability, consensus and ethical acceptability.

As to the second instance of plan evaluation concerning the *relations* between the several demands, we can be short. Here the central criterion to be fulfilled by the designer is *fit*. For instance, do the structural specifications [S] fit the design criteria [C] and the design assumptions [A], and via these, do they fit the design goal(s) [G]? Referring to the fact that it is a goal – sub goal structure, we distinguish three components: (a) are lower demands in the hierarchy sufficient to achieve the demand next higher in this hierarchy. If not this is an *error of omission*, and we have either to improve the elements lower in the hierarchy, or to extend their number, or both. (b) Are the "lower" demands not achieving more than intended by the designer. That is, these demands may be too far reaching in both a qualitative and a quantitative sense, with respect to achieving the "higher" demands. If yes this is an *error of commission,* and we have to reduce the scope of the elements lower in the hierarchy. (c) Are the sub goals logically consistent with their goals? Furthermore the designer/evaluator may use a set of general criteria, such as effectiveness (in contributing to a "higher" demand) ethical norms, opportunity, acceptance for the stakeholders, and practical criteria such as feasibility and affordability.

Process evaluation: A next step is the evaluation in/of the stages (4) (prototype) and (5) (implementation). Here in principle a set of actors become involved in the designing process. Quite a few criteria may be relevant to evaluate their activities such as: the degree and quality of their co-operation and of their communication, their expertise and skills, the aids and infrastructure that they use, the efforts they do, to mention only a few.

Besides, these criteria which are specific for the stages (4) and (5), there are also general process criteria that should be followed during the whole process of designing. In the literature on designing these are often called design *guidelines*. The reader should realise that design guidelines are essentially an articulation of a design methodology.

We distinguish *generic* and *specific* guidelines. For instance in designing instructions for students a *general* guideline may be to conceive the design as a compound of modules. This involves the designing of several relatively independent components of the artefact to be designed that later on may be plugged as independent units into the design (artefact) as a whole.

As designing an artefact involves decision-making, one category of generic design guidelines can be derived from decision theory. An example of a generic guideline borrowed from decision theory is the following. The effort invested in collecting information by the designer must be in balance with the expected importance of the decision that will be based on it. This design guideline will ensure that limited capacity for searching information will not be wasted on minor decisions. It can be formulated as follows in a generic way: "Allocate resources to design stages in proportion to the expected benefits of improving the design". It implies that the designer should not spend too much design effort during the design process on a minor aspect of the design. Another generic design guideline that can be derived from decision theory is that the scope of decisions during the designing process will be explicitly formulated and underpinned by arguments as well. This helps to reduce the number of design decisions with overlapping scope, and thus helps to reduce the number of potentially conflicting design decisions.

Another criterion for process evaluation borrowed from decision theory is whether the designer makes a distinction between *strategic*, *tactical* and *operational* decisions, and whether he or she does not make a mistake as to the order in which these decisions have to be taken. Strategic design decisions are essentially decisions about the artefact as a whole. They will be taken in the initial stages of the designing process. Reversing the order of strategic and tactical decisions incurs relatively high costs. In our aircraft example a strategic decision is the choice of both the type and the material of the fuselage and wings. This decision is strategic for several reasons. First of all it requires an overall view of the field of aircraft design. This in contrast to decisions about the form and the material of the fuel tanks, which domain is a rather narrow one. The latter is a tactical decision that can be delegated to someone whose expertise is much more limited, but maybe also much more detailed within its limitations. Second, changing a decision about the fuselage and wings will have consequences for many other decisions: another fuselage will have more or less mass, the mass will be distributed differently, the form of the aircraft will be different, the way it behaves in the air currants may be different et cetera. Finally operational decisions are decisions of a very detailed and recurrent nature such as decisions about the lay out of the hydrolics throughout the aircraft.

Next, we have to decide how to *check* process guidelines in the context of a process evaluation. Empirical measurement of departures from design guidelines may be difficult, because of the many, often not formalised and moral aspects that may be involved and that ask for human consideration. Currently there seem to be two options: we can invite experts to check if predefined process guidelines are being obeyed during the process or we can translate guidelines into requirements for intermediate products and check if the intermediate products satisfy these requirements. The latter is often feasible as has been shown for instance in Quality Function Deployment (Hauser and Clausing, 1988). However, it may slow down the design process considerably because of the extra effort that must be invested in the production of intermediate products. The modular design guideline is easy to test by checking if intermediate products are modular. But it is difficult to check whether the designing process is constantly aimed towards a modular design, i.e., by checking if the designer consciously thinks in terms of modules with minimal interdependencies and maximum internal coherence. On the other hand this form of expert review is exactly what an advisor of a student does in design-oriented studies. So we all will recognize that it is a feasible approach.

At the end of this stage of process evaluation both the *utilisation* and the *performance* of the prototype has to be evaluated in order to improve it before the large scale production of the artefact starts. As will be pointed out in Section 3 below this is an instance of *formative* evaluation. Ideally components of the design will have been prototyped earlier and tested as well. For instance the climate control system in our aircraft should have been tested before the first test flight. However, further tests during test flights will be necessary as well, in order to know whether everything works *as an integrated whole*. This implies that also the assumptions [A_f], [A_u] and [A_c] should be tested once again at this stage.

Product evaluation: Once the prototype is realized and implemented the designer has to find out whether it comes up to our expectations, or whether in another sense it has favorable effects or outcomes. This is to be done in stage (6). First of all we refer to the four generic criteria for product evaluation in the last section, i.e., change, goal achievement, effectiveness and efficiency. As the two latter types of product evaluation are goal based, the design goal [G] of course functions as a main touchstone. Moreover, the question of causality has to be solved. In other words, the designer has to find out whether the artefact directly or indirectly is responsible for the outcomes or not. In principle for this we need to do measurements in the domain of [G] at least at two points in time; one before the artefact comes into being and one after it has been functioning or used for a while. For more information on this type of effectiveness assessment see Pawson and Tiley, 1997.

Thus far the product evaluation regards summative evaluation. However, if effectiveness assessment at stage 6 (or some time later) points out that the artefact fails to achieve the goal(s) partially or totally, then it may be worthwhile to start a *formative* evaluation. That is, we have to look at all those parts of the designing process in order to see (a) whether there is a reason for this failure, and (b) how we can improve the artefact. Especially, we have to check whether the requirements [R] and assumptions [A] are adequate and correctly derived from the goal(s) [G], whether the structural specifications [S] are adequate and correctly derived from the requirements [R] and assumptions [A], and whether the requirements and specifications were correctly transferred to or translated into the prototype. This in fact means that we start a second run of the designing cycle.

A final remark is that besides or instead of goal based evaluation we may also want to see whether the artefact has favorable outcomes that have nothing to do with the design goal [G] and thus were not envisaged, but that nevertheless have a positive value. This is an instance of goal free evaluation. This especially may happen if the goal(s) are shifting over time, that is if circumstances and demands of stakeholders have been changing during the process of designing.

5. Evaluation in Stages

Once it is clear how the designing process should be structured, what types of evaluation are relevant, what is the role of design guidelines and requirements, and what type of criteria may be used, we can start a discussion on evaluation as part of the designing process. For each stage in the designing cycle we discuss the evaluation that should at least make part of the designing process. Here may be at stake local process and product evaluation on the one hand, and overall plan, process and product evaluation on the other.

1. First hunch: In this first stage of the designing cycle the designer/evaluator has to answer the question whether all conditions were fulfilled to have a fruitful idea about the creation of a new artefact. He or she especially must check whether the design goal [G] really covers the desires of the stakeholders. If the desires where studied by means of empirical research such as an interview, a questionnaire or documents, then the evaluators have to check the *validity, reliability, researcher-independence and verifiability* of the research. These are standard scientific quality criteria, largely elaborated in methodological handbooks.

Although many people may take this first stage of the designing cycle for granted, being more a matter of common sense, intuition and arts rather than of systematic thought, rationality and science, it is worthwhile to consider this stage from a methodological and an evaluative point of view. For obvious reasons it is very important that, right from the start, the designer has a clear idea and overview of all the recent social and technological, material and immaterial commodities, raw material, (half) manufactures, modules and subroutines that are available and from which he or she can make use in the process of designing, especially in the stages 4 and 5. It is very unlikely that new and fruitful ideas come out of the blue. (Alexander, 1964; Brown, 1988; Nonaka and Takeuchi, 1995; Csikszentmihalyi, 1996; Simon, 1996). If the designer is not up to date and well informed in this respect, the artefact to be designed and produced most probably will not be sufficiently innovative. (Csikszentmihalyi, 1996). It even may be superseded before it is produced. Thus a design guideline for this stage is that the designer(s) should invest sufficient effort in acquiring knowledge and information of all those aspects and details that may be important to construct the prototype, i.e., to realise the design. If in the helpdesk example the business process designers do not understand in detail the variety of problems with respect to office programs, if they are not familiar with office programs, and neither with the many different types of users, nor with the different forms of pressure that can be exerted on help desk employees, it is very unlikely that their concepts will turn out to be effective. In short, local process evaluation during stage (1) involves at least a check whether the designer is knowledgeable or invests efforts in knowledge acquisition.

However, a test whether relevant fields and disciplines were taken into consideration is not enough. An unexpected and innovative hunch becomes more probable whenever we bring together experts from (totally) different fields who normally have no contact with each other. For an iterative brainstorm of experts the researcher may make use of participatory research techniques such as Delphi, workshop techniques, focus group interviewing, gaming and scenario building with experts to elicit their relevant knowledge and information. The choice of experts and the communication and fine-tuning between them, must be balanced against the importance of the design problem and the opportunities of the designer. In many cases this implies that the construction problem at hand is already sufficiently well formulated in order to entice a number of people to invest attention in it.

However, this initial stage of the designing process is seldom satisfactorily formulated. Thus it is advisable to invest efforts in reformulating the design challenge and looking at the problem from different perspectives. It is clear that ex post evaluation of the result of stage (1) will not easily convey to what extent the designers have looked at the problem from different perspectives. So we have to ask them questions (interview or questionnaire) as part of the evaluation. At the same time it is clear that during the initial stages of the process of designing the guideline to look at the design challenge from different perspectives, may lead to the formulation of different partially mutual exclusive sets of functional requirements (R_f). For instance, in the case of the helpdesk, one line of reasoning is that the helpdesk provides information to employees who encounter a problem. Another line of reasoning is that the helpdesk should monitor what problems cost much time in the organisation, and then should come up with proposals for alternative ways of proceeding. If many employees ask helpdesk support with respect to the data handling options in the spreadsheet program, the helpdesk can try to find out why so many employees do individual data handling. It also may (help) answer the question why they want to do this with a spreadsheet program and what might be an alternative for the organisation as a whole. A helpdesk with this function will differ considerably from a helpdesk that only gives information with respect to a specific detail of an office application.

A designer looking for input from different perspectives should realise that these perspectives often stem from different underlying paradigms. Here communication between experts may invoke a problem, as it requires at least some overlap in language and conceptual knowledge. This will require a skilled facilitator and/or a learning effort of the experts involved. This too may be part of the designing task. In general, as is well known in literature on technical designing, design methodology must provide a vehicle for communication between users and designers as well as among designers themselves. Such vehicles are for instance a pattern language (Alexander, et al., 1977), entity type relationship type modelling (Chen, 1976), object role modelling (Halpin, 1995), unified modelling language UML (Booch et al., 1998; Rumbaugh et al., 1998) , blueprints and diagrams (Cross, 2000), logical constraints (Chandrasekaran, 1990) and the house of quality (Hauser and Clausing, 1988).

Thus, if the designer communicates with peer reviewers during this stage, these peers can observe and judge a line of reasoning and give feedback on it. For instance, if the designer aims at designing a system that supports the construction of schedules for schools, universities, transport companies, trains et cetera, someone with a background in classical optimisation techniques may implicitly assume that it is always possible to define an object function that satisfies all stakeholders. Peers with different disciplinary backgrounds such as artificial intelligence or organisational sociology will soon make explicit such an [A] and question the corresponding line of reasoning and the underlying paradigm as well.

Finally, with regard to a local product evaluation of this stage we formulate some dimensions for criteria to be put at the goal(s) [G], besides the one of clearness already mentioned. To be mentioned are feasibility, affordability, opportunity, acceptance by the stakeholders, moral justifiability, and opportunity costs. The latter criterion regards the question whether the adoption of another goal would have more profit.

In conclusion, we can say that the criteria for ex post evaluation at the end of this first stage of the designing process may be captured in the following questions: (1) Was the involvement and variety of experts well balanced against the expected impact of the design? We could set a quantitative norm for "well balanced" when we can assume that there is a relation between the total design effort and its expected impact. Based on this assumption we can require that the investment of different experts in the first stage should at least be budgetted at x% of the total design budget. (2) Is the goal [G] sufficiently sharp defined in order to derive the functional [R_f], contextual [R_c] and users requirements [R_u], and to give direction to the next stages in the designing process? (Feed forward or ex ante evaluation). (3) Have standard methodological guidelines for empirical research been followed during the process that led to the formulation of [G]? Or has empirical research been excluded in this stage on the basis of sound arguments? See also other criteria to be used at this stage, as these were formulated in the last section.

2. Requirements and assumptions: In this stage empirical research should be carried out in order to find out what are the users [R_u] and the contextual requirements [R_c]. Thus standard criteria for evaluation here are the empirical validity and reliability, as well as the researcher independence and verifiability of the results. As to R_f, at this stage the designer/evaluator has to carry out a logical test, whether the functional demands really fit the set goal(s) [G] of stage 1. The set of functional requirements [R_f] should cover this/these goal(s), no more (error of commission) no less (error of omission). (see Verschuren and Zsolnai, 1998). Next the designer needs insight in the user requirements [R_u]. It stands to reason that whenever there is a very large group of potential users, we may draw a random sample out of the target population, and send them a questionnaire by post or by email. Interviews may be better because they offer an opportunity to interact with the respondent, but these are more time consuming. Again compliance to standard scientific criteria is required. The questionnaire or interviews may yield a big number of different and even contrasting demands. If these demands cannot be reconciled in a satisfactory way, the designer either has to reduce the target group or to design different variants of the artefact, or both. The same holds for an operational translation of initial formulations of requirements into design criteria [C].

If the members of the users group are supposed to interact and communicate intensively in employing the artefact once it is realised, then once more in many cases the researcher preferably uses participatory techniques such as focus group interviews and workshop techniques. Because in these methods interaction and communication between the participants play an important role, this gives a better opportunity for obtaining relevant data than individual face-to-face interviews. A good option is gaming. This is a form of human simulation. By building a game of something that resembles the artefact as the object of gaming, followed by playing the game with the future users of the artefact, the researcher in principle obtains a clear and detailed insight in what is important or not. However, gaming is very expensive and time consuming. In any case it is important at this stage that the stakeholders get a clear idea of the artefact to be designed and of the context in which and the purposes for which it will be used.

Of course, all this is an instance of *ex ante* evaluation as the artefact still is not realised. Thus, if gaming is not opportune the questions must be answered on the basis of imagination and a "mental eye" of/on a future state of affairs. In principle methods and strategies for empirical research are of little use here. In fact this means that the specification of users requirements [R_u] may remain a problem (Dasgupta, 1991). Experience has shown that for innovative design users often do not know what they want, which makes validation of user requirements ex ante very difficult if not impossible. This is also the main reason for a designing strategy that involves forms of rapid prototyping in order to enable the prospective users to experience the opportunities and threats of the proposed innovations (Stapleton, 1997).

Finally the researcher has to find out what are the relevant contextual requirements [R_c]. What practical, social, political or juridical side conditions are at stake? Here to a certain extent the same arguments are valid as is the case with the assessment of the user's requirements [R_u]. Besides, as sometimes requirements are laid down in official documents, the researcher should gather these documents and carry out a systematic content analysis.

To derive and verify the separate design requirements [R] is a necessary but not a sufficient task at this stage. On top of it, as a part of a plan evaluation the designer has to check the logic of the *combination of* and the *relations between* the three classes of exogenous requirements [R_f], [R_u] and [R_c]. Here not only expert knowledge may be used. Besides, some special procedures are available. One of these is Quality Function Deployment (QFD) (Hauser and Clausing, 1988). This is a methodology that supports the process of making explicit the relations between [G], [R_u], and [R_c]. Central in the description of these relations is a series of matrix like constructs. Such a construct is called "the house of quality" in which the rows describe detailed user requirements in the language of the user, and the columns describe engineering variables in the language of the engineer. The roof makes a connection between engineering variables from different engineering aspects. This methodology may function as a design guideline and thus as a testing criterion for evaluation at this stage. However, as long as there is no experience with QFD in a sufficiently wide range of different design areas, application of it is certainly not a trivial exercise (Costa et al., 2001).

Another approach in the process of specifying exogenous requirements is to use a pattern language (Alexander et al., 1977; Gamma et al., 1994) to describe design patterns. A pattern language enables us to make explicit those design requirements that became clear during a series of previous design efforts in the same design area. Alexander (1964) sees "the process of achieving good fit between two entities as a negative process of neutralising the incongruities, or irritants, or forces, which cause misfit". His approach of defining design patterns aims at describing advice for neutralising those incongruities and forces. As an example we present one of the design patterns that are typical for a farm house in the Bernese Oberland:

"North south axis//west facing entrance down the slope// two floors//hay loft at the back//bedrooms in the front//garden to the south//piched roof//half hipped end//balcony toward the garden//carved ornaments//. (Alexander, 1979)". The use of design patterns as a checklist for evaluation has called growing interest in completely different fields (Gamma et al., 1994). The minimal form of process evaluation for this stage is a check whether they have used a methodology in order to establish the functional requirements [R_f], the user requirements [R_u] and the contextual requirements [R_c], and whether they have used a methodology to translate these requirements into operationally defined requirements, i.e., design criteria [C_f], [C_c] and [C_u]. As pointed out in Section 3, this entails unraveling key variables in parts and aspects, setting modalities or scores on these variables that should be achieved, and finding criteria that make clear how a design can be improved if formative evaluation urges this. The next step is to decide whether the designers did select a useful methodology for this stage, followed by a check whether the designers did correctly follow the guidelines in the chosen methodology.

Product evaluation in this stage involves answering the question whether the output of this stage consists of operationally defined design requirements, and if these requirements really cover the exogenous requirements and at the same time match the goals [G].

If all requirements are operationally defined, and if it could be established unambiguously that the requirements fully cover the goal(s) [G], then empirical evaluation in the later stages can be straightforward. In that case, the soft spot in the evaluation is localised in stage 2. In practice the output of stage 2 seldom completely satisfies all stakeholders. A number of design requirements is often formulated ambiguously or there is doubt whether these requirements [R] cover the exogenous demands and the goals. Both shortcomings of the outcome of stage 2 will lead to proliferation of soft spots to other stages of the designing cycle and make evaluation in other stages more difficult. This may force the designer to go back to stage 2 and to improve the formulations of the requirements. This again is an instance of an iterative designing strategy.

Finally at this stage, the designer/evaluator has to check the credibility and acceptance of the assumptions [A_f], [A_s] and [A_c]. This of course primarily is a matter of empirical research. In case of insufficient credibility the designer either has to induce changes in reality, for instance by giving information and instructions, or to adapt the design or both.

3. Structural specifications: In this stage evaluation aims at an assessment of the quality of the translation of the design requirements into the structural specifications. This is a logical rather than an empirical test. Here, we especially have to look at *structural alternatives*. That is, mostly a given functional requirement [R_f] may be served by several alternative structural characteristics of the artefact to be designed. For instance, in our example in Section 2, the functional requirement of rapid charging the aircraft may not only be satisfied by a large cargo door as has been proposed. A structural alternative is the use of containers on a rail system that may be charged in advance. For several reasons it is seldom feasible to select one and only one alternative as "the best". For that reason (Simon, 1996) introduced the term "satisficing" as the most typical characteristic of this stage of the design activity. The choice of the alternative depends on the infrastructural circumstances, desires of the users and stakeholders, and financial costs. So these too may be used as criteria for evaluation.

Besides, in this stage we have to evaluate as part of an iterative designing strategy whether the functional requirements [R_f] from stage 2 can be mapped to a composition of (sub) systems or should be adapted. The latter essentially implies a reiteration of earlier stages. Such iteration is quite common.

As already said, the output of stage 3 is a design on paper, i.e., a detailed outline of the artefact. It has the form of a blueprint that allows direct implementation of the outline into a prototype in the next stage. Process evaluation at this stage involves a methodological check of the specifications that are used during this stage, after they have been unraveled and operationalised (if necessary). A product evaluation implies a check whether the results of this stage are compliant with the design criteria [C_f], [C_u] and [C_c], and a check whether the results of this stage are sufficiently clear for those who have to work with the structural specifications in stage 4 (feed forward).

One of the differences between experienced and young designers is that the first mostly are able to directly evaluate and exclude at face value possible structural alternatives. Thus they efficiently allocate their resources to that set of alternatives that matters. A guideline that may be helpful notably to inexperienced designers is the following: First look at those constraints that cut off the largest parts of the set of alternatives but at the same time leave as much options open as possible. In other words search for those constraints that make many solutions not worth considering. But be also sure that you do not throw away a solution before you are convinced that the solution is not feasible.

By means of a continuous reflection on guidelines for efficient evaluation of possible alternatives ("the problem space" (Simon, 1996)), we may guarantee that the design resources are not spoiled on unpromising corners of this space. It is unlikely that a good design will result if most design efforts were wasted in the wrong parts of the problem space.

With respect to product evaluation in this stage, it is clear that the results of both the preliminary and the detailed design stage must be evaluated against the design criteria. As the design is not yet implemented in this stage, an empirical test of these criteria is still not feasible.

As to an eventual modular structure of the design, evaluation may involve a check by experts in order to make sure that the interfaces of the modules have been defined according to a specific formalism. Thus, for a modular course in higher education the interface of a module usually should consist of a specification of the prior knowledge that the student is assumed to have before he or she starts with the module and a specification of the type of assignments the student will have to be able to complete satisfactory when finishing the module. Both for prior knowledge description as well as for assignment types a specific format can be defined. A check on modular design will mainly involve a check on the interfaces. In this case, it will involve answering questions such as: is for each module the required prior knowledge listed and listed according to the specified format? In fact this part of the interface is identical to [A_u].

4. Prototype: In stage 4, the designers have to assure that the structural specifications are actually preserved in the prototype, no less, no more. Process evaluation for this stage implies a check on what the designers did to assure that the structural specifications are preserved. More specifically, a formative process evaluation at this very stage aims at realising the as yet best possible transformation of the structural specifications into a prototype.

When the structural specifications are completely unambiguous the evaluation implies that the designer or external experts review the transformation process along formal lines of reasoning. However, in reality structural specifications are seldom complete or unambiguous. This implies that the transformation process involves decisions. A review of the decision making moments in stage 4 should therefore be based on general decision making guidelines.

This involves a preliminary consideration of the functional criteria [C_f], the user's criteria [C_u] and the contextual criteria [C_c], as well as the designing guidelines [D]. If the prototype or the process to realise this do not satisfy these criteria, a diagnosis must be made in order to find out what exactly lacks or does not fit, and why this is the case. This in principle leads to a revision of the structural requirements [R_s] in order to improve the match with the design goal(s) [G], the users' requirements [R_u] and contextual requirements [R_c]. The designer has to adapt either the structural specifications [S] themselves, and/or the way these have been realised or materialised, and/or the functional requirements [R_f]. If this still does not work, most probably the functional, the users and/or the contextual requirements have to be changed, which again is an instance of iterative designing. This of course is rather sweeping as it may touch the roots of the design. Nevertheless this happens quite often, especially in situations where there is much uncertainty.

At this stage, product evaluation involves a check whether the prototype actually comes up to the structural specifications [S]. And if it differs as a consequence of an iterative designing strategy, we have to check whether this deviation is based on (a) sound arguments and (b) still fits the goal [G] and the requirements [R_f],[R_u] and [R_c]. If the latter appears not to be the case, then the researcher has to adapt the structural specifications or the way these are realised, or both.

In short, product evaluation at this stage regards at least the relation between the design (on paper) on the one hand, and the prototype on the other. This evaluation is an analogy of the verification step in modeling (Schlesinger et al., 1979). A mismatch between the symbolic representation and the prototype may be detected in an expert review or as a result of an empirical test of the behaviour of the prototype in the context where it should function in the next stage of the designing cycle. Also a focus group interview with the future users may shed light on the question how to improve the prototype.

Although other forms of prototyping do not fit the scope of this article, a few remarks should be made here. If the design does not aim at mass production, such as the design of a skyscraper, or a nuclear waste storage facility, or a law), or in case the costs of mass production are virtually zero (such as the design of digital materials) it does not make sense to realise a full-blown prototype. In such cases a scale model or partial product may substitute the role of a prototype. Testing whether a scale model or a partial product satisfies the design requirements is then the only realistic option, even though such a test is based on assumptions that relate the test results to the behaviour of the full blown product. Evaluation of the process that led to the scale model or to a partial product, implies an assessment of the theoretical line of reasoning that leads to the conclusion that the scale model or partial product both fit the structural specifications.

5. Implementation: In this stage, the process and the outcome of the implementation of the prototype has to be evaluated. In the context of a process evaluation, primarily formative rather than summative, we try to answer the question whether this implementation process was properly carried out. The designer next has to follow the adapted implementation process guidelines, leading to an improvement of the prototype.

Evaluation in stage 5 means that we must check whether the conditions under which the prototype is supposed to operate all have been realised. This boils down to a check whether the elements in set [A] have been satisfied. In the example of the helpdesk it is likely that the designers have made assumptions [A] about the way employees will access the helpdesk. Thus, the implementation of the helpdesk may imply that everyone will receive instructions or follow a training in how to access the helpdesk. Thus, implementation guidelines will involve a systematic check on all assumptions, and some action of the designer whenever an assumed condition turns out not to be fulfilled.

During the implementation stage or even during the test stage we often will detect assumptions which were made implicitly and must first be made explicit, or assumptions that were fulfilled when we started the design but are not fulfilled any more because the environment has changed while we were in the designing process.

Evaluation of the implementation stage of all this involves a check whether all the contextual design criteria [C_c] and contextual assumptions [A_c] have been satisfied. For instance, as to the latter, in our helpdesk example one of the assumptions [A_c] may be that every employee has a sound card on his or her desktop. Other assumptions may be that every user speaks and understands English, or that employees are willing to transfer responsibility for decisions at a detailed level to the helpdesk. For the latter they must have the right attitude. All these assumptions should have been made explicit.

Sometimes a problem with assumptions about the environment can directly be solved. For instance, in our example a soundcard can be added to those desktops that turned out to have no soundcard after all. Of course, it should be evaluated if the employees who will take part in the prototype test, got the right and sufficient instructions. This instruction must also raise the right expectations with these employees. Again these "right expectations" should have been formulated as elements in the set [A].

At this stage, the evaluator often has to rely on the opinion of practical and theoretical experts in relevant domains. For doing this in a methodologically sound way we may bring together these experts in a workshop or a focus group, thereby making use of appropriate participatory techniques. Besides, qualitative research methods such as systematic observation, in depth interviews and qualitative content analysis of written and audio-visual documents may be useful, rather than quantitative methods. The reason for this is that several aspects of the context have to be balanced. This can hardly be done in a quantitative or reductionalistic way, such as by means of paired comparisons of experts. We rather need a holistic approach, i.e., the use of group techniques and qualitative research methods (Verschuren, 2001, 2003).

At this stage, the evaluator should also check whether the right users were selected with respect to knowledge, skill, experience and attitude. And also whether the users have access to a relevant infra-structure and logistics. For checking compliance with these guidelines in a professional way the designer again preferably uses qualitative and participatory methods for data gathering.

At the end of this stage, the prototype is set into operation in an environment that is compliant with [A]. Then the behavior of the prototype and its environment is compared with the design criteria mentioned in Section 3. (Notice that some design criteria may refer to environmental variables! For instance, in case of the aircraft there could be a design criterion defining maximum wake turbulence and in case of the helpdesk there could be a design criterion defining minimum employee satisfaction). Evaluation of the test process implies reference to guidelines for evaluation of the outcomes of the tests (Pawson and Tiley, 1997). In this article, we include interpretation of the test results in this stage. Product evaluation entails answering the question whether the prototype functions well, given the design goal(s) [G], the design criteria [C_g], [C_u] and [C_c] and the assumptions [A_g], [A_u] and [A_c]. Of course, this is an instance of *formative* evaluation, as the objective is the improvement of the prototype. This may be helped by asking how the users and other stakeholders experience the (use of the) prototype, what problems they encounter, how they cope with them, and with what results. In case the prototype does not yet function very well the evaluator/designer has to check the adequacy of requirements [R], assumptions [A] and specifications [S], as well as the way they are derived from items higher in the goal – sub goal hierarchy.

All these elements should have been captured in the design criteria [C]. And if the designers have been able to formulate a fully clear and operational design, testing will be straightforward. We just have to execute the operations defined in the design criteria and compare the outcome with the criterion at hand.

However, in practice this is not probable, often as a consequence of shortcomings in stage 2 that are proliferated to this stage 5. To evaluate and tackle this we in principle prefer the use of qualitative and participatory research methods. Once more the most important qualitative research technique to be carried out by the designer is *systematic observation* of the users while they use or apply the artefact. Of course, this may be reinforced by means of interviews with the users and other stakeholders, and of qualitative content analysis of relevant written and audio-visual documents, in order to know what they think and feel about the artefact. The advantages of such a combination of methods are described in the methodological literature under the heading of *triangulation*.

6. Evaluation: In the words of Simon: "Everyone designs who devises courses of action aimed at changing existing situations into preferred ones." (Simon, 1996). In most cases, this implies answering the question whether, and if so to what extent, the construction problem at hand has been solved, i.e., the goal(s) [G] is/are achieved. In such cases evaluation implies goal based evaluation. It is also *summative* evaluation of the product, i.e., the artefact that is the result of the designing process in the first five stages of the designing cycle. More specifically this is an assessment of the effects of the artefact. Here the researcher has to find out to what extent the artefact leads to a preferred new situation or new processes, and what the benefits of this new situation are, *as a consequence* of (the use of) the artefact. So here we have to corroborate a causal relationship between the (use of the) artefact on the one hand, and the positive and negative results and affects of this use on the other. For this causal principle we need a randomised experiment. Next best is a correlational design in which we keep constant suspect variables that may bias a causal conclusion or which we analyse by calculating partial correlations. Still another possibility is a case study design, where an intensive and qualitative study of the process of causation must test the plausibility of the causal hypothesis.

Assessment of the direct effect(s) may be followed by an evaluation after the artefact has been used for a while in its real life context, in order to assess the long-term qualities of it. This is still goal-based evaluation. However, because of rapid changing conditions and circumstances, and or in case of not operationally defined goals [G], the evaluator may be forced to carry out a *goal free* evaluation. In that case he or she resigns from the "official" design goal or problem(s) to be solved [G]. Instead he or she sets other professional criteria derived from theory or proposed by stakeholders and/or experts in the field.

6. Summary and Conclusions

Throughout this article, it became clear that a logical and empirical evaluation of the designing processes and products is an essential prerequisite for designing an artefact that will be compliant with the expectations of the stakeholders and/or with professional standards. Evaluation of the designing process asks for a highly systematic approach. Designing processes can be unravelled in a number of stages. Each stage should be evaluated with respect to designing rules and research methodology on the one hand, and to the results (product) of this stage on the other. In this article, we have made explicit much of what is often implicit in design oriented research. Furthermore, we have offered additional criteria for evaluating the designing process and its results. As to each particular stage of the designing process there appear to be design guidelines and requirements. At many points in the process empirical evaluation requires qualitative and participatory research techniques, rather than quantitative ones. This holds in particular for the process that leads to operational definitions of design requirements, for the evaluation of the way in which design guidelines were selected and for answering the question to what extent the selected design guidelines have been followed.

However, although there is a wealth of design-oriented publications, which hide many implicit methodological design guidelines, an explicit critical appraisal of generally accepted design guidelines and types of design requirements is still missing. In our view such an appraisal should be high on a research agenda of methodologists in the domain of design-oriented research. The same holds for questions such as how to attribute different weights to different design guidelines. More specifically, the designer/evaluator has to ask her or himself, whether the structural specifications make a design that comes up to the criteria [C_f], [C_u] and [C_c], as well as to the assumptions [A_u] and [A_c]. Of course, these mainly are criteria to be checked at face value by means of logical reasoning.

Several types of evaluation appear to be relevant for the designing process. In particular, the distinction between goal-based and goal-free evaluation is important for those designs for which insufficient operationally defined design requirements can be formulated. To the extent that the goals are fully captured in design requirements, goal-based evaluation is essentially empirical requirement testing. However, in much design oriented research the proof that the design requirements are a correct reformulation of the goals is not trivial. One of the reasons is that the design requirements are often much more detailed than the goals. Given a goal [G] mostly more than one set of design requirements and structural specifications are possible.

Designers/evaluators should be well aware that exogenous requirements [R] and structural specifications [S] must be unravelled in different dimensions and aspects, and from these operationally well defined criteria must be derived. As long as we do not succeed in defining such criteria, adequate formative evaluation will either involve expert reviews or will not be possible at all.

For empirical evaluation of designs normal scientific criteria should be used: i.e., validity, reliability, researcher independence and verifiability. Priority on the methodological research agenda should be how to evaluate validity of the process that leads to operationally defined design requirements and the validity of goal-free evaluation.

References

Alexander, C. (1964). *Notes on the Synthesis of Form.* Cambridge: Harvard University Press.

Alexander, C. (1979). *The timeless way of building.* New York: Oxford University Press.

Alexander, C., Ishikawa, S. et al. (1977). *A pattern Language: Towns, Buildings, Construction.* New York: Oxford University Press.

Asimow, M. (1962). *Introduction To Design.* Englewood Cliffs, NJ: Prentice-Hall.

Babbie, E. (1998). *The Practice of Social Research.* Belmont, CA: Wadsworth.

Booch, G., Rumbaugh, J. et al. (1998). *The Unified Modeling Language User Guide.* Amsberdam, Addison-Wesley.

Brown, K. A. (1988). *Inventors at Work : Interviews with 16 Notable American Inventors.*
Redmond, Washington: Tempus Books of Microsoft Press.

Chandrasekaran, B. (1990). Design problem solving: a task analysis. *AI-Magazine* 11: 59–71.

Chen, P. P.-S. (1976). The entity-relationship model-toward a unified view of data. *ACM TODS* 1.

Costa, A. I. A., Dekker, M. et al. (2001). Quality function deployment in the food industry: a review. *Trends in Food Science Technology* 11: 306–314.

Creswell, J. W. (1994). *Research Design: Qualitative and Quantitative Approaches.* Thousand Oaks: Sage.

Cross, N. (2000). *Engineering Design Methods: Strategies for Product Design.* Chichester: Wiley.

Csikszentmihalyi, M. (1996). *Creativity: Flow and the Psychology of Discovery and Invention.* New York: HarperCollins Publishers.

Dasgupta, S. (1991). *Design Theory and Computer Science.* Cambridge, UK: Cambridge University Press.

Dasgupta, S. (1991). *The Theory of Plausible Designs. D. T. a. C. Science.* Cambridge, UK: Cambridge University Press.

Dasgupta, S. (1996). *Technology and creativity*. New York: Oxford University Press.

Denzin, N. K. & Lincoln Y. S. (1994). *Handbook of Qualitative Research*. Thousand Oaks: Sage.

Gamma, E., Helm, R. et al. (1994). *Design Patterns: Elements of Reusable Object-Oriented Software*. Amsterdam: Addison and Wesley.

Halpin, T. (1995). *Conceptual Schema and Relational Database Design*. Sydney: Prentice-Hall.

Hauser, J. R. & Clausing D. (1988): The House of quality. *Harvard Buisiness Review* 1988 63–73.

Mohr, L. B. (1995). *Impact Analysis for Program Evaluation*. London: Sage.

Nonaka, I. & Takeuchi, H. (1995). *The Knowledge-Creating Company*. New York: Oxford University Press.

Patton, M. Q. (1997). *Utilization-Focussed Evaluation. The New Century Text.* London: Sage.

Pawson, R. & Tiley, N. (1997). *Realistic Evaluation*. London: Sage.

Rossi, P. H., Freeman, H. E. et al. (1998). *Evaluation: a Systematic Approach*. London: Sage.

Rumbaugh, J., Jacobson, I. et al. (1998). *The Unified Modeling Language Reference Manual*, Amsberdam, Addison and Wesley.

Schlesinger, S., Crosby, R. E. et al. (1979). Terminology for model credibility. *Simulation* 32: 103–104.

Simon, H. A. (1996). *The sciences of the artificial*. Cambridge, MA: MIT Press.

Stapleton, J. (1997). *DSDM, Dynamic Systems Development Method: The Method in Practice*. Harlow, England ; Reading, MA: Addison-Wesley.

Verschuren, P., Somers, N. et al. (1997). The need for qualitative methods in agricultural economic research. *Tijdschrift voor Sociaal-Wetenschappelijk Onderzoek van de Landbouw. (Gent, België)* Jrg. 12(4): 367–376.

Verschuren, P. & Zsolnai, L. (1998). Norms, goals and stakeholders in program evaluation. *Human Systems Management* 17: 155–160.

Verschuren, P. J. M. (2001). Holism versus reductionism in modern social science research. *Quality and Quantity* 35(4): 389–405.

Verschuren, P. J. M. (2003). Case study as a research strategy: some ambiguities and opportunities. *International Journal of Social Science Methodology* 6(12): 121–139.

Vincenti, W. G. (1990). *What Engineers Know and How They Know It*. London: The Johns Hopkins Press.

Yin, R. K. (1984). *Case Study Research: Design and Methods.* London: Sage.

Yin, R. K. (1994). *Case Study Research : Design and Methods*. Thousand Oaks: Sage Publications.