The methodology of man-machine systems: Problems of verification and validation

Forskningscenter Risø, Roskilde

Publication date:
1981

Document Version
Publisher's PDF, also known as Version of record

RISØ-M-2313

THE METHODOLOGY OF MAN-MACHINE SYSTEMS: PROBLEMS OF VERIFICATION AND VALIDATION

Erik Hollnagel

Abstract. This paper provides an elementary discussion of the problems of verification and validation in the context of the empirical evaluation of designs for man-machine systems. After a definition of the basic terms, a breakdown of the major parts of the process of evaluation is given, with the purpose of indicating where problems may occur. This is followed by a discussion of verification and validation, as two distinct concepts. Finally, some of the practical problems of ascertaining validity are discussed. The general conclusion is that rather than rely blindly on a well-established procedure or rule, one should pay attention to the meaningfulness of the aspects which are selected for observation, and the degree of systematism of the methods of observation and analysis. A qualitative approach is thus seen as complementary to a quantitative approach, rather than antithetical to it.

INIS Descriptors: MAN-MACHINE SYSTEMS; NUCLEAR POWER PLANTS; PLANNING.

UDC 65.015.1 : 621.039.56

October 1981

Risø National Laboratory, DK 4000 Roskilde, Denmark

ISBN 87-550-0798-8
ISSN 0418-6435

Risø Repro 1981

TABLE OF CONTENTS

INTRODUCTION

The complexity of the man-machine systems (MMSs) employed in the control of nuclear plants and similar intricate systems, and the gradual realization of the complexity of the events which may occur in such systems, has led to an increase in the number of designs that aim to insure the safe functioning of such systems. The many suggestions for improvements of existing MMSs must, however, be assessed in some way before they are actually introduced. One has in advance to know, or at least to feel reasonably certain, that the new design will function as planned, in the sense that the expected consequences for the functioning of the MMS will actually occur. This has led to a substantial increase in the research in MMSs in general, and in the development of methods for observation, measurement, and analysis in particular.

There is, however, as yet only little in the way of a systematic methodology for accomplishing the assessment of a design, and there seems also to be a lack of clear concepts in this area, as well as a confusion in the use of terminology. The reason for this state of affairs is probably that the assessment of a design for an MMS is an undertaking which is neither pure engineering nor pure behavioral science, but rather a mixture of both. The mixed nature of an MMS as well as the sheer complexity of it limits the usefulness of the traditional methods from natural and behavioral science. In the assessment of designs for MMSs one can, for instance, normally not use control groups or the traditional parametric method of varying one variable at a time while keeping others constant. Thus the established principles for experimental design (e.g. Campbell & Stanley, 1963) are of little direct avail.

The present paper has the purpose of clarifying some of the concepts and principles which play a part in the assessment of designs, in particular the aspects of verification and validation. It purports to discuss some of the essential aspects of

verification and validation, and to identify the major methodological problems which derive from them. There will, however, not be given any fixed and final solution to the problems. First of all, the problems are not such that clear-cut solutions should be given on a general basis. And secondly, the important point is precisely to be aware of the problems, before one begins to solve them. What is presented here is thus a list of things to keep in mind when making an assessment of a design, rather than a set of instructions for doing it. The present report is based (and partly biased) by the experience gained during the joint Scandinavian project on Control Room Design, which was carried out in 1978-1981. A number of reports describing this project, referred to as the NKA/KRU project, may be found in NKA/KRU-(81)14. Some previous reports which go more into detail with the problems of verification and validation have already been published (Note 1, Note 2, and Hollnagel 1981).

A DEFINITION OF THE BASIC TERMS

Since the terminology used for MMSs is a mixture taken from a number of different sciences (such as engineering, systems theory, cybernetics, psychology, etc.), it is necessary to begin by defining the way in which the terms are used here, as this may prevent misunderstandings later on. Hopefully, the definitions are not too different from the reader's normal use of the terms.

We shall use the term <u>design</u> to indicate any particular arrangement of a part of an MMS, in contrast the whole of the MMS, which has the purpose of shaping and improving the performance of the MMS, either by preventing the occurrence of some activities or by facilitating others. The design may relate to a physical part of the MMS, such as a panel or a

group of instruments, or it may be a part of the system software, the procedures, etc. Implementing the design in the system is normally expected to bring about a specified change in the performance of the system. If the specified change is actually obtained, we may say that the design was correct. If not, it was obviously incorrect or wrong. (I am deliberately talking about the performance of the system rather than the performance of either the man/men or the machine(s), because the plant policy normally is concerned with the output of the MMS, i.e. the way in which it functions as a whole.)

The implementation of the design normally takes place in a number of steps. Thus the design may be tested in various versions and stages of development, before it is finally implemented. It is this test which we refer to as the experimental evaluation. There may actually be a number of tests, since the design develops gradually rather than emerging in its final form out of nothing. The whole process of getting ideas for designs and gradually filtering the good ideas from the bad ones is in itself extremely interesting. However, we cannot go into it here, but may just mention that it can be described in analogy with the normal process of generating or producing scientific hypotheses. In many cases, particular in relation to nuclear plants, the impetus for the design lies in some event or some demand for a new and improved design. This also means that some of the criteria which can be used to separate good designs from bad designs are given from the start.

We may mention in passing that the whole process of developing a design should be seen not only from the technical but also from the organizational or sociological point of view, since it is not always the ideal, rational process that the scientific description assumes. From a theoretical point of view such considerations as economy, pressure from special interest groups, commitments, attitudes and prejudices, etc., are normally excluded. But in actual fact they may have a significant influence, as anyone who has tried to work outside a scientific laboratory can testify. We shall, however, refrain from enter-

ing into this here, but only point out that this kind of description, i.e. the organizational one, cannot be disregarded when the complete picture has to be given.

The experimental evaluation itself accentuates some of the essential problems. One of them is how the evaluation or testing is carried out. Another is the relation between the results of the test and the results one expects from the real-life setting. It is the latter which properly speaking is the problem of validity. We shall return to that in a later section. For the present let us start by taking a closer look at the problem of testing as such.

THE COMPONENTS OF EVALUATION

Evaluating a design means that it is implemented in some model system instead of in the real-life object system. The model system thus represents the essential aspects of the object system, either by being a copy of it on a smaller and less detailed scale, or by being a representation of it in a different domain, e.g. a mathematical model of a physical process. The results of using the design in the model system are then compared with something else, and it is the result of this comparison which is the core of the evaluation. Making the comparison points directly to the following three important aspects of the evaluation:

1) How do we determine what the results of the implementation are, i.e. how do we observe, record, or measure them?

2) How do we determine the "something else", i.e. how do we specify the requirements or the criteria for comparison?

3) Finally, how do we make the comparison, i.e. is there a specific method or set of rules, which can be used?

Determining the Results of the Implementation

The first aspect is concerned with finding the proper indicators for the results of the implementation, to be used in the comparison. If, for instance, the purpose of the design was to increase the speed of certain activities, we could simply measure the appropriate response og reaction times, since they would be the needed indicators. If the purpose was to reduce the level of stress, we might find some suitable physiological measure of stress and then use that. But even this example makes it clear what the difficulties are. In many cases it is not easy to find suitable indicators and to measure them in a reasonable way. Since the definition of the indicators is based mainly on the definition of the criteria, cf. the second point below, the crucial question becomes how and to what extent an independent assessment of the indicators can be made.

It is generally considered desirable to have some kind of measurement in the traditional sense, but one may often have to refrain from that and rely rather on some other kind of observation or systematic description of the performance. Yet whether the assessment is a quantitative measurement as we know it from the natural sciences, or another kind of observation, is of minor importance, as long as the assessment is made systematically, using a set of well-defined categories. The widespread belief in the superior quality of assessments given by means of number is mistaken, as anyone with a good knowledge of scientific methodology will readily acknowledge. Thus we need not be concerned with that aspect here. The really crucial point is that one knows what the indicators stand for, i.e. that one is able to interpret them in a meaningful, consistent, and unequivocal way. So, in the case of an experimental evaluation, it is important that a systematic way of observing, recording, and assessing the chosen indicators is developed or found.

Defining the Criteria for Comparison

Concerning the specification of the criteria, this is important in two ways, as discussed above. One is that it is significant for determining the indicators to be used. Another is that it is significant for the outcome of the comparison. However, the criteria may be expressed in an apparently clear way and yet be hard to operationalize. Thus, for instance, one criterion may be that the operators no longer make any incorrect diagnoses. That seems clear enough in the sense that we can easily understand what it means. But how does one use that in an evaluation? One problem is that it may be difficult to say in an absolute way whether a diagnosis is correct or incorrect. Another that it may even be difficult to say what constitutes a diagnosis and what does not. In many cases the criteria are vaguely given, simply because the purpose of the design, i.e. the expected result, is vaguely given. It may be to enhance the operators' decision making, or to reduce the mental load; to improve visual discriminability, or to reduce mistakes in the use of buttons, etc.; it may be to prevent spatial misorientation on the panels, or to increase the legibility and comprehensibility of the procedures.

Similar examples are easy to find. In each case the criteria appear clear enough, because we easily see what is meant by them. But when it comes to the point where the criteria are going to be used to define the proper indicators, and to serve as a basis for comparison, it may turn out that there are many problems. Again, a preferred solution seems to be to find some kind of measurement, and rely on that. Measurements are, of course, attractive because they are so simple and so - seemingly - objective. But their simplicity is deceptive. It is not a problem to find a measurement as such. But it may be hard to find a measurement which is meaningful, cf. the above. Finding such a measurement requires that one has access to a clear, qualitative description of what one wants to measure, i.e. that one has a good qualitative conceptual model. It is the establishing of this model which is the real problem. Once it has been established, it is of minor importance whether it

can be expressed in terms of actual measurements or not. Quite often systematic observations are far better than measurements, both to make and to use. That is certainly so in the cases where the operators' behavior plays an essential part.

The Rules for Making the Comparison

The third aspect was how the comparison was going to be made. In this case it is clearly an advantage, if the results and the criteria are present in the form of measurements, because it is easy to say whether one number is smaller or larger than another. But that is an advantage only if the measurements are simple numbers - and, of course, if the numbers can be interpreted in a meaningful way. As soon as it becomes necessary to combine numbers, or to use statistical tests, the advantages of having measurements may completely disappear. Clearly, it is only an asset to have a rule for comparison as transparent as a statistical test if it is used properly. But the general experience is that this is not so (cf. Tversky & Kahneman, 1971 or Brigham, 1974). It is a sad fact that many scientists or researchers do not know what they are doing when they are using statistical tests, but become spellbound by the magic of numbers and the myth of objectivity which is connected with numbers. This has not been reduced by the advent of computers and packages of computer programs that can produce endless pages of statistics. However, if the statistics are not used sensibly, they are damaging rather than supporting.

The lesson to be learned from this is that it is important to have a systematic method for making the comparison, and a method that one understands in the detail. Such methods may be found for all kinds of observations, whether they are verbal data or numerical measurements. They may appear less exact in the former case than in the latter, but that is not necessarily so. As long as the basis for the comparison is systematic and interpretable, and the comparison can be carried out according to a specified rule or set of rules, the result is equally reliable whether it is given by words or by digits. One should

not forget that if the result is given by digits, these will have to be interpreted verbally anyway. This means that the comparison may be carried out as a match among descriptions and observations, just as well as in any other way. The really important point is that the comparison is systematic, hence reproducible. It must be possible to treat results from similar implementations in the same way.

We have thus seen how the evaluation of a design raises some important problems - which one may find in any scientific methodology, and which are related to the very basis for having an empirical science. Put very briefly, the important point is that one knows what one is doing. Whatever observations are made, one must be sure that they are interpretable and made in a systematic way. Whatever comparisons are made, one must know and understand the rules by which they are made. And whatever results one gets, one must know what they mean, i.e. be able to interpret them consistently and unequivocally in relation to the type of situation for which the design was intended. The reliance on methods qua methods, without understanding their significance, can only lead the researcher astray. In order to be able to make an experimental evaluation, one must have an adequate qualitative model of the phenomenon under investigation. It is this model which guides the interpretation at all levels, hence insures the meaningfulness of the results.

VERIFICATION AS THE DEVELOPMENT OF A DESIGN

Apart from the problems in the process of evaluating a design, as we saw above, one must also consider the purpose or function of evaluating a design. Therefore, in addition to the how one should also speculate about the why. These two aspects are, of course, interrelated, since specifying the purpose for doing something may often have ramifications for how one can go about doing it.

It is here that one must make a distinction between verification and validation. Finding and developing a design is very much like solving an unfamiliar problem. The solution that first comes to mind may not be a usable one, and a proper solution is only gradually developed. However, there is a difference from problem solving because in the development of a design the "problem" may be only vaguely stated (corresponding to the so-called ill-defined problems), and the criteria for evaluating the solution may be similarly uncertain. There is therefore no assurance that the solution that is acceptable is also the best solution, or even among the best. But the longer the process of evaluation is carried out, i.e. the longer the period in which the design is developed, hence also the more suggestions for design that are developed, the greater is the chance of ending up with a good design.

We cannot go into the details of how one gets an idea for a design. Certainly, a considerable influence comes from the ideas that are generally floating around, so to speak, in the research institutions which occupy themselves with the proper subject matter. Another influence is of course coming from the concrete instance or problem which has lead to the realization of the actual need for a new design. But let us just assume that we have come up with a new design, in answer to a recognized need. Obviously, one cannot implement the design straight away in the object system. It is necessary before that to make certain that the design functions according to expectations, i.e. that is it indeed a solution to the initial problem. It is this phase which is called the verification - which literally means the process of establishing the correspondence between a theory and the facts, to confirm or prove the truth of a hypothesis or a theory, i.e. the assumptions which support the design.

Since the design cannot be implemented directly in the object system, it has to be evaluated in some other way. This is what we refer to as the experimental evaluation. The evaluation generally takes place by means of a model of the real system, whether it is in the form of a computer simulator or in the

form of a small-scale experimental system. The model system is a copy of the object system, but with a smaller amount of details. This is, of course, very important. The details that are left out are excluded because they are considered to be of little significance for the problem. But obviously, if that is not correct, the evaluation suffers, and may even become completely useless and misleading. We shall return to this aspect when we discuss the problem of validation. In the case of the verification, we are simply interested in finding out whether the results of the experimental evaluation correspond to the hypotheses or ideas behind the design. If they do so, we say that the design has been verified. If not, we may either say that the design has not been verified or - if we have a strong trust in it - that there has been some flaw in the experimental evaluation, so that it should be repeated in other, and better controlled, conditions. (Note, by the way, that if the design is verified, one is less inclined to consider whether it was due to fortuitousness, although that may play an equally important role in the positive and the negative instance.)

The verification is thus the process by which the good designs are filtered out from the bad designs, based on the degree to which the results match the expectations. It is, in other words, the test-bed for the ideas and the method for weeding out the incorrect ideas from the correct. We have already considered the more formal aspects of this testing. The purpose obviously is to reduce the number of ideas or designs that must be seriously considered and perhaps implemented, so that the implementation will only take place for those designs that actually work. A complete success can, of course, not be guaranteed. But mistakes should be avoided as far as possible, mainly because the costs (in nuclear systems) can be tremendous. It therefore seems reasonable to have a gradual process of verification, perhaps using model systems that are more and more realistic, and which contain more and more details. A simple conceptual experiment may be sufficient to discard some designs, while others show their deficiencies only in a test on a full-scale simulator. The quality of the evaluation is

obviously increased the more detailed the model system is, but so is the cost. The determination of how the experimental evaluation shall take place is therefore very much a result of a trade-off between these two factors, hence representing a practical compromise rather than a scientific ideal.

The basic steps in the verification have been shown in Figure 1, which is based on the previous discussion of the components of the evaluation. Figure 1 suggests the basic nature of the verification, rather than the concept of repeated verification at different levels of complexity.

VALIDATION AS THE TEST AND PROOF OF DESIGNS

The obvious question is, of course, whether a design which has been verified has thereby also been validated. And the answer to that is clearly a no. That is obvious simply from the condition mentioned above that the verification takes place by means of a model system which has a reduced number of details compared with the real system. Since one cannot know with absolute certainty whether any of the details that have been left out is crucial, one cannot directly transfer the results from the experimental evaluation to the real-life situation. One can, rather, be certain that there always will be a difference between an experimental evaluation and the real-life situation, even in the case of a full-scale replica simulator. The mere knowledge of the persons who participate, that they take part in an experiment - and there is no way in which they can be fooled not to believe that - will affect the result.

Validation is, technically speaking, the extent to which the results from the experimental evaluation correspond to the results from the object system, i.e. from the real life. It is thus actually a question of the validity of the experimental evaluation. If it is valid, there will be no difference between

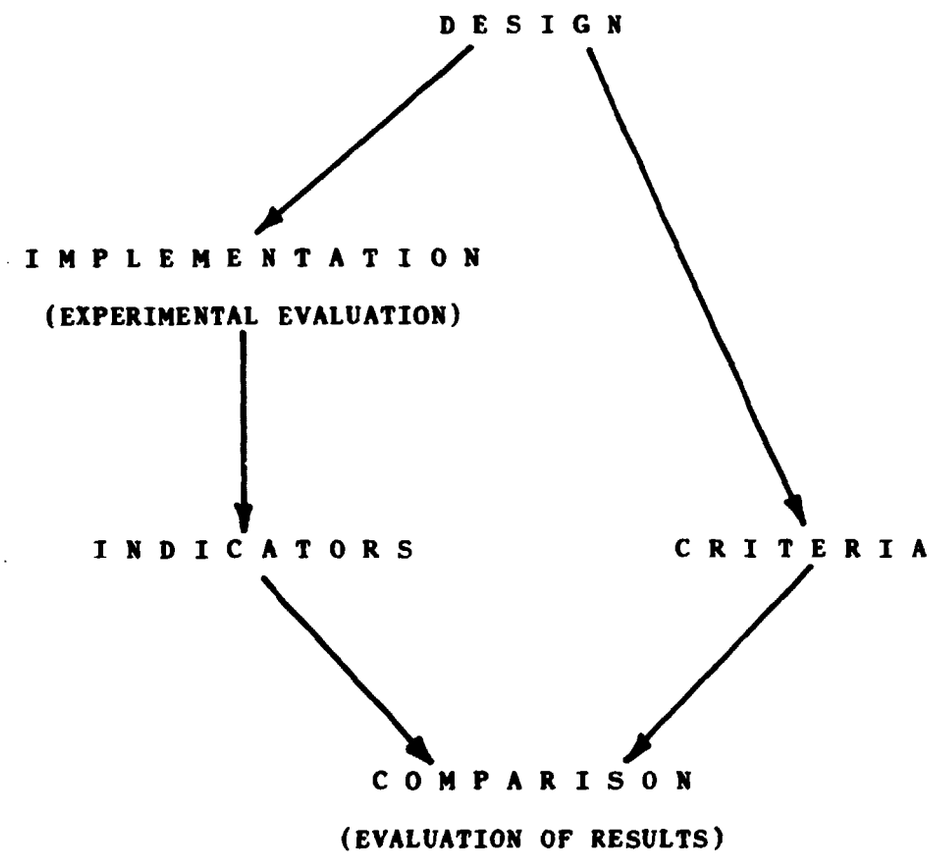

Fig. 1. The basic steps in the verification of a design.

the results from the two types of situations, at least no difference that is relevant with respect to the consequences of the design. If the validity is less than perfect, there will be differences. And the validity is always less than perfect, barring the unlikely case where the two systems are identical, i.e. one and the same. The problem is therefore whether the differences are sufficiently small to be considered insignificant. Or whether they are so large that the design is useless in practice, despite the fact that it has been verified.

Types of Validity

There are, however, more than one type of validity. We need not go into the technical details here, but note only three different types of validity, which all are relevant for the consideration of experimental evaluations.

The first type of validity is content validity. This refers to whether the content of the model system corresponds to the domain it is intended to model or measure. A traditional example is that a test for arithmetic performance should consist of arithmetic tasks or problems rather than e.g. reading problems. In so simple an example, content validity is obvious and easy to ascertain. However, it becomes less obvious when the object system and the performance is more diffuse, as it well may be in the case of an MMS. Still, content validity is usually taken for granted although that may be unjustified. And if more explicit attempts of establishing content validity are made, they are based on conceptual arguments rather than measurements. It is quite often a question of whether the test situation appears to be convincing to the critical observer, i.e. a sort of subjective assessment. Establishing a measurement for content validity in a complex system would obviously run into the same type of fundamental problems that we have for the whole concept of validity and measurement as such.

The second type of validity is construct validity. This is a more sophisticated form of content validity. It requires a more

precise definition of the construct, i.e. the aspect of behavior that one wants to study or which the design is expected to affect. If, for instance, we are concerned with operator decision making, then we must be able to specify precisely what types of behavior represent decision making, and which do not. Next we must find some suitable measure of the extent to which each test item or part of the test, as well as the test as a whole, corresponds to or correlates with the construct. This measure then is an expression of the construct validity. It is thus a statistical type of validity, in contrast to content validity. Yet it is easy to see that they are similar in kind. And since in the case of an MMS there will more often be talk of content validity than of construct validity - simply because it may be extremely difficult to find a sensible or reasonable definition and measurement of the construct - we shall concatenate the two and in the following speak only of content validity. But it must be noted that the better specified the circumstances are, i.e. the more specific the design is, the easier it may be to use the construct validity in a meaningful way. So the concept is by no means discarded, but rather to be considered as one end of a continuum.

The third type of validity which we shall be concerned with here is the empirical validity. This is, as the name indicates, the degree to which the results from the experimental evaluation corresponds to the results from the real-life application. It is often a specific relation between two sets of measurements, hence a statistical or quantitative type of validity. The external measurement is often called the criterion, and the name criterion-related validity may be used. Put differently, it is the extent to which the results of the evaluation corresponds to the empirical facts. In the case of an MMS using a simulator, it is the extent to which the results of implementing the design in the simulator corresponds to the results of implementing the design in the real-life system. It is, obviously, this last type of validity which corresponds to what people have in mind, when they just talk about validity in an unspecified way.

We can show how the relation between validity and verification is by Figure 2, which also tries to make clear the difference between content validity and empirical validity. It should come as no surprise that verification, content validity, and empirical validity each presents its own problems. And also that it is quite important to make a distinction between them, since that has implications for how one plans to go about establishing either. Certainly, one should not make the mistake of improving the content validity when it is the empirical validity that is needed, or vice versa.

PROBLEMS OF CONTENT VALIDITY

Content validity and empirical validity both have importance for experimental evaluations involving MMSs, in particular where simulators play a role. We shall therefore consider them both, and the order in which they are mentioned is in no way an indication of their relative importance. There are two aspects which are especially relevant for the problem of content validity. The one concerns the relation between simulators and the real-life situations, i.e. the nuclear plants. And the other involves the question of the complexity of the systems.

Content validity was the question of whether the model system corresponded to the object system - or rather the degree of correspondence, whether it was merely like it, quite similar to it, or a replica of it. The ultimate likeness will of course be in the case where the object system can be used as a model system, i.e. when it could be made available for the purpose of the experimental evaluation. Note, however, that even in this case there would be a difference, unless one could make the persons conducting and participating in the test believe that it was a real-life situation rather than "just" a test. This would correspond to the traditional method of double-blind

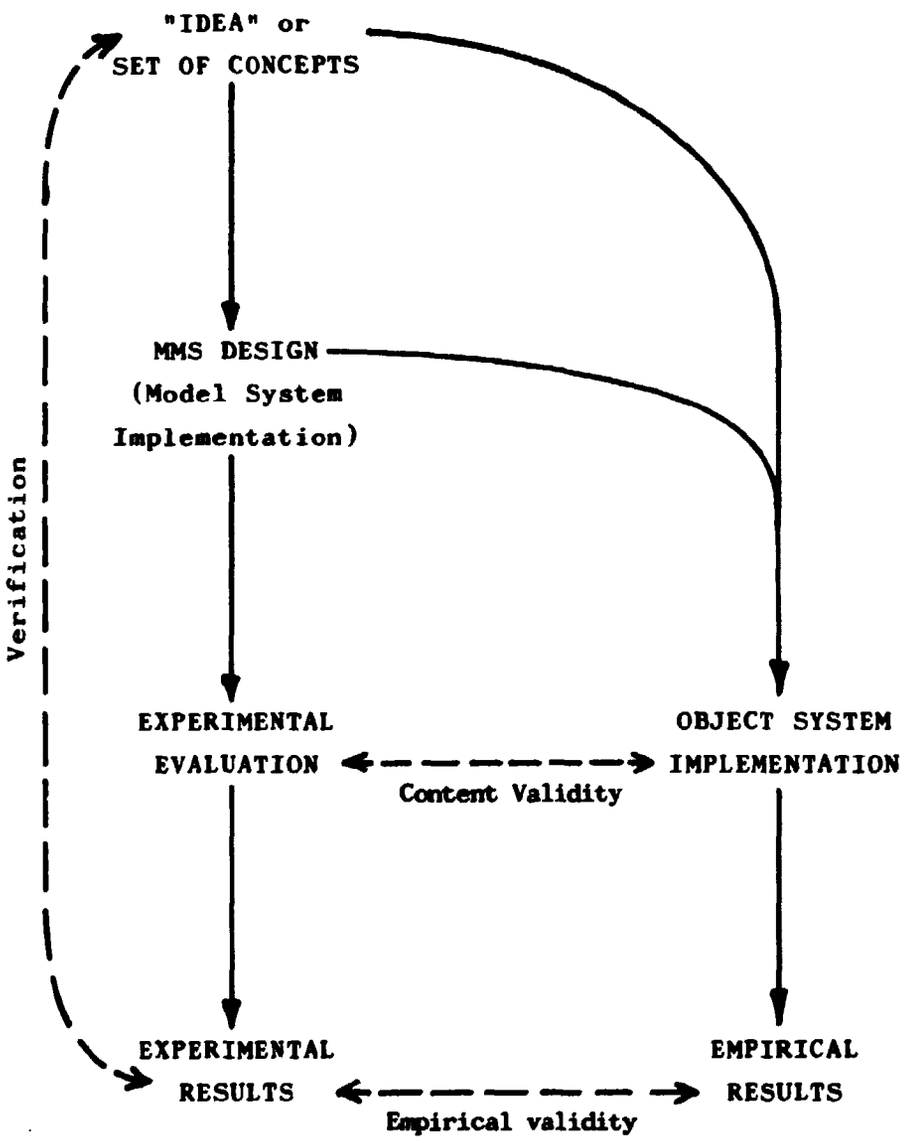

Fig. 2. The relation between validity and verification in experimental evaluation.

tests known from other areas. But it would still be concerned only with the content validity.

In the case of experimental evaluations of designs for nuclear reactors, the commonly used vehicle is a simulator, either a full-scale simulator made available for the purpose, or a specially designed experimental simulator. These represent different sources of data, with different methodological problems (cf. e.g. Hollnagel & Rasmussen, 1981). But the essential point is that in neither case is the simulator ever a true replica of the object system. This means that one has to consider seriously to what extent the content validity of the experimental evaluation is assured. One problem is the "feel" of the simulator, i.e. whether it responds in the same way as the plant. An operator quickly develops a "feeling" for the process - a sort of personal knowledge or experience of how the system acts and reacts to him. This process feeling may change with experience, but even a moderately experienced operator will have some kind of "process feeling". (Note, however, that it need not be correct in an objective sense.) And it is definitely important that the simulator matches this process feeling. Otherwise the operator will have to develop a special process feeling for the simulator, which naturally reduces the content validity. Another aspect is the degree to which the task in question can be realistically implemented, i.e. whether the operators can be assumed to take the task seriously. If, for instance, the design concerns an alarm handling system, it is quite essential that the operators take the task seriously, and respond to the alarms with the same kind of seriousness that they would do in real life. If they simply react to them as artificial disturbances, and not as representing anything serious for them and the system, the content validity of the test will probably not be very good. However, it is fortunately rather easy to get the operators to take the task seriously, since a sense of personal prestige may well take the place of the responsibility of a real plant. But it does mean that one should give this ample consideration, i.e. consider the psychological side of the evaluation, and not just the technical side.

The content validity can, of course, never be perfect, because a simulator - even a replica simulator - is different from the real plant. But I believe that an acceptable degree of content validity can be ascertained, if one tries to do so. Since the content validity essentially corresponds to the subjective experience of the task, and of the difficulty of it, it very much depends on the way in which the operators are prepared for the test, i.e. informed, trained, instructed, etc. The question of providing adequate operator training may be crucial indeed, cf. Note 2. All in all it means that the problem of the content validity of the simulator is not just a problem of having a technically perfect simulator, with all the buttons and lamps in the right place, but also (and perhaps even more) to ensure a trustworthiness of the human side, i.e. the social and psychological aspects of the simulator. It is, after all, the content validity of the MMS - the operators and the technical system - which is essential, and not just the technical side. The latter appears to be given an undue amount of emphasis, probably because it is easier to handle and measure than the human side. But technical fidelity alone will never be sufficient.

Another aspect of the content validity is concerned with the degree of complexity of the systems. Since the number of parameters that are necessary to describe a nuclear plant is very large, and since the response of the plant in every conceivable situation is beyond reach, one cannot hope ever to grasp and emulate the full complexity of it. It belongs to what Beer has termed the exceedingly complex systems (Beer, 1964). It seems sensible therefore to ask whether it is necessary to try to emulate the complexity. The argument against it is that from the psychological point of view, the performance of the operator may be described by means of a relatively simple set of strategies or activities, that are combined according to the demands of the situation, and that the apparent complexity of the performance is a reflection of the complexity of the environment, rather than the complexity of the psychological "mechanisms" of the human, as suggested by Herbert Simon (Simon, 1969).

This view is correct to the extent that it concerns the description of the performance, leaving out for the moment the hypothetical psychological "mechanisms" that may be used to explain the activities. The performance of an operator in a nuclear plant and the performance of an experimental subject in a far more simple situation may be described in the same technical language and by means of the same psychological concepts. In other words, MMSs at different levels of complexity may be described by the same terms and the same principles. Thus, making a diagnosis in a simple problem and in a real plant may be described on a general level, cf. e.g. Rasmussen (1978). And this carries over to other instances as well, e.g. so that some of the training, the teaching of the basic principles, may be carried out just as good (if not better) on a simplified system as on the real system. From this point of view then, the complexity of the object system need not be reproduced in the model system. The problem of content validity does not rest on that, but rather on the possibility of finding the same types of activity, i.e. the same types of performance in the subjects on the object system as on the model system. This makes establishing the content validity altogether an empirical matter. If it turns out that the types of performance elicited by the two types of system are psychologically identical (in the sense that they conform to the same generic description of the behavior), then we may say that from this point of view the content validity is acceptable. Or at least negatively, that if this condition is not fulfilled, then the content validity is clearly not acceptable. On the other hand, the content validity cannot be established analytically, i.e. in advance of actually making a test. But since the model system and the object system are relatively fixed, a content validity once established can be used in future tests. So the analytical difficulties are not really serious.

This view of content validity is, in a way, the other side of the coin, in relation to the technical identity of the systems. We claimed that the technical identity was insufficient to guarantee a satisfactory content validity of the MMS. A

technical identity only gives a face validity which often is deceptive, and hides the important problems. The discussion here of the psychological aspects of the content validity supports that earlier conclusion. Yet it should also be quite clear that the psychological identity of performances is insufficient by itself. For one thing it depends very much on the level on which the descriptions are given. They may be on so general a level that playing chess is seen as identical to handling a LOCA. But that is clearly of little interest with respect to the experimental evaluation and the problem of content validity. Though the formal aspects of the two different performances may be similar or even identical, it is easy to show that a description on a slightly more detailed level will reveal more differences than likenesses. It therefore seems reasonable to demand at least a likeness on the technical level, i.e. that the tasks are formally and realistically the same. The physical manifestation of the systems may be different, varying from a replica to a computer VDU. Taken together this means that the content validity will have to depend upon technical as well as psychological considerations, and that neither alone is sufficient. This is just another way of emphasizing that an MMS is precisely what the name says: A Man and Machine System, requiring due consideration of the factors that influence each part as well as the parts as functioning together. Neither pure technology nor pure psychology will suffice.

PROBLEMS OF EMPIRICAL VALIDITY

When we turn to empirical validity, we find also here a number of problems. The first of these is that it may be difficult to identify the empirical basis, i.e. the empirical data which function as the criterion for the results from the experimental evaluation. And the other that it may be difficult to define the appropriate measurements.

The difficulty of finding the empirical basis is largely due to the nature of the domain, i.e. the functioning of nuclear plants. For those designs which are intended for normal operation, the problem may be quite small. But it is precisely the designs intended for the off-normal situations that attract the most attention, since they concern situations with grave potential consequences, economic as well as social. However, because these serious off-normal cases are very rare, the appropriate empirical basis is obviously difficult to establish. In certain cases it may even be so that one hopes that the appropriate empirical basis never materializes. Although that is fortunate from several points of view, it is most unfortunate from the point of view of the empirical validity of the design, since there is then no way in which it can be established. One might, perhaps, talk about an expected empirical validity, which would refer to the expected consequences of a given scenario. But that would, of course, be no better than the assumptions on which it was based, hence provide nothing that was not already present in the case of content validity.

The most evident solution in this case is to use a simulator or a kind of game, and evaluate the design in the simulated situation. This will provide what we could call a simulated empirical validity. Although this can provide a good indication of what the empirical validity might be in the real case, there is a number of problems connected to it. First of all, is, of course, that any simulation will be reduced with respect to the number of details in comparison with the object system. And that has as a further consequence that the more unlikely the situation is which is simulated, the less certain may one be of the trustworthiness of the simulation. Even for a full-scale training simulator, designed as a replica simulator, there are many uncertainties concerned with transients that are a bit unusual. The proper functioning of the simulator is to a large extent based on a calibration with empirical data, both for normal conditions and for transients. But it is obvious, then, that there can be no calibration worth mentioning for situations for which there are no empirical data, hence not for the off-normal situations that the design may be intended for. This

means that even in the case of the simulator, the uncertainty of the correctness of the simulator's functioning makes the validity dubious. It is a simulated empirical validity, and should under no circumstances be mistaken for the proper empirical validity. The best way to avoid simulated empirical validity is, perhaps, not to use it at all, but rather to remain satisfied with the content validity and the known limitations of that.

The second aspect, which is related to this, is the difficulty of finding appropriate measurements on which the empirical validity can be based. This is the case both for the very complex and for relatively simple, normal cases. It is not that it may be a problem to find some measurement and to carry out the required calculations. The problem is rather to find a measurement that is inherently meaningful, hence sensible to use.

This is, of course, related to the previously mentioned problem of defining the purpose of the design. Take, for instance, an alarm handling system. The purpose of that may be to improve the presentation of the alarms in order to reduce the number of operator mistakes, false detections, or misses. But even though this description of the purpose may be sufficient for a general characterization, how does one go about finding a measurement that adequately captures it? Such a measurement is, however, required if the empirical validity shall be calculated. This is, of course, the same problem that we saw in the case of the evaluation and verification of the design. Only now it is emphasized even more. It must not only be a proper measurement, but also one that can easily be made or obtained under normal working conditions. And this, clearly, is a very difficult restriction, which makes a traditional empirical validity very hard, if not impossible, to use.

CONCLUDING REMARKS

We have now taken a look at some of the major problems related to establishing the validity of a design, whether it be the content validity or the empirical validity. The nature of the problems indicates that in many cases it is necessary to use only the content validity, since the empirical validity is beyond reach, at least if a sensible measurement is required. This, of course, poses the problem of whether validation is possible at all, and if so in what sense.

Even if it may be difficult to establish the validity of a design, it is the responsibility of the designer and the researcher to do so, both because that is a basic part of empirical science and because of the serious consequences that the design may have. The difficulty of establishing the validity is to a great extent dependent upon the type of situation for which the design is intended. The more specific the design is, i.e. the better defined its expected consequences are, the easier it is to evaluate the validity, content validity as well as empirical validity. A design which through a change in a display system has the purpose of reducing the visual fatigue of the operators is clearly quite easy to assess and validate. It is furthermore so that the empirical basis for the validity is easier to establish and to measure for the more restricted designs. So there may be many cases where the establishing of the validity is no major problem. The reader should thus not despair over the problems presented here. But it is essential to realize that one cannot refer to the concept of validity in an indiscriminate way. The problem of validity must be assessed separately for each case, and no standardized answer can be given to questions in this respect. Most important of all is perhaps that one cannot simply transplant methods used for establishing validity in well-defined environments to the type of complex MMSs that we are talking about here. The <u>principles</u> behind the establishing of validity may be used, but as a basis for separate considerations rather than as a mechanical rule.

The question of validity touches upon the question of measurement of performance of such. Measurements of this kind are used not only in connection with validity, but also as parts of the various ways in which the performance of the system can be assessed, whether it is the total system or the individual parts (the man or the machine) of it. This is, of course, the problem that traditionally is known as quantification versus qualification, i.e. whether one should use a quantitative measurement or a qualitative description. Put thus simply the answer, of course, must be neither. It is not a question of using one method or the other, but rather of realizing the complexity of the situation, and to find or develop the measures which are appropriate for that. As I have stated repeatedly in this paper, making a measurement requires that one knows what one measures, and that the measurement can be interpreted in a consistent and sensible way. One must, of course, always start with a description of the system under consideration. This description must identify the characteristics of the system, its parts, structure and function, as well as the environmental constraints, etc. Such a description is necessarily qualitative, in the sense that it describes the qualities of the system (although that does not exclude the use of quantities as well). But it is a mistake to believe that this is just a necessary evil, and that it should be translated into a quantitative description instead. Rather, quantifications should be made on the basis of the qualitative description or model, and the two should never be separated. A quantification, or a quantitative model, can only make sense if it is related to a qualitative description or model of the system in question. That goes for the definition of the measurement, as well as the ensuing analysis and interpretation of it. It is, after all, the qualities of the system that are measured in some way, directly or indirectly.

The discussion of the advantages of the qualitative and the quantitative approach, respectively, is, however, an issue which deserves a separate treatment. I hope that the present discussion of the concepts and principles which are part of making an experimental evaluation has made it clear what the

major problems are, and that they deserve serious consideration in each particular case. And although no concrete solutions have been suggested, the reader has hopefully been convinced that solutions are possible, if only the psychological and methodological aspects of making new designs are given as much attention as the technical ones.

REFERENCE NOTES

Note 1. Hollnagel, E., 1981, On the Validity of Simulator Studies: Problems and Preliminary Precepts. Roskilde, Denmark: Risø, Electronics Department N-39-80 (NKA/KRU-P2(80)33).

Note 2. Hollnagel, E., 1981, Experimental Evaluation and Validation. Roskilde, Denmark: Risø, Electronics Department N-20-81 (NKA/KRU-P2(81)40).

REFERENCES

Beer, S., 1964, Cybernetics and Management. New York: Science Editions Inc.

Brigham, F. R., 1974, A Critical Note on the Application of Analysis of Variance Techniques in Ergonomics Research. Ergonomics, 1974, 17(2), 259-265.

Campbell, D. T. and J. C. Stanley, 1963, Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally.

Hollnagel, E., 1981, Verification and Validation in the Experimental Evaluation of New Designs for Man-Machine Systems (Risø-M-2300). Risø National Laboratory, Roskilde, Denmark: Electronics Department.

Hollnagel, E. and J. Rasmussen, 1981, Simulator Training Analysis. A Proposal for Combined Trainee Debriefing and Performance Data Collection in Training Simulators (Risø-M-2301). Risø National Laboratory, Roskilde, Denmark: Electronics Department. (NKA/KRU-P2(81)38.)

NKA/KRU-(81)14 Enlarged Nordic Cooperative Program on Nuclear Safety. NKA/KRU Project on Operator Training, Control Room Design and Human Reliability: Publications List.

Rasmussen, J., 1978, Notes on Diagnostic Strategies in Process Plant Environment (Risø-M-1983). Risø National Laboratory, Roskilde, Denmark: Electronics Department.

Simon, H. A., 1969, The Sciences of the Artificial. Cambridge, Massachusetts: M.I.T. Press.

Tversky, A. and D. Kahneman, 1971, Belief in the Law of Small Numbers. Psychological Bulletin, 1971, 76(2), 105-110.

Risø National Laboratory **Risø-M-** 2313

Risø-M- 2313

Title and author(s)

The Methodology of Man-Machine Systems: Problems of Verification and Validation

Erik Hollnagel

Date October 1981

Department or group

Electronics

Group's own registration number(s)

NKA/LIT-3(81)106

R-13-81

EH/bs

30 pages + tables + illustrations

Abstract

This paper provides an elementary discussion of the problems of verification and validation in the context of the empirical evaluation of designs for man-machine systems. After a definition of the basic terms, a breakdown of the major parts of the process of evaluation is given, with the purpose of indicating where problems may occur. This is followed by a discussion of verification and validation, as two distinct concepts. Finally, some of the practical problems of ascertaining validity are discussed. The general conclusion is that rather than rely blindly on a well-established procedure or rule, one should pay attention to the meaningfulness of the aspects which are selected for observation, and the degree of systematism of the methods of observation and analysis. A qualitative approach is thus seen as complementary to a quantitative approach, rather than antithetical to it.

Copies to

Available on request from Risø Library, Risø National Laboratory (Risø Bibliotek), Forsøgsanlæg Risø), DK-4000 Roskilde, Denmark
Telephone: (03) 37 12 12, ext. 2262. Telex: 43116