ANALYSIS OF SYSTEMS OF CONSTRUCTED FACILITIES

by

ANDREW C. LEMER

S.B., Massachusetts Institute of Technology
1967

M.S., Massachusetts Institute of Technology
1968

Submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology
August 1971

Signature of Author. . . . .
Department of Civil Engineering, Aug. 16, 1971

Certified by . . . .
Thesis Supervisor

Accepted by. . . . .
Chairman, Departmental Committee on Graduate Students

ABSTRACT

ANALYSIS OF SYSTEMS OF CONSTRUCTED FACILITIES

by

ANDREW C. LEMER

Submitted to the Department of Civil Engineering on August 16, 1971, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

An approach to the analysis of systems of constructed facilities is presented. This approach is intended to provide assistance in design decision, that is, in the selection of particular actions to be undertaken to shape the physical characteristics of a facility's service. The principal concept developed is that of user-based performance of the system of constructed facilities.

The goal of the decision-maker is to provide the user with a system which will exhibit qualities of satisfactory performance throughout a certain design service life and in a relatively efficient manner. To this end, actions are chosen in planning, design, implementation, operation, and maintenance, based upon their predicted impact upon service life performance and cost. Costs must be broadly defined in terms of foregone opportunities for increased benefits to the various users, to include social and political factors as well as economic.

Performance is evaluated in terms of three measures of effectiveness; serviceability is a measure of the degree to which the facility provides satisfactory service to the user, where user is broadly defined to include the range of recipients of the benefits of the system. Reliability is a measure of the probability that serviceability will remain at adequate levels throughout a given design service life - A recognition of the physical uncertainties inherent in such systems. Maintainability is a measure of the degree to which continued effort is required during the facility's service life to keep performance at a satisfactory level.

Implementation of these measures requires application of ideas from diverse areas. Serviceability is evaluated through uses of psycho-physical and psychometric scaling techniques. Reliability and maintainability analysis requires stochastic simulation, often implemented with the help of

Digital computers. In particular, simulation using the Markov process appears promising.

The concept of performance and techniques for its use in design decision-making are illustrated through application to highway transportation. The area of urban housing is also briefly examined.

Thesis Supervisor: Professor Fred Moavenzadeh

Title: Associate Professor of Civil Engineering

TABLE OF CONTENTS

CHAPTER I

THE PROBLEM OF DESIGN DECISION

A. Introduction

Systems of constructed facilities -- highways, bridges, office buildings, houses, etc. -- are physical units which must be planned, designed, built, operated, and maintained, subject to complex and often far-reaching interactions with the social, political, and economic systems which they serve. The physical units, with their typically long service lifetimes and large size, represent a major comittment, not only in the recognized economic terms of capital, but also in terms of social and political possibilities.

Resources are required for constructed facilities: committments are made through allocations of resources to particular activities. The manner in which resources are allocated to a particular facility, and thus to gaining the services which this facility provides, includes several levels of detail.

On a national scale, decisions are made that particular sectors of activity are worthy of receiving resources. Comparisons are made on a broad policy basis among such concerns as education, transportation, and city government. Little thought is given at this stage to the individual projects eventually to be undertaken within each sector.

The outcome of possible choices (for allocation of resources) at this stage are evaluated by very general measures of welfare and development, viewed on the regional or national level. Per capita income, level of unemployment, literacy rate, and population are examples of the types of parameters which become the object of decision. The increasing need to have and use such criteria is reflected in the growing usage of performance budgeting in government (see Novick (1) for discussion), while the recognition that the ability to assess social factors lags sadly behind this ability in economic factors has led to work on social accounting and the possibilities for development of "social indicators" (2).

Once it has been decided that resources should be used in a sector of activity, decision-making proceeds to the more detailed problems of allocating these available resources to particular projects. If, for example, transportation is chosen as a desirable means for encouraging regional growth and development, then questions such as what links should be established, and what mode should be employed for each link must be approached.

Typically, decision-making at this second level will focus upon the equilibrium of economic supply and demand. Predictions of the market or demand for a facility's services are attempted, based upon the price which a user might pay and certain parameters describing the level of services provided. These parameters - factors such as trip time and frequency of service for a transportation facility are stated with varying degrees of explicitness, depending upon the sophistication of the prediction (see Manheim (3) for an example in the field of transportation). Similarly, attempts are made to determine what it would cost to supply these services. A comparison of the two predictions leads to a decision that a certain type of facility should be provided to supply the desired volume of services.

The exact details of the facility are left to be decided at the final stage of decision. Physical characteristics and operational policies must be selected. If, for example, it was decided at the higher level of decision that a highway would be the preferred way of providing transportation, the pavement and road alignment must now be determined. The criteria now are such factors as maximum vehicle load, numbers of vehicles, and anticipated weather conditions to be resisted.

The statement that a particular highway is preferred to other possible highways, and the decision that this particular system configuration should be implemented as the "optimal" use of resources, are then predicated upon decisions that a highway is the preferred mode of transport, and that transport is a desirable area of activity.

For systems of constructed facilities, one may conveniently refer to the higher two levels of decision as planning stages, and to the final level as the design stage. The terms planning and design, through their traditional usage, connote the more or less abstract terms in which the constructed facility is treated by former, while the latter deals with the "hardware" of system behavior.

The present work is concerned with this last stage of decision. The primary objective of this thesis is to present an approach to the analysis of systems of constructed facilities, analysis directed toward providing information about the design, implementation, operation, and maintenance activities which comprise the means by which a facility is realized. The results of this analysis should be useful at the design stage of allocation of resources, allocation made through selection of actions.

The question to be addressed in this thesis is not whether a particular transportation link should be a highway or a train, not whether a particular plot within the city should have low income or luxury housing, but rather, given this decision, what physical characteristics are required of the constructed facility in order that it may fulfill its role in the larger planning framework, and how may these characteristics be provided. That is to say, this analysis is undertaken with the assumption that what have been termed the planning decisions have initially been made.

This is not to say that the planning decision is a fixed and unchangeable constraint upon design. While the three levels of decision have been presented as distinct stages of activity, there is substantial overlap of planning and design. There should be significant exchange of information, both up and down. Indeed, it is incumbent upon decision-makers in the design stage to show when and why the system cannot meet the requirements implied in the planning decision within the limits of the assumptions which led to this decision. Similarly, it may be shown when there are design features of the facility which may permit the accomplishment of objectives in a more desirable way than that forseen in planning.

An example of this interaction is in the changing view of the use of maintenance activities for highways in developing nations. Once seen as something to be avoided because of the possibility that neglect would lead to failure, the maintenance-intensive alternative is now viewed as a desireable means of increasing employment and changing social patterns (4).

One of the intentions of the analysis presented in this thesis is then to draw additional attention to the need for this exchange of information between planning and design levels. The planner must recognize that his decisions are dependent upon the qualities of the system of constructed facilities, qualities which may be impossible to obtain in a manner commensurate with the previously stated planning goals. In turn, the designer must recognize that he is providing these qualities, and that a sacrifice of service for the sake of reducing resource requirements may not represent a satisfactory solution to the problem.

Even further, the exchange should continue during the service life of the facility. There should be a feedback of information on how the facility's behavior really is influencing the systems which the facility serves, and what changes in behavior might be possible and desireable.

This work will present an argument that constructed facilities can be analyzed in such a way as to permit and encourage this exchange of information, and more particularly it will explore and illustrate means whereby this analysis may be undertaken. The approach developed here is presented not as the only possible way of attacking design decision problems, but rather as a valid and useable means of undertaking what has not been and perhaps cannot be done with other, traditional, techniques.

The approach presented in this study is intended to be applicable to a broad range of types of constructed facilities. It must, however, be pointed out that the examples used throughout the development and exposition of the work are biased toward the particular area of transportation facilities, and especially highways. Thus it may only be suggested that what is done here for highways might be extended to other types of facilities: a beginning for urban housing is presented in an Appendix. It is hoped that further exploration will be encouraged.

B. A Concept of Performance

The system of constructed facilities is intended to supply a particular set of services. The nature and volume of the services desired are determined largely in the planning decisions which preceed actual design (subject, of course, to information from past design actions). The design decisions problem is one of allocating resources to undertake actions which will describe and bring into being a constructed facility which will deliver these desired services, and to make this allocation in a most desireable manner. Questions of what is "most desireable" will also depend upon planning decision and upon the broad role of the system of constructed facilities within the social, political, and economic systems which it serves.

Design decision will be concerned with how well a particular facility provides service - its performance - and with the resources required to obtain this service - the facility's cost. One will in general attempt at this stage to achieve the highest possible level of performance for any given level of resource usage, and must face major problems arising from the complex, multidimensional, character of both cost and performance.

Performance may be described in terms of three parameters - serviceability, reliability, and maintainability:

Serviceability is a measure of the degree to which the constructed facility provides satisfactory service to the user, from the user's point of view. The user, the recipient of the services of the constructed facility, is broadly defined to include not only the direct user, such as the driver of a highway vehicle, but also indirect users (merchants who receive goods shipped via highway) and subsidiary users (the eventual purchaser of the goods). Serviceability serves as an evaluation of the present service behavior of the constructed facility, and includes such factors as the quality of ride of a highway or the degree of comfort in the environment of a house (i.e., heat and humidity).

In many cases in evaluating serviceability, one is dealing with users' perceptions, and is thus dependent to a great extent upon subjective judgements. This dependence makes it necessary that continued attention be given to what the user - people - want and need from the constructed facility. Techniques for measuring serviceability are thus based heavily upon ideas from psychology and economics, particularly concepts of utility. In particular, it is suggested that a good measure of serviceability is the probability that any one of the community of users will find the present physical service behavior of the constructed facility to be satisfactory. In practice, this parameter will be estimated by the fraction of users who adjudge the facility adequate.

Reliability is a measure of the probability that a facility will provide adequate service - e.g., exhibit adequate serviceability - throughout the design service life of the constructed facility. The constructed facility is an uncertain physical system serving in an uncertain environment, and to make decisions based upon seemingly certain, deterministic, predictions of future service is unrealistic.

Predictions of future behavior should be made in a probabilistic fashion. Stochastic models, to be used in making such predictions, must be developed to describe the way in which the facility responds to its environment, and must be able to deal not only with the physical phenomena of weather and service usage, but also with the varying effects of operating and maintenance policies. These models may be developed through an analytical approach, where there is an understanding of the processes involved when failure occurs, or through an activities approach, when one knows only the events which lead up to an observed loss of serviceability. In either case, it may be expected that statistical data, gathered from observation of the environment and of facilities in service, will be of importance as input to the models.

Closely related to reliability is maintainability, proposed as a measure of the degree to which a facility will be sensitive to the uncertainties associated with future human activities. Specifically, maintainability may be defined as a measure of the degree to which continued effort is required throughout the service life to assure that serviceability remains at adequate levels. While maintenance activities, and the possible consequences of their neglect, represent the principal factor influencing maintainability. Other factors however, such as the possible political uncertainties associated with future funding, will influence maintainability and the designer's view of its importance.

Because of the primacy of maintenance in this component of performance, it is often convenient to think in terms of two distinct aspects of maintainability. One is maintainability with respect to normal maintenance, the scheduled, repetitive action, which is primarily preventive in character. To the extent that normal maintenance is effective, its neglect may be expected to lead to losses of reliability and subsequently of serviceability. The other aspect is maintainability with respect to repair, referring to the actions which are required if premature losses of serviceability are observed or are felt to be impending.

Together, reliability and maintainability serve to provide means for evaluation of the future availability of a facility. While serviceability refers to the present services provided by the facility, these other two components refer to the possibilities that these services will remain adequate throughout the remainder of the design service life.

At any instant of time, one may consider these three parameters - serviceability, reliability, and maintainability - to arrive at a judgement of the facility's value as a means of providing desired services throughout a specified period of time. Because of the way serviceability is defined, as the fraction of users who will find the physical characteristics of service to be satisfactory, this value will ultimately refer to the chances that the constructed facility will fullfill its role in the larger context of social, political, and economic systems. This idea will be amplified somewhat in the next chapter. The point to be made here is that at the design stage attempts are made to achieve a facility which will exhibit the highest possible value throughout the design life, for a given level of resources. That is, at each instant in the design service life, a facility with high value is one which exhibits adequate characteristics of present service, and the promise that these characteristics will so continue. Such a facility will be considered an acceptable solution to the design decision problem.

In particular, it may be suggested that the performance of a facility, how well that facility provides the services for which it is intended, will be evaluated by the predicted lifetime trends of value, in terms of serviceability, reliability, and maintainability. It will be proposed that performance is estimated by an integral, with respect to time, of value over the design life.

C. The Nature of Costs

As, suggested, one is concerned not only with performance, the way in which a facility provides service, but also with the resources required to realize the facility, and in particular with the costs derived therefrom. While this thesis is concerned primarily with the definition and evaluation of performance of systems of constructed facilities, a brief look at the nature of costs is in order. The treatment of costs has received considerable attention in the economic literature, and a more extensive discussion would be beyond the scope of present interests.

Basically, the cost of a resource in any particular use is determined by the most productive alternative use for that resource (5). For example, if the gravel to be used in a highway pavement could as well be sold for use in building construction, the cost of that gravel as used in the pavement must be at least as high as this market price. The value of a resource is thus measured in terms of an opportunity cost, signifying the foregone opportunities for alternative uses of that resource.

In view of previous discussion, it may be seen that any particular allocation of resources within a constructed facility may involve foregone opportunities which are not strictly economic in nature. Social and political impacts should also be considered. Costs, like performance, are complex and multidimensional.

The assessment of alternative uses for resources is, of course, often a rather difficult task. Possible uses, and thus costs, will in general be different in the short run from what they are in the long run. Over a longer period of time, social and political systems can change in response to physical stimulus: people will learn new ways of doing things, which could not be done in short periods of time. The planning decision must contend with the relatively long service life of constructed facilities; indeed, it would seem that many facilities depend quite strongly upon the "long run" for their justification.*

Another difficulty lies in the use of a common scale of measure for costs. It would be nice if all elements of resource evaluation could be translated into monetary equivalency, but this is seldom the case. Attempts to place values on such factors as travel time and loss of life (see Reference (7), for example) are generally open to serious practical as well as moral question. It seems doubtful that one measure of cost will accomodate all aspects of resource usage for constructed facilities.

In spite of such difficulties, costs must still be treated. Some suggestions as to how this task may be approached will prove useful.

It should immediately be stated that the resource requirements, and thus costs, for a constructed facility should be predicted for the entire design service life of the facility. The total cost of a facility includes not simply the initial implementation expenses, but also all future expenditures associated with the facility., To the extent that these costs are expressable in monetary terms, future expenditures may be made commensurable with current ones by expressing them in terms of their present value.

* The whole field of social overhead capital and economic development induced by this form of investment could be approached with this view. (See Reference 6)

That is, future costs are discounted by opportunity cost of capital (9), which is usually expressed as a percentage rate. (Under certain conditions this rate will be identical with the market rate of interest of loans). Most generally this discounting is done by referring all economic costs to a common time, usually the present, and summing to find the present value of total (economic) cost.

The problem of short versus long run is reflected in the basic dilemmaof whether costs should be judged by past experience or by what one feels is possible for the future. Use of past experience may be biased by conditions which do not or need not hold in the future, with the possible result that these conditions are reinforced (10). If little is expected, little is likely to be achieved. The second course runs the risk of being unrealistic, but can provide a stronger test of a facility's ability to fulfill the role for which it is planned. Further, the projections will in some degree serve to pace activity over the design service life. For example, allowing for higher maintenance expenses for a highway may encourage better quality maintenance activities in the future.

The problem of costs which are not readily expressible in economic terms is to some extent circumvented by comparing alternatives to a single base-level alternative. Increased costs are measured in terms of the sacrifices made in some aspects of service to achieve increases in other aspects, relative to the base level values. In this way it may be possible to avoid having to define a distinct scale of measure for the costs in question. For example, the social costs of several alternative highways, represented in community disruption and undesirable growth patterns, might be rated through of qualitative judgements as to whether each is better or worse than the mean or minimum economic cost alternative.

Of course, this idea of relative cost may be applied to economic factors as well. In some cases, especially at early stages of decision-making, much of the work involved in deciding upon the actual values of the proper costs of resources can be avoided with no loss of validity.

A subsidiary point to be made is the distinction in design between costs as they are perceived by the user and the actual total cost of a facility. The point is perhaps best made through an example.

Suppose that the problem is to design a highway pavement for a toll road, and that the resources available to be used for this pavement system are limited to 50% of the revenue received. The level of toll and an anticipated number of users who will travel the road at this toll are determined in planning through a projection of demand.

Further, suppose that comfort is the only component of the serviceability measure and that pavement roughness (as recorded with a standardized instrument) is the only indicant required for the prediction of comfort. As Figure 1 suggests one might expect serviceability to differ with varying levels of toll. At a higher toll, the user is less likely to tolerate roughness and discomfort.

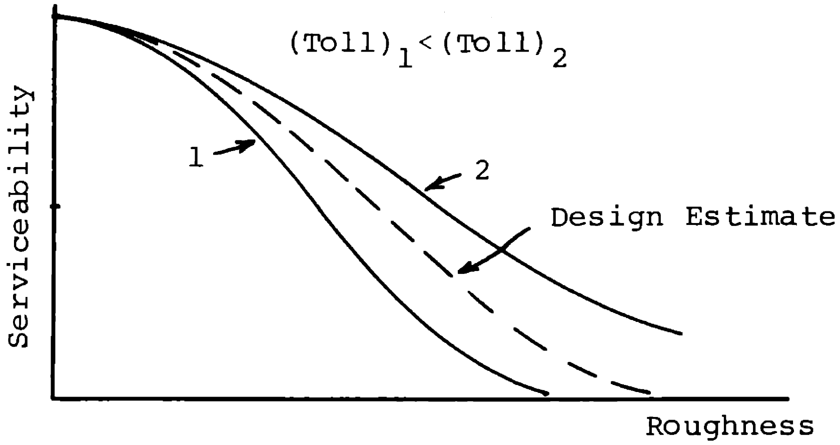


FIGURE 1: Change in Serviceability with Economic Factors

In practice, it will generally be the case that exact placement of these curves will not be known. Available information and experimentation will produce an estimation of serviceability versus roughness, assuming that other factors (i.e., toll) are roughly the same as the user's past experience. This estimate is shown as a broken line in Figure 1.

Recognizing this approximation, it will still be expected that at the given toll, there is some roughness above which the predictions of numbers of users, and thus of revenues, become quite questionable. The failure value S_f is determined in this way, as an attempt to have adequate physical characteristics.

In any case, the designer may find that the maximum serviceability he can achieve at the level of resources indicated is insufficient. It may be required that resources allowed the designer be raised to 60% of the toll receipt. Or, this additional funding could come from other sources, external to the system in question. The point is, that as long as the toll is not increased, the users' ideas of resources are unchanged, and the planning prediction holds. If the designer does not provide the required smoothness, because of constraints upon resource usage, the planning prediction is likely to become invalid. In short, the extent to which users receive and are concerned with resource allocations - costs - is reflected in the performance measure. Design decision, however, is based upon a broader view of all of the resources required for the facility's realization.

A discussion of costs could be expanded substantially by going into such things as the particular formulas to be used in computing present value of future costs, the ways in which opportunity costs of resources may be extracted from historical data and mathematical models, means for projecting longer run consequences or actions, more detailed explanation of the difference between consumers' and producers' views of costs, and so on. But this is beyond the scope of the present work. The principal points to be made about cost may be quickly summarized:

The costs associated with a system of constructed facilities must be recognized as occuring throughout the design service life of the facility, and their distribution over time can have impact upon decision. These costs are determined by the most productive alternative uses for the resources required in realization of a facility, and as such are often complex and difficult to evaluate. The total cost of a facility cannot always be evaluated in strictly monetary terms.

Capital has a time value; to the extent that resource expenditures are expressed in terms of monetary values, they should be referred to a common time through discounting. Where recognized expenditures are difficult to quantify or to evaluate in absolute terms, this approach of expressing costs in relative terms by comparing alternatives to a common base offers a possible solution.

A facility's total cost must be compared with its performance to give the decision-maker a basis for rational choice among alternatives. The following section will examine the activities of this comparison at the level of design decision.

D. How Decisions Are Made

The system of constructed facilities is evaluated in terms of its performance and cost, i.e., the qualities of its service and the resource requirements to provide this service over its design service life. It has been stated at several points in previous pages that an effort will be made to find a facility exhibiting the best performance possible at a given level of resources, and it has been explained that performance and cost are complex and multi-faceted. In what manner then can alternative facilities be compared, and how can a selection of the "most desireable" facility be made?

The design decision is made in a progressive manner, at several stages in the process of analysis. The first stage is a part of the search for an acceptable solution. Once a possible alternative has been proposed and checked to assure that its qualities of service are at least adequate (i.e., its predicted service behavior), an effort may be made to balance the resource allocation more effectively. There will generally be several individual subscales of serviceability, as it is estimated in practice. Unless there are prestated dominance relations among these subscales, such that higher levels are desired on some than on others, it will be most efficient to have equal estimates of serviceability on all subscales. That is, an allocation of resources which produces, for example, a high predicted fraction of satisfied users with respect to comfort on a highway, and a relatively low value with respect to safety, will be considered a somewhat inefficient use of these resources because the overall serviceability is likely to be limited by the lower subscale. A most efficient allocation would be made by sacrificing serviceability on the higher subscale to raise it on the lower one.

It will in many cases not be possible to balance the allocation of resources in this way, and a various alternatives will be characterized by their high predicted serviceability on certain subscales. The critical point however is that the balancing must be carried out, to the extent that it is necessary, to assure that at least minimum levels of serviceability are provided. This phase of the decision problem may be termed, in Simon's words (8), "satisficing".

The next stage of decision is reached when a number of alternatives have emerged from the designer's search, each having survived the satisficing strategy, and each having characteristics of cost and performance. Now one will search for those alternatives which exhibit the best performance at each level of cost. If the statements of cost and performance were straightforward and single valued, this search would present no difficulty and would be hardly worth mentioning. But this is seldom the case. The performance function will often be implicit in the actions of design decision, and not explicitly stated. In any case, examples of the types of criteria expressed in the actual statement of performance might include the following: minimum direct user cost at a given reliability, maintainability preferred to reliability at any particular expected serviceability, highest utilization of labor. By comparing alternatives with one another, within the context of such criteria, one will be able to identify best performance at a given cost, and will build up a sort of production function of increasing performance and cost.

This second screening results in a set of alternative possible solutions which are all acceptable and represent a relatively most efficient use of the particular resource requirement. The final selection of one facility from among this set of possibilities cannot be made within the context of the design decision problem, but must be made in view of the social, political, and economic systems with which the facility interacts, that is, at the planning level.

E. Structure of the Thesis

This first chapter has of course been intended to serve simply as an introduction to and review of the major points of the thesis. In Chapter II, the general framework of the analysis will be examined in more detail. In particular, the specific steps to be taken in the analysis will be reviewed, the concept of performance will be further formalized, and the view of the design decision problem presented here will be placed in the perspective of broader economic analysis (as representative of the planning decision).

The next section of the thesis, Chapters III and IV, will examine in detail the concepts of serviceability, reliability, and maintainability, and the use of these parameters as measures of service value. Formalized definitions and techniques for the application of the ideas developed will be presented.

Chapter V will present an example of the application of these ideas, to the area of highway pavements. This case study is intended not only to illustrate the approach to analysis but also to provide useful information for pavements. Another case, urban housing, is examined quite briefly in an Appendix.

Finally, Chapters VI and VII will close the presentation with a summary and evaluation, and with suggestions for areas in which the ideas presented here might be fruitfully extended.

CHAPTER II

A FRAMEWORK FOR ANALYSIS

This chapter will focus in particular upon a concept of performance, a concept which may be used at the level of design decision to compare alternative facility configurations as possible solutions to the design decision problem. A detailed description of performance will be presented, with an examination of what an evaluation of performance may reveal about design alternatives. Attention will be given to the relation of this approach to the broader context of economic analysis. Finally, a brief discussion of the decision process - the explicit steps to be taken in the analysis of systems of constructed facilities - will be given as a means of placing subsequent discussion in perspective.

A. A Concept of Performance

Performance has been defined as the manner in which a facility provides the services for which it was intended, and has been described as a function of serviceability, reliability, and maintainability. A central and distinctive point of the approach to analysis of constructed facilities presented here is the user-based description of serviceability, and thus of performance. The importance of serviceability, reliability, and maintainability at the design decision level lie in their estimation both of users' response to present conditions of service and of possible future response throughout the design service life. An important aspect of performance is that it includes consideration of the entire time period for which service is to be provided to users. The statement that a facility exhibits adequate performance will mean that its qualities of service behavior are now acceptable to users, and may be expected to remain so throughout the design service life.

A quick review of terms will serve to introduce some details which will be useful in gaining an understanding of performance (more complete discussion of these details is for the most part deferred to later chapters), and will provide a vehicle for establishing a bit of symbolic notation.* The statements so presented are intended to serve primarily as aids to intuitive understanding, although there is no case where operations indicated could not be handled in the more complex manner. Later chapters will re-examine this point.

Serviceability, defined as the degree to which a facility provides satisfactory service from the user's point of view, may be designated S(t), where t is time. This parameter is estimated by the fraction of users expected to adjudge the physical service characteristics of the constructed facility acceptable, and is thus measured on a scale from zero to unity. High serviceability, $S(t) \to 1.0$, indicates that at that instant of time there is a high liklihood, that a user will find the facility's service satisfactory.

In practice, serviceability is approximated by a multi-dimensional function estimating user response relative to a

* It should be understood that the following discussion is simplified by using single letters and symbols to denote what will often be multi-dimensional parameters.

number of measurable and apparently independent aspects of a facility's behavior. S(t) will thus in general have a number of component subscales, each predicting the fraction of users satisfied with that particular aspect of service.

Normal usage and aging of a facility will be expected to produce a deterioration of the qualities of service, reflected in decreasing S(t). From the higher level of planning decision, there will be designated a minimum acceptable level of serviceability, S_f. Failure is said to occur if serviceability fall below this level during the design service life, i.e., if $S(t)<S_f$. The basic requirement to be satisfied in design is that $S(t)\geq S_f$, $0\leq t\leq T_D$, where T_D is the design failure age, the end of the design service life.

Reliability, the probability that a facility will give adequate service, is thus defined relative to serviceability. Specifically, reliability may be written as $R(t) = \text{Prob}\ [S(\tau) \geq S_f,\ t < \tau \leq T_D]$. As a probability, a measure of the uncertainties associated with a particular facility, R(t) is measured on a scale of zero to unity, with $R(t)\rightarrow 1.0$ indicating a high probability that a facility will provide acceptable service throughout its design service life.

Maintainability, the extent to which continued effort is required throughout the design service life, is measured by the inverse of the fraction of the design life lost if failure occurs. Ease of maintenance will mean rapid repair and a low fraction of time lost. The coefficient of maintainability, M(t), will then increase on a scale from 1.0, indicating that repair cannot restore service before the end of the design life, to ∞, indicating instantaneous renewal and no time lost.

Together, reliability and maintainability can provide information about the future availability of a facility's services. Reliability approaching 1.0 or maintainability approaching ∞ will indicate a very low probability of failure or very little time lost in the event that failure does occur, in which case it may be considered that there is a good chance that the facility will provide adequate service throughout the design life. Conversely, low reliability and maintainability indicate high risk of failure and the likelihood that failure will mean significant losses of service time.

One may now define the value of a facility, V(t), as the estimate of how well that facility is meeting its goal of providing adequate service throughout the design service life, at that instant of time. At any instant of time, one alternative will be considered to be better than another, with respect to services provided, if its value V(t) = V[S(t), R(t), M(t)] is higher. This means that the facility exhibits good qualities of present service and good liklihood that adequate service will continue. High value is associated with high levels of serviceability, reliability, and maintainability.*

* At equal costs, the higher value is also clearly preferred. At this point, nothing is said about the costs of higher value and subsequent improved performance.

The performance of a constructed facility is then defined as a measure of the value of that facility over the entire design service life. In particular, performance Ψ is given as

$$\Psi(t) = \int_{t}^{T_D} V(\tau)\,d\tau$$

Good performance is associated with high value at every instant of the design service life.

The actual form of the value function, and thus of the measure of performance, will to some extent depend upon the nature of the particular design decision situation. The extent to which tradeoffs are allowed among serviceability, reliability, and maintainability, and the possibility that one of these parameters may be preferred to the others will influence how value is determined. These forms may now be explored.

B. Forms of the Value Function

Two opposite approaches may be taken to formulating the specific forms of the value and performance functions: It may be assumed that there is complete tradeoff among the three parameters comprising value (serviceability, reliability, and maintainability), allowing increases in one parameter to offset decreases in another. Or, it may be assumed that no tradeoff is allowed, in which case one parameter will generally be considered more important than the others in judging value.

It may be expected that there will be a full range of possible value functions between these two extremes.

1. Complete Tradeoff

The assumption that there is complete tradeoff is subject to the condition that the values of all three parameters comprising value are at or above any minimum of acceptability. Subject to this provision, there will be an equivalency between present service and future availability. One may begin by investigating availability.

The parameter $\frac{1-R(t)}{M(t)}$, which is the product of the probability of a failure (1-R) and the expected value of time lost if this failure occurs (1/M, a fraction of the service life), will estimate the expected value of the event that failure occurs and time is lost. The value $(1 - \frac{1-R}{M})$ is then the estimated fraction of the service life during which the services of the facility will be available. High reliability and ease of maintenance (high maintainability) will cause this fraction to tend to unity.

The value function would then have the form

$$V(t) = S(t) \; [1 - \frac{1-R(t)}{M(t)}]$$

This form indicates that value is equal to the present serviceability of the facility, modified by the future availability.

If serviceability is equal to the minimum acceptable, $S(t) = S_f$, then the condition for value to be considered adequate would be reliability equal to 1000 or maintainability going to infinity, or both (i.e., no chance of failure or no time lost if failure occurs). In either case, $V(t) = S(t) = S_f$. This minimum level for adequate value will then imply minimum acceptable levels of reliability and maintainability.

There will be a minimum acceptable reliability, R^*, given that $S = 1.0$ and $M = 1.0$. That is, this level of reliability is the limit to which tradeoff against increased serviceability will be allowed, given that a failure will lead to complete loss of service life, and assuming that the facility's value is to remain constant, $V(t) = S_f$. Substituting into the expression for value, it is found that numerically $R^* = S_f$. This equality may be interpreted as giving a measure of the amount of risk of failure which is tolerable, where risk arises from physical factors of system and environment, or from the possibility that users will find a given quality of service unacceptable.

Similarly, a minimum acceptable level of maintainability to keep value constant may be inferred. Again assuming that $S = 1.0$, reliability is allowed to drop to zero. At this point, substitution into the expression for value shows that $M = \frac{1}{1-R^*}$ is this minimum to keep value $V(t) = S_f$ (= R^*).

The contour of constant minimum value derived from these arguments is illustrated in Figure 1. If the value of a facility falls below this surface, it is considered to be unacceptable at that instant of time. Higher levels of value

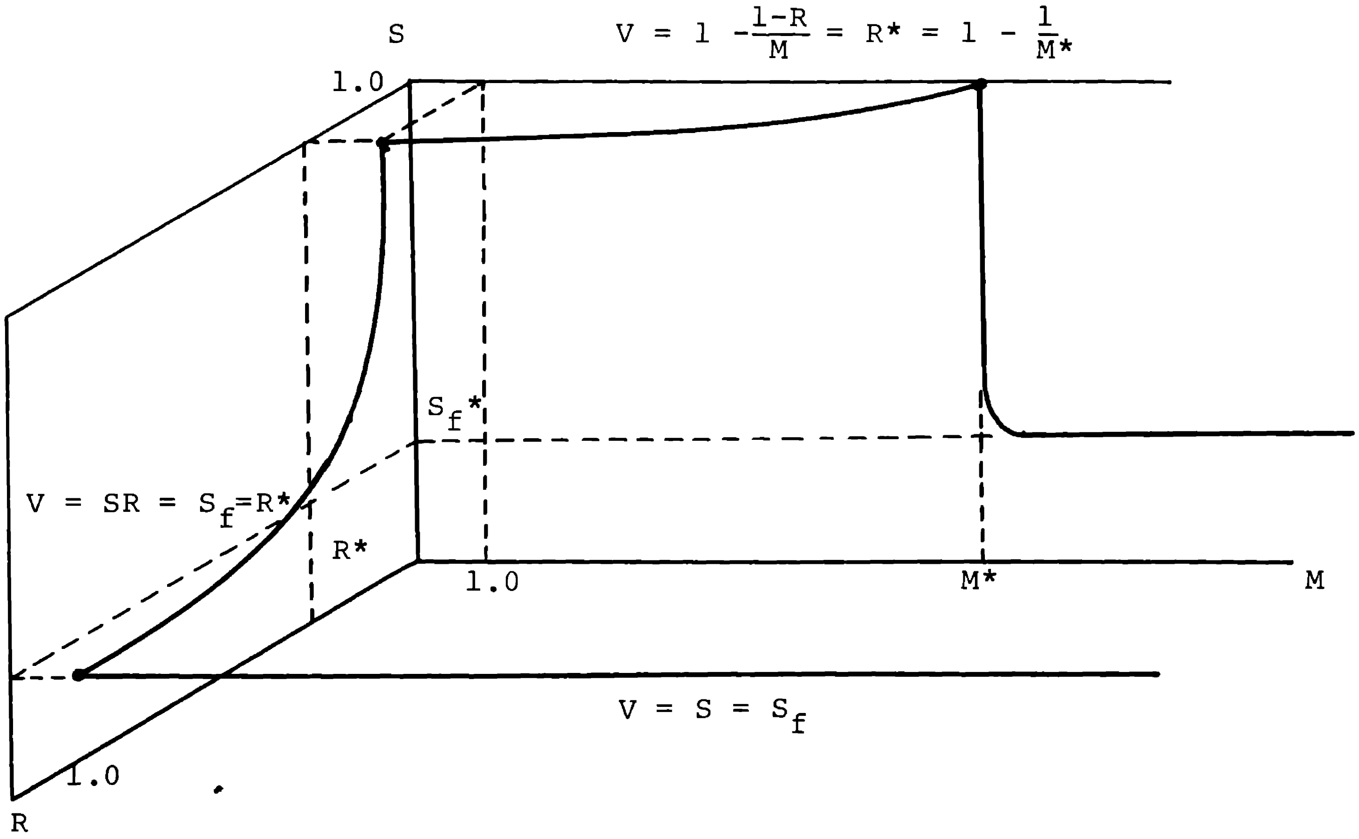


FIGURE 1: Contour of Minimum Acceptable Value, Assuming Complete Tradeoff,

$$V(t) = S(t)[1 - \frac{1-R(t)}{M(t)}]$$

will describe similar surfaces above this minimum. At any instant of time, a facility will be preferred to another, with respect to services provided, if it falls on a higher value contour above the minimum.

As suggested, this value function allows for complete tradeoff among serviceability, reliability, and maintainability within the concept of performance. A facility may have predicted qualities of service, on a day to day basis, which lead to a relatively low fraction of users likely to be satisfied, linked with very steady service and good estimated availability, and be considered to deliver performance equivalent to another facility having much higher serviceability but greater uncertainty. Further, maintainability and reliability may be freely exchanged within the framework of useful service time. At the other extreme of performance evaluation is the possibility that no tradeoff is allowed, that there are definite preference for one aspect of performance over another.

2. No Tradeoff Allowed, Dominance Among Components

If no tradeoff is allowed among components of performance, it will generally be the case that there is one or several components which are felt to be especially important, and which thus provide the basis for evaluation of the value function. For example, high reliability may be preferred to high maintainability and high serviceability (given that both are above any minimum of acceptability which might be set) in a situation where an inopportune failure would be especially undesireable. Such a case might occur with a transportation facility intended to serve heavy traffic at peak hours, but which is relatively idle at other times. Maintenance can be undertaken at leisure during the idle periods, and increased serviceability, while it would perhaps be nice, is definitely not as important as assuring that the facility is delivering at least adequate service at peak times.

In contrast, very high serviceability, linked with moderate levels of reliability, might be desired if the facility is new and must attract users. At a later stage in the service life, when patronage has been built up, reliability might become relatively more important.

The value function in such situations would assume a form such as the following:

$$V(t) = \begin{cases} 0 \text{ if } S(t) < S_f,\ R(t) < R^*,\ M(t) < M^* \\ R(t) \text{ if } S(t) \geq S_f,\ R(t) \geq R^*,\ M(t) \geq M^* \\ \text{and } R > M > S. \end{cases}$$

This means that if any of the three parameters comprising performance is below what is defined to be the minimum acceptable, then value is equal to zero; i.e., the facility absolutely is not fulfilling its role. If all parameters are above the minimum, then value is determined by the dominant, or preferred parameter. The last statement, R > M > S, is intended to give this dominance relation. If two alternatives are being compared, both above the failure levels, and are found to have equal reliability, then value is assessed on the basis of maintainability. Should M(t) also be equal for the two, serviceability is compared. If this parameter too is equal for the two, then the facility alternatives are said to have equal value at that instant. Very high maintainability in one alternative will not overcome preference for another alternative with slightly higher reliability.

It may be noted that in cases such as that above, where users are to be attracted to a new facility, the explicit form of the value function may change at some time during the design service life. This change, embodied in a shifting of the dominance relation, is a reflection of a change in the services required of the system of constructed facilities as part of a larger system.

C. The Measure of Performance and Modification for Time Value

The preceding section looked at specific forms the value function might take, in terms of the two extremes of allowable tradeoff among serviceability, reliability, and maintainability. It may be seen that, depending upon the particular design situation, the actual form of this function might lie between these extremes, incorporating points of each. For example, complete tradeoff might be allowed if reliability is above some relatively high level, while below this level (and above failure levels) a straightforward dominance relation would hold. This might be the case if there are constraints upon resource availability, but these fall above what is required to achieve a barely adequate facility, and there is value placed upon a particular aspect of performance, but only up to a point.

In all cases, however, the value function will give a dimensionless (i.e., having no natural unit of measurement) numerical rating for the facility's quality of service at each instant in the design service life. The performance function, as the integral of value with respect to time, will then assume the form of time weighted by value and summed over the design service life. As value can in general be defined so that it will vary from zero to 1.0, as illustrated above, the highest level of performance would be indicated by a numerical evaluation equal to the length of the design service life. This indicates that value is at its highest at all times during the service life.

There is a modification of the basic definition of performance which should be mentioned. As it has been presented, equal weight in the performance function is given to service provided at all times during the service life. There is reason to propose that the value of future predicted services should be discounted, as is normally done with the economic aspects of cost. In this case, the performance of a constructed facility would be given as

$$\psi(t) = \int_{t}^{T_o} \frac{V(\tau)}{(1+\delta)^{\tau-t}} d\tau$$

The factor δ is a discount factor, and reduces the apparent value of service as time increases to the failure age. The concept is entirely analogous to the discounting of future expenditures in the economic sense.

There are two principal justifications for this modification of the performance function. First, predictions of future behavior are in general less reliable as the time horizon of prediction increases. Discounting reduces the impact of more distant predictions.

Second, there is a general tendency of people to prefer present goods and services to future possibilities. This is observed to be true for a broad range of economic goods, and should apply to the services of constructed facilities as well. The possibilities for technological and social obsolesence of a constructed facility's services are a concrete example of why the preference for present services should hold.

D. The Links with Higher Levels of Decision

The previous discussion has been concerned primarily with describing the approach to decision at the design level, which will be developed herein. It has been stated, however that there is interaction among the levels of decision-making, that it is in fact impossible to isolate decision at any one level, ignoring other levels. In this section, an effort will be made to gain insight into the nature of this interaction by exploring the concept of performance within a context of economic analysis.

The constructed facility may be viewed as a production process, delivering services to users, the consumer. Decisions are then to be made about prices, quantity, and quality of services provided, based upon comparisons of supply and demand.* This view is represented in transportation planning by such work as that of Soberman (1), Lago (2), and Manheim, et al. (3).

Within this context, decision-making is generally directed toward allocating resources in production in an optimal fashion. Optimality is defined as a maximization of profits, the excess of revenues over costs. One might object to the application of this criterion to the general case of constructed facilities, which in many cases provide a public service. It is a basic theorem of economics, however, that consumers will be best off when the relative prices of goods are equal to their relative social costs (see Stigler (4)), and this theorem may be understood to imply sufficiently broad definitions of cost and revenue that there should always be a profit to an activity, albeit this profit may not appear in conventional economic terms. The selection of projects in international development for the highest social rate of return (5), just as in business one would invest in projects having the highest financial return, is

* Here again, the caveat must be given that prices and services are quite complex, and the notation is useful primarily as intuitive argument.

a reflection of this idea.

In the simplest and most classical case, costs of production are assumed to vary only with quantity produced, as all other factors, notably quality of product, are assumed to be fixed and constant. Similarly, demand is determined as a function of price with this same assumption of constant quality. The profit maximization problem is then one of setting price and quantity, and is solved at the point when the marginal costs of production are equal to the marginal revenue of sales. That is, output is increased until the cost of producing the last unit is exactly equal to the increased revenue associated with the sale of this last unit. This equilibrium solution is illustrated graphically in Figure 2, a familiar picture in any economics text.

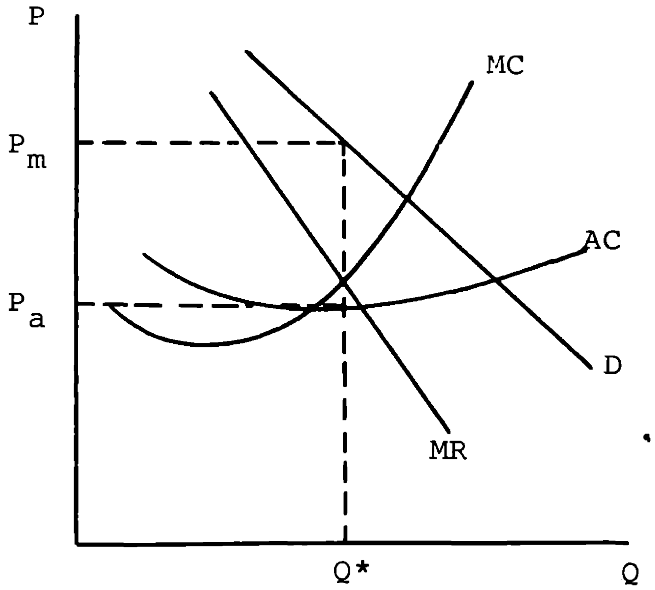


FIGURE 2: Profit Maximization with Price and Quantity Only

The marginal cost (MC) and average cost (AC) curves are the statements of a producer's production function, telling the costs of producing any quantity of goods of an assumed quality and input mix. Because there are so many ways that resource inputs could be combined to produce any given output, it is often implicit in arguments of this sort that the technologically most efficient means of production is being used (4). To be compared with this production statement is that of demand (D), the quantity that will be consumed at any particular price. The marginal revenue curve (MR) is derived from the demand curve. Profit is maximized at the point at which MR = MC. At this point, the selling price is P_m, bringing revenues of P_mQ*, as opposed to total costs of P_aQ*.

The difficulty with this conclusion, or rather with the argument leading to it, is that there are a number of other variables available for decision. In particular, the quality of the goods can be varied, effecting both cost and demand. The problem is complicated by the fact that the actual form of this variation is not in practice known, but must be found through the activities of design. Each new facility represents a new production situation, so that planning decisions made on the basis of past experience with constructed facilities cannot avoid the problem of incomplete information.

One may begin to explore further the role of the physical characteristics of the facility. The so-called Dorfman-Steiner theorem provides a useful basis for this inquiry, by incorporating product quality into the optimality conditions. The theorem was derived to show the proper allocation of resources among quantity, quality, and advertising in the production and marketing of goods or services.*

One starts with the following definitions:

Q = unit sales during the period of analysis
P = price paid per unit, by the computer (user)
c = average cost of production (AC in Figure 2)
s = advertising outlay during the period
y = an index of product quality.

$$Q = Q(P,s,y)$$
$$c = (c(Q,y)$$
$$\pi = PQ(P,s,y) - Qc(Q,y) - s$$

Profit (π) and all other functions are assumed continuous and differentiable.

The above definitions may be somewhat elaborated to place them within the context of constructed facilities. Sales, or the quantity of goods demanded, may be seen as the number of units of service usage delivered by a facility. For example, total trip miles might be an appropriate unit of measure for a transportation facility. Price and average cost would be viewed in these terms, although in the case of price there is a problem of the difference

* The derivation which follows is based upon Palda's discussion (6).

between actual and perceived values. Plourde (7), for example has looked at this question in some depth.

The concept of advertising is a bit unusual in the present context. By advertising is meant any activity not directly a part of the constructed facility, which has the effect of influencing the user's perception and judgement of service and thus of shifting demand. Besides the generally understood forms of advertising, quite frequently used, for example, in housing, there will be a range of other activities which will fall into this part of the analytical model. There are educational efforts directed toward increasing the user's understanding of what he is getting, as with a new form of facility, and what he is paying, as with the concern for pollution as an encouragement to use mass transit facilities.

Subsidies too might fit into this part of the model. Especially in the case of a direct subsidy to reduce the price charged to the user, a subsidy will effectively increase the demand for services. Further, the subsidy may be viewed as reducing the broadly defined "profits" of the facility by serving to divert resources from other possible uses. The consideration of advertising expenditures is thus left to the present discussion for the interesting suggestions it may provide about actions which may be taken outside of the physical system to effect changes within this system.

The final variable, the index of product quality, may be seen to refer to precisely those qualities of physical service with which design decision is concerned. It has been argued in previous discussion that there are a number of factors which the user perceives and judges in determining the adequacy of a facility's service, and that these factors are estimated by measurable indicants of service. One example used was the role of roughness, as measured with a standard instrument, as an indicant of serviceability with respect to comfort for a highway pavement. These judgement factors and their indicants are then an effective index of the quality of service, and thus are among the principal components in the evaluation of performance. For further illustration, one may compare the above example with other indices typically suggested in the economic literature to illustrate this parameter. Such suggestions include the number of cylinders in a car and the load capacity of a washing machine.

In making the derivation, the first step is to maximize the profit function. This is done by taking partial derivatives of the expression for profit, with respect to price, advertising, and quality, and setting them equal to zero.

$$\frac{\partial \pi}{\partial P} = Q + P \frac{\partial Q}{\partial P} - C \frac{\partial Q}{\partial P} - Q \frac{\partial c}{\partial Q} \frac{\partial Q}{\partial P} = 0$$

$$\frac{\partial \pi}{\partial s} = P \frac{\partial Q}{\partial s} - c \frac{\partial Q}{\partial s} - Q \frac{\partial c}{\partial Q} \frac{\partial Q}{\partial s} - 1 = 0$$

$$\frac{\partial\pi}{\partial y} = P\,\frac{\partial Q}{\partial y} - c\,\frac{\partial Q}{\partial y} - Q\left(\frac{\partial c}{\partial Q}\,\frac{\partial Q}{\partial y} + \frac{\partial c}{\partial y}\right) = 0$$

Simplifying and expressing the equations in like terms of price and cost the following equilibrium condition may be written:

$$\frac{-Q}{\partial Q/\partial P} = Q\,\frac{\partial c/\partial y}{\partial Q/\partial y} = \frac{1}{\partial Q/\partial s} \tag{2}$$

This condition may be made more meaningful, in economic terms, by introducing the concepts of elasticity.

Price elasticity of demand is defined as

$$\eta = -\,\frac{\partial Q}{\partial P}\,\frac{P}{Q} \tag{3}$$

This parameter gives the fraction decrease in demand resulting from a fractional increase in price; i.e., with transposition, one finds

$$\frac{\partial Q}{\partial P} = -\,\frac{Q\eta}{P} \tag{4}$$

The elasticity of demand with respect to changes in quality is given as

$$\eta_c = \frac{\partial Q/\ y}{\partial c/\ y}\,\frac{c}{Q} \tag{5}$$

This is the fraction change in demand relative to the fraction change in production cost, both induced by a change in quality.

The marginal effect of advertising on sales is given by the parameter

$$\mu = \frac{\partial Q}{\partial s} P \qquad (6)$$

This is the incremental increase in revenue due to a small increase in advertising expenditure.

With appropriate transporsition, the three parameters defined in (3), (5), and (6) above may be placed in the conditions of optimality (2). One then finds that

$$\frac{P}{\eta} = \frac{c}{\eta_c} = \frac{P}{\mu} \qquad (7)$$

are the conditions for profit maximization. Inverting and multiplying by price, one has the Dorman-Steiner theorem;

$$\eta = \frac{P}{c} \eta_c = \mu \qquad (8)$$

This theorem states that the producer of a product will maximize his profit if he can manipulate his allocations of resources to the point where the numerical value of the price elasticity of demand, the value of the marginal effect of advertising expenditure on sales, and the value of the product of quality elasticity and sales markup (over average cost) are all equal. From the definition of marginal revenue as

$$MR = P(1 - \frac{1}{\eta}) \qquad (9)$$

it is found that these conditions are linked to the basic MC = MR rule as follows:

$$\eta = \frac{P}{P - MR} = \frac{P}{P - MC} \qquad (10)$$

Refering back to statement (5), it is seen that the second term of the Dorman-Steiner theorem may be rewritten as

$$\frac{P}{c}\,\eta_c = \frac{\partial Q/\partial y}{\partial c/\partial y}\,\frac{P}{Q} \qquad (11)$$

It was stated that the quality parameter y is viewed as the same set of factors upon which the estimation of serviceability is based, as will be explained in more detail in the next chapter. As has been explained, serviceability is estimated as the fraction of users finding the qualities of service to be adequate, and who are therefore presumably willing to serve the role of consumer. Then ,the variation of serviceability with qualities of service is written

$$\frac{\partial s}{\partial y} = \frac{\partial Q}{\partial y} \frac{1}{Q} \tag{12}$$

where s is the serviceability function.

Replacing the quantity terms in the right hand side of statement (11), one finds that

$$\frac{P}{c} \eta_c = \frac{\partial s/\partial y}{\partial c/\partial y} P = \eta \tag{13}$$

This is replaced in the theorem statement (8) to arrive at the following criterion for optimal physical conditions:

$$\frac{\partial s}{\partial y} = \frac{\partial c}{\partial y} \frac{\eta}{P} \tag{14}$$

That is, optimality is achieved when the slope of the serviceability function is numerically equal to the product of price elasticity of demand and the rate of change of average cost with respect to changes in quality, as a fraction of price.

This equation then states the desired conditions of quality of service relative to the price-quantity decision traditionally considered in economic planning. In principle, if there were complete knowledge of the factors determining demand for the facility's services, all aspects of the constructed facility would be decided at once, as implied in the above derivation. In practice, however, there is the previously discussed separation of decision-making into planning and design levels, with the latter concerned primarily with service quality. The serviceability function is estimated at the design level assuming that factors such as price are in normally expected ranges, just as demand is estimated at the planning levels assuming past experience regarding quality. This condition relating serviceability, and thus the entire design concept of performance, to demand (reflected in elasticity) and price is then the desired link between these levels of planning and design.

This link may be examined in more detail. Figure 3 illustrates a typical serviceability function, or rather, a typical serviceability subscale. (The nature of the serviceability function is explored in more detail in the next chapter). In general, there will be an S-shaped curve of serviceability with respect to the judgemental qualities designated by "y". This function is estimated at the level of design decision.

As suggested in Figure 2, the average cost of a facility is initially estimated for the purpose of making the price-quantity decision, Assuming past experience regarding quality of service, it should be possible to estimate the variation of costs with small changes of service quality, without explicit judgement of the location of the cost curve relative to service quality (on an absolute scale). Then with the planning-derived estimate of price elasticity of demand and the decision regarding price, and optimal value for the slope of the serviceability function may be computed. As suggested in Figure 3, there may in general be two points at which the slope is equal to this value. If the quality of service is below the lower of these values, then resources devoted to increasing quality will yield increases in demand, leading to increased revenue in excess of increased cost. Above the upper limit, increased serviceability (implying increased demand) is insufficient to justify increased resource expenditure.

In defining the performance of a constructed facility, the failure level of serviceability will then be set in the interval between these two points. The performance of possible alternatives will be predicted (in design) relative to this failure level. As suggested in Figure 4, the final set of alternatives which emerge from design decision-making will fall somewhere above or below the level of the original planning estimate. A preponderance of alternatives well above the planning estimate will suggest that the planning decision should be reviewed because costs were under-estimated - perhaps the facility in question does not represent the best use of resources. Alternatives below the planning estimate indicate that there are unexpected savings to be had in this project.

The spread of design alternatives suggested in Figure 4 arises through the design concept of performance. Serviceability reflects the users' views of service quality, and

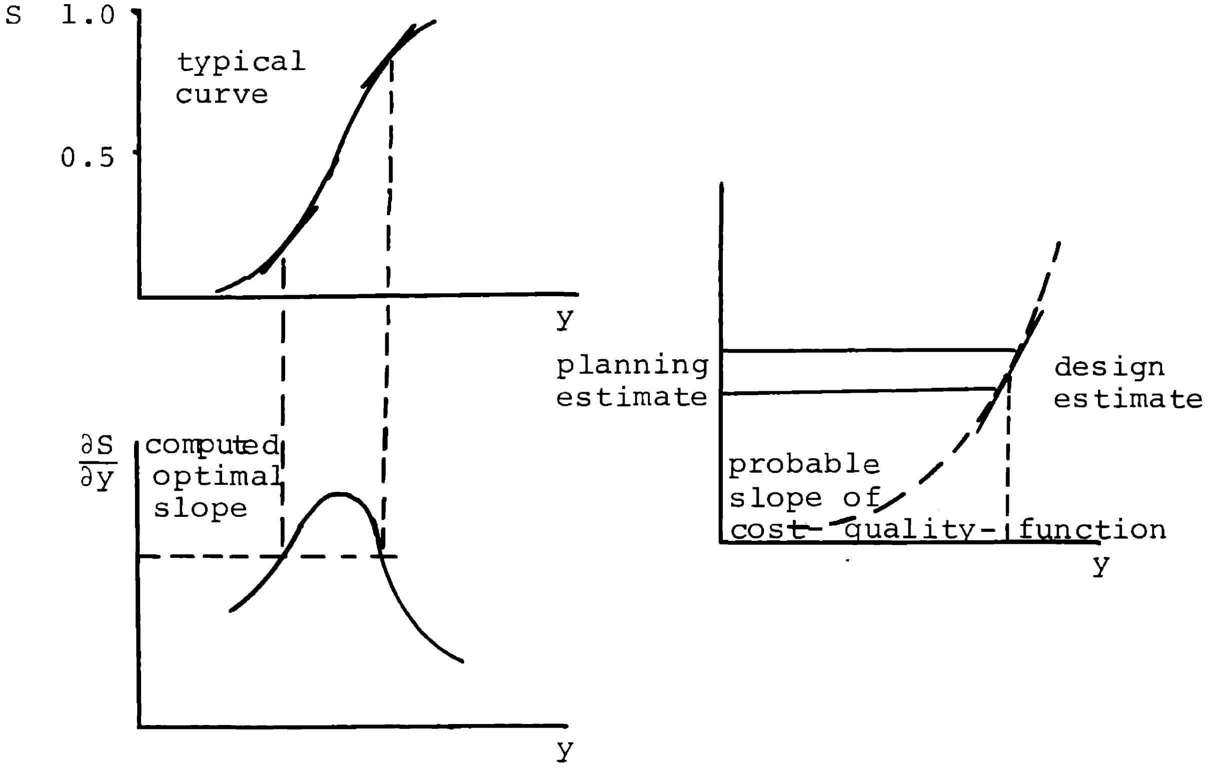


FIGURE 3: The Link Between Design and Planning

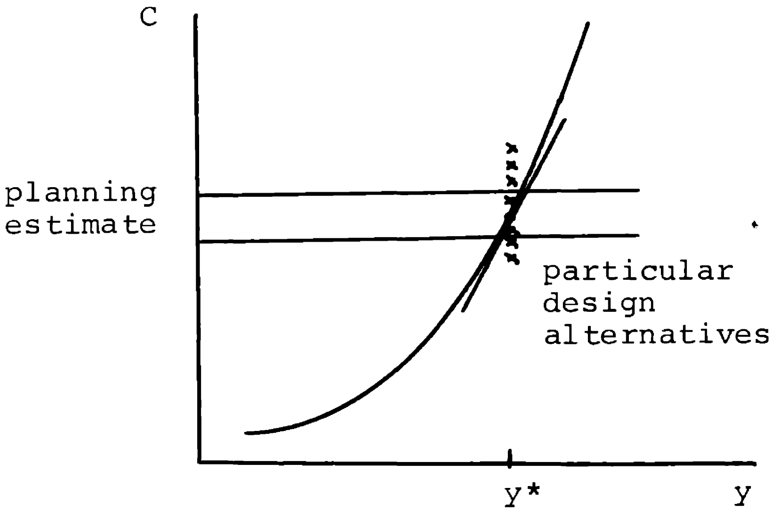


FIGURE 4: Design Estimates of Cost

thus provides a link to planning decision, as explained. But there is no consideration on the part of the user of the full impact of system reliability and maintainability. For a given minimum serviceability, as implied by the service quality y*, increased resource usage can be directed toward increasing reliability or maintainability, and thus performance. This set of cost estimates is then the set of efficient alternative facilities found in design analysis, and is a reflection of the multi-faceted aspect of performance.

Following the line of reasoning suggested above, an iterative application of planning and design activity may achieve a balanced allocation of resources. The planning price-quantity decision leads to design quality conclusions, which in turn tell something about planning assumptions.

Now, having looked at the tools for design decision, and their possible operational links with the planning levels, a brief examination of the process of design decision will be made.

E. The Process of Design Decision

A discussion of the decision process serves two purposes. First, it provides a skeleton for the framework for analysis being developed here. Outlining the specific steps to be taken will illustrate at what point each of the concepts presented is used, and how they fit together to give information of use in decision. Second, there is considerable opportunity in this work for use of the computer as an aid in analysis. The description of the design decision process may serve as a basis for devising particular computer programs to undertake particular aspects of the analysis. Although the computer has been used in a very limited way in this work, there is much room for contribution in this area. (See Alexander (8) or Guenther (9) for examples in transportation; others abound).

It must be pointed out at the outset of this discussion that although the process of decision is here described in terms of distinct steps, in practice these steps are seldom so clear and may follow in different sequence from that shown. The model suggested here is by no means the only way to approach the problem.

Figure 5 presents a picture of the process of design decision, with steps numbered in order of possible occurrence. These steps will be examined individually.

The first step is identification of the specific component subscales of the serviceability measure. As explained previously, serviceability is multidimensional in character, reflecting the various facets of service required of the facility. Identification will typically proceed through some combination of judgement and experimental technique, as will be discussed in Chapter III.

Once subscales of the serviceability function, suitable indicants must be found to permit prediction of serviceability on these subscales (Step 2). These indicants are

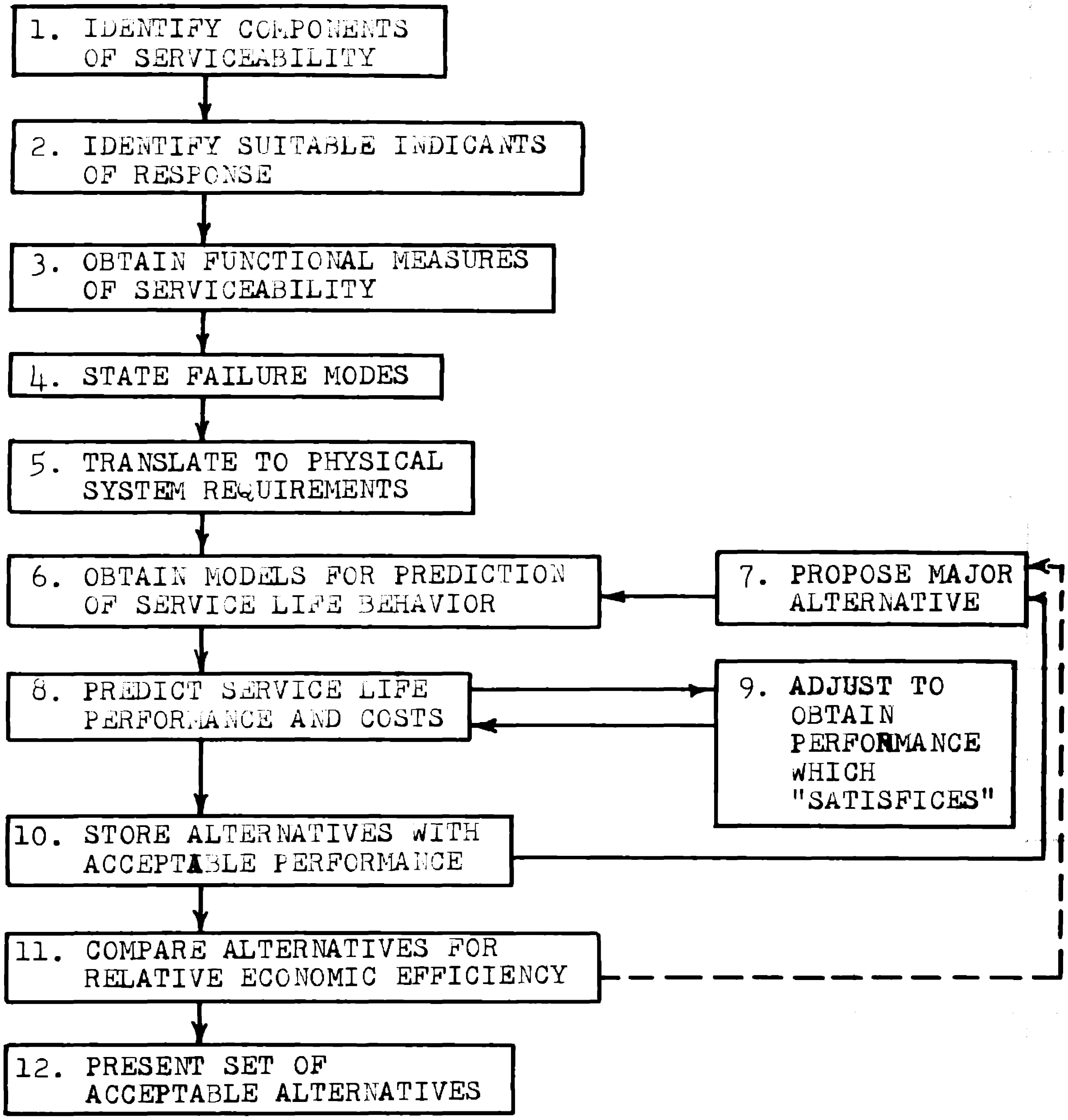

1. IDENTIFY COMPONENTS OF SERVICEABILITY
2. IDENTIFY SUITABLE INDICANTS OF RESPONSE
3. OBTAIN FUNCTIONAL MEASURES OF SERVICEABILITY
4. STATE FAILURE MODES
5. TRANSLATE TO PHYSICAL SYSTEM REQUIREMENTS
6. OBTAIN MODELS FOR PREDICTION OF SERVICE LIFE BEHAVIOR
7. PROPOSE MAJOR ALTERNATIVE
8. PREDICT SERVICE LIFE PERFORMANCE AND COSTS
9. ADJUST TO OBTAIN PERFORMANCE WHICH "SATISFICES"
10. STORE ALTERNATIVES WITH ACCEPTABLE PERFORMANCE
11. COMPARE ALTERNATIVES FOR RELATIVE ECONOMIC EFFICIENCY
12. PRESENT SET OF ACCEPTABLE ALTERNATIVES

Figure 5 : Steps in the analysis of systems of constructed facilities

parameters which are based upon measureable features of physical behavior. Referring to a previous example, one component subscale of serviceability of highway pavements is found to be quality of ride perceived by direct users. In turn, it is found that quality of ride is predicted by measurement of the macroscopic roughness of the pavement and the use classification of the road (i.e., high speed expressway, secondary roads, etc.).

Now one may proceed to develop the actual function of serviceability with respect to the various indicants (Step 3). This step is most difficult when it involves assessment of direct users' response, dependent upon subjective perceptions of these users. A range of scaling methods have been devised in psychology and economics, which offer assistance in this task.

Minimum acceptable levels of serviceability are specified (Step 4) with reference to the planning levels of decision. These minima provide the basis for computations of reliability, beginning with statement of the possible modes of failure for the facility (Step 5). Failure may occur on any one of the several component subscales of serviceability, generally in any of a number of physical manners.

The next step (Step 6) is to find or devise models which will permit prediction of the physical service behavior of particular alternative facilities. These modesl will be stochastic in nature, making probabilistic predictions of behavior as a function of actions taken in implementation, operation, and maintenance of the system. An alternative is proposed in terms of such actions (Step 7), and predictions are made (Step 8). Resource requirements for the system are also predicted at this stage.

It will often be the case that there will be some inadequate aspects of serviceability or apparent inefficiencies or resource usage revealed in this initial prediction. A sub-optimization procedure may be undertaken (Step 9) to adjust the alternative to deliver at least satisfactory performance in a balanced manner. For example, an otherwise satisfactory pavement system might be quite likely to show loss of safety due to polishing of aggregate at the surface. A slight adjustment of resource allocation, via use of a different aggregate or scheduling of special maintenance activities, will raise this aspect of serviceability, and thus the alternative, into the satisfactory range. An alternative which has passed this step is then set aside (Step 10) as the search continues, in a cyclical fashion, to develop a range of acceptable alternatives.

When a number of alternatives have been so prepared, they may be compared for relative efficiency of resource usage. The alternative exhibiting the highest levels of performance at any given level of resource usage are judged to be most efficient and define the so-called production function of performance versus cost. Of course, it is possible that further search would produce more productive alternatives, and this possibility must be considered in deciding when to terminate the search procedure.

Through the comparison of alternatives for relative efficiency, the output of the design decision process is generated (Step 12). The set of alternatives defining the efficiency envelope are expected to deliver satisfactory service in an efficient manner, through the design service life of the facility. Selection of one from among this set, for actual implementation, must be made with reference to a higher level of decision.

CHAPTER III

SERVICEABILITY AND ITS MEASUREMENT

A. Introduction

Serviceability has been defined as the degree to which a system of constructed facilities provides adequate service to the user, from the user's point of view. Within the broader context of performance, this parameter is presented as a means for evaluating the present qualities of the system's physical behavior.

Specifically, it was suggested that serviceability is estimated as the probability that the user will judge service to be satisfactory. In practice, this parameter will be measured as the fraction of users finding service to be adequate. The level of serviceability required to render the facility satisfactory with respect to the design level of decision was presented as derived from the planning levels: serviceability must be sufficient to give assurance of the feasibility of the planning decision.

In this chapter, the serviceability function will be discussed in detail. Section B will present a detailed treatment of the definition and development of the function. Section C is then a combination of literature review presenting background and justification for the proposed measure of serviceability, and a synthesis and extension of ideas to be applied to the evaluation of the behavior of systems of constructed facilities. In Section D, the serviceability function will be explored to see what may be said in general about its behavior and possible tradeoffs among the dimensions of users' judgement. Finally, Section E will draw upon previous sections to make explicit statements about how the serviceability of systems of constructed facilities might be evaluated for a particular problem.

B. The Service Problem and A Measure of Success

1. The Approach

An individual user will perceive certain qualities of the physical service characteristics of a system of constructed facilities, and will judge the facility on these perceptions. There will be one or more internal judgemental factors which the user considers. For example, the roughness of a highway pavement will be perceived by a direct user as vibration and noise in the vehicle. He will then derive a feeling of comfort or discomfort arising from this vibration and noise, and it is this feeling which he would use, among others, to judge the adequacy of the pavement.

It may be suggested that the user has a judgement space $[Z_i]$ of i independent factors against which he judges a constructed facility. The user derives some value or pleasure, or utility, from increased amounts of these factors $[Z_i]$. This utility is derived according to a function $U_m(Z)$, where m indicates that this is a particular individual, denoted m, and Z is a particular vector within $[Z_i]$. A constructed facility will be judged by the user in terms of such a vector Z.

The function $U_m(Z)$ is posited to be a monotonicly increasing function of Z, and there exists a value $U_m(Z_m*)$ above which the individual will feel that the facility is generally satisfactory. The value Z_m* which is judged to be satisfactory is termed the individual's aspiration level. Z_m* will define a surface in $[Z_i]$ such that $Z \geq Z_m*$ will indicate that the particular example of service, judged as Z, is satisfactory to the individual user m.

The user will make these judgements based upon his perception of a set of system service qualities $[Y_j]$. In the example mentioned above, noise and vibrations are such qualities. It is suggested that the facility will exhibit a vector y of such service qualities, and that there is a relation.

$$Z = B_m(y) \tag{1}$$

$B_m(y)$ is termed the individual's perception function. In the current example, comfort is a function of perceived noise and vibration.

Finally, there are certain measurable system characteristics $[x_k]$ which may be used to predict the value of y for a particular service situation. Macroscopic pavement roughness, vehicle speed and suspension system are examples of $[x_k]$. These characteristics may be termed indicants of perceived system service qualities. One may propose a function

$$y = A[x] \qquad (2)$$

to make predictions of these qualities.

Then if one wishes to provide a facility which a particular individual user will judge to be satisfactory, one must assure that the facility will exhibit characteristics x such that the following is true:

$$Z \geq Z_m \qquad (3)$$

$$Z = B_m(y) \qquad (1)$$

$$y = A(x) \qquad (2)$$

That is, in the example, the roughness of the road and the vehicle characteristics must be such that the direct user will feel adequately comfortable. However, he is not concerned directly with vehicle and road, but rather with qualities of noise and vibration which these characteristics induce.

A decision-maker attempting to satisfy the user by meeting condition (3) will have to allocate resources for the constructed facility. If one defines a function $\pi(x)$ as the cost of achieving system characteristics x, then the decision-maker will want to have a system of constructed facilities which solves the following program:

$$\text{Minimize } \pi(x) \tag{4}$$

such that;

$$Z \geq Z_m{}^*$$

$$Z = B_m(y)$$

$$y = A(x)$$

This program is a restatement of the basic goal for systems of constructed facilities, for an individual user, and neglecting explicit considerations of the effects of time.

2. Serviceability as A Measure of Success: The Problem of Many Users

The system of constructed facilities must serve many users. Each one of these users will perceive and judge the service of the facility in the manner described above. The decision maker's problem is made more complex by the introduction of additional constraints, representing the judgement of each of these users. That is, the problem is now to

$$\text{Minimize } \pi(x) \tag{4a}$$

subject to

$$Z_1 \geq Z_1{}^*$$

$$Z_2 \geq Z_2$$

$$\vdots$$

$$Z_M \geq Z_M{}^*$$

$$Z_1 = B_1(y)$$
$$Z_2 = B_2(y)$$
$$\vdots$$
$$Z_M = B_M(y)$$

for a total of M individual users.

As the number of users increases, the problem becomes rapidly more complex. Not only does each new user have a new perception function and aspirations level, but also there is the possibility that new qualities $[y_i]$ will be needed to predict Z_m. Note that the subscript m has been applied to indicate that each individual perceives and judges the facility in his own way. It is suggested here that one cannot in practice solve this problem with certainty.*

It is proposed that the decision-maker must evaluate his success in terms of a probabilistic measure. Assume that a function can be found to generate a vector

$$Z = B(y)$$

Z is a vector in a space $[Z_i]$ which includes all of the judgemental components of all of the M users, and the function B is an abstracted tool to be used by the analyst. Then one may define a new function S(Z), which will be

* Indeed, the problem may in fact not be solveable under conditions of free choice (see Arrow's General Possibility Theorem (1)).

termed the serviceability of a system of constructed facilities, thus:

$$S(Z) = \text{Prob}[Z \geq Z_m^*,\ m = 1,2,\ldots,M] \qquad (5)$$

This function is then a measure of the probability that the constraints of the decision problem (4a) are met.

Serviceability is thus suggested as a measure of the degree to which a facility provides satisfactory service to the user, from the user's point of view. This measure may be used in analysis as an indication of how close a particular alternative is to solving the basic problem of (4a). This problem might be recast as

$$\begin{aligned} &\text{Maximize } S(Z) \qquad (6) \\ &\text{subject to } \pi(x) < \pi_o \end{aligned}$$

where π_o is a budget constraint. This type of statement will be convenient in later discussions.

In practice, as will be shown, the function S(Z) will be approximated by another function S'(y), which will estimate the fraction of users who will be satisfied with a given level of physical service qualities. This function will be found in the form of several separate subscales estimating fraction satisfied with particular aspects of service, as a practical approach to the overall measure.

The above is the rationale for serviceability as a measure of effectiveness for systems of constructed facilities. Use of the measure of course involves a number of problems. Analysis requires identification of the components of $[Z_i]$ and the functions A(x) and B(y) which will permit an evaluation of Z for a facility. One must also be able to find the distribution of individuals' aspiration levels Z_m*, implying that one might wish to know something about the individual functions $B_m(y)$. The following pages will attempt to present arguments for the possibility of overcoming these problems and for the validity of this approach applied to a broad range of physical behavioral characteristics.

C. Bases of the Serviceability Function

1. Sources in Psychophysics and Psychological Scaling

An idea basic to the above development is that of utility. It is thus appropriate to review this idea and its range of applicability. This review may best proceed within an historical framework.

While the concept of utility as a measure of subjective response was initiated in the field of economics, the developments in that field and applications of the concept to explain consumer's behavior may be viewed as a part of more extensive work in psychology. With the broader acceptance in the early nineteenth century of the idea that response might not only be discussed as a means of explaining otherwise anomalous behavior, but also might actually be measured, psychophysics and the statistical approach to individual differences was underway. From these two areas of psychology have grown the principal ideas of what subjective response is, how it might be measured, and how it will enter into the individual's behavior.

Psychophysics was described by Fechner (who may be thought of as its founder) as "an exact theory of functionally dependent relations of body and soul" (2). The primary interest of this field is to measure sensitivity and discriminatory capacity of the senses-physiological response. Mental testing, or psychometrics (3), is concerned statistically with the variations among individuals, and in particular with measurements of opinion and intelligence. Following independent paths, these two divisions of psychology have developed substantial knowledge in the field of psychological scaling. It is from this knowledge that basis and techniques for serviceability may be drawn.

The concepts are not completely new. In 1760, Bouguer performed an experiment in visual perception (4). He took two lighted candles and a vertical rod, and moved one of the candles away from the rod until the shadow of the rod was just barely noticeable on the background screen upon which both candles shone. This procedure was repeated with another arrangement of the candles, to determine another barely noticeable difference. He found that the difference in illumination intensity between background and shadow was about 1/64 of the intensity of the backgrounds at all levels of illumination. This difference in intensity is an illustration of a threshold of perception, the smallest difference which can be detected. In the terminology more fully developed later, by Fechner (2), this is the just noticeable difference, designated jnd.

Bouguer's experiments and others like it anticipated the statement in 1834 by Weber of the law which came to bear his name (5). This law states that the amount of change in intensity of stimulus - in the above example, light intensity - representing a jnd is constant over all ranges of intensity. Symbolically,

$$\frac{\Delta M}{M} = \text{Constant}$$

where M is the magnitude of stimulus. This law received extensive attention and these constants were measured for many psychological dimensions.

The primary importance of Weber's Law was that it represented the first comprehensive quantitative measurement of sensory judgements. Such measurement is important here as indication of the extent of the individual's ability to detect differences or judge similarity between two stimuli. But it is important to recognize that Weber's law implies no scale of response, no measure of what the subject feels. There is simply the statement that feeling is present.

Fechner introduced the idea of a subjective scale (4) by proposing that all jnd's are subjectively equal. That is, the jnd is the basic unit of subjective measurement. This assumption is applied to Weber's law:

$$\frac{\Delta M}{M} = \text{Constant}$$

then

$$\Delta S = K\frac{\Delta M}{M}$$

where ΔS is the (constant) change in subjective magnitude of sensation, i.e., the response to stimulus. K is a constant of proportionality.

Solving the equation for response, one integrates the expression

$$\frac{\Delta S}{\Delta M} = \frac{K}{M} ;$$

and Fechner's law is given as

$$S = K \log M + a$$

Fechner's book, published in 1860, presented this law and opened the way to the field of psychophysics.

Thurstone (6) put the measurement of subjective response on broader footing by suggesting, in the late 1920's, that perhaps the techniques of psychophysics could be applied to the study of attitudes. "Instead of asking a person which of two cylinders is heavier, we might as well ask something interesting, such as, "Which of these two nationalities do you in general prefer to associate with?' or, 'Which of these two offenses do you consider to be in general the more serious?' or "Which of these two pictures or colored design do you like better?'". This generalization of thought did much to revitalize and extend psychophysics and mental testing.

Objections to the ideas reviewed above have been several, and seem to fall esentially into two classes. The first class includes criticism of the attitudes and assumptions which lead to the laws stated. The philosophical background of psychology was such that in the mid 1800's there were many people who felt that these qualities of subjective response were beyond measure, that the only path to psychological knowledge was through introspective investigation. This feeling waned as time passed. More serious was criticism of the assumptions used. Fechner's law is quite vulnerable because of its basis of equal values of subjective magnitude.

Indeed, although a good deal of data was manipulated in the late 19th century to yield the parameters of Fechner's law, slightly changed assumptions yield different formulas, which also may be supported. Stevens has found such wide verification of the power law

$$S = kM^n$$

where n is a constant, that he suggested that it may be useful in achieving broad concensus on social matters (7,8). He refers to this law as the "psychophysical law" (9).

The second class of criticism is concerned with measurement and scaling methods. Closely allied with the psychophysical law are the experimental techniques of direct estimation or direct scaling. These techniques and the results of their use are basically different from the confusion techniques of Fechner's jnd. The individual is asked to judge the difference between two stimuli, or their relative intensities. Hence numbers are directly applied rather than imputed from the number of jnd's occuring along the scale between two magnitudes (10). The subject might be asked to rate his feelings along a scale from one to five, or to determine what magnitude of stimulus he considers to be twice as strong as the previous one. Thurstone and Stevens have done a great deal of work in demonstrating such procedures for mental testing and psychophysics, respectively (4).

The use of direct scaling has been encouraged by Stevens' extensive analysis of scale types. There are several ways of scaling response, each scale more useful and, as one might expect, more difficult to apply than preceding ones in the hierarchy. Stevens defines four principal types of scale (7). These are the nominal, ordinal, interval, and ratio scales.

The nominal scale, as the term implies, simply assigns names to the elements of the group being scaled. No measurements in the usual sense are implied. The scale is simply a means of identification of differences among elements. Examples of a nominal scale are the numbering of players on an athletic team, or the way a taxonomist classifies plants or animals.

Next in terms of ordering is the ordinal scale. In this scale the progression of names or numbers indicates a set order. For example, successive street numbers tell in what order one might expect the houses to appear. However no other information is implied, and one does not know if successive houses are one foot or one mile apart.

The interval scale not only indicates order, but also expresses the difference between elements in terms of a standard, the unit interval. Thermometers and calendars are examples of the interval scale - given two points on the scale, one may compute the distance between them in terms of the unit interval. A particular characteristic of this scale is that it has no natural origin. The origin is set by choice and the scale defined by adding away from it.

The most powerful of the scales is the ratio scale

This scale of which most physical scales such as mass, density, pressure, and voltage are examples - possesses not only the measuring qualities of the interval scale but also a natural zero point. Each element in the scale is expressed as a ratio of the unit interval.

There are other types of scales which may be hybridized from these four. An example which is of particular interest here is the ordered metric. This scale is an ordinal scale which also has order in the intervals between elements. That is, it is known that the difference between, say, the second and third elements is greater than that between first and second, which in turn is greater than that between the third and fourth, and so on. As no unit of measure, or unit interval, is implied, this is not as strong as an interval scale.

As suggested, each scale, besides being mathematically more powerful and more useful than its predecessor, is more difficult to obtain for psychological parameters. A great deal of work has been done trying to set general rules or techniques for scaling (see for example Winkler (11) or Galanter (12)). But for many of the applications of interest here, the ordered metric scale is adequate. Such a scale is adequate for definition of an aspiration level (13). This level will become more important in later pages.

These scales, and the "laws" discussed earlier, are a means for predicting subjective response to a stimulus.

If the stimulus is measured in terms of some set physical parameter, for example true weight, then on tries to scale the response, in this case perceived heaviness. Questions are asked (sometimes implicitly by measuring some physical response characteristic) to yield the scale desired. "Which weight is heavier"? yields a ratio scale. The utility upon which the serviceability measure is based will be portrayed as a type of response.

Sometimes a response measurement is made in terms of a substitute for the factor of interest. For example, the skin's electrical resistivity is found to vary with the subject's anxiety (a finding used in lie detector tests). Then for a given stimulus, anxiety response is measured by the proxy of electrical resistance. It is said that the resistance value, ohms, is an <u>indicant</u> of anxiety felt (5). Or, anticipating the next topic of discussion, it might be said that the equilibrium price in an economic study is an indicant of desireable qualities of a consumer good.

Thurstone led the way in the step from measurement of response to physical stimulus to the measurement of attitudes, response to social or emotional stimulus. With this short step to the measurement of tastes and values, psychologists found themselves involved in the attempt to predict human choice behavior. How do people's tastes and values influence them in their behavior, in their decision making? Here the development of utility theory in the field of economics was encountered.

2. Utility Theory and Consumers' Behavior

In the middle to late 1800's, a veritable revolution in the field of economics was under way. Samuelson (14) has suggested that if one criterion is to be found to distinguish the modern field which emerged from its classical background, it might be the introduction of the subjective theory of value. Such outstanding economists as Jevons, Walras, and Menger tried to explain comsumer' behavior-motives, decisions, and actions - in terms of the idea that rational men will try to maximize their happiness, or, as Bentham termed it, their utility. The theory was at its height with Edgeworth's Mathematical Psychics, published in 1881 (15).

An assumption basic to the entire theory was that the amount of satisfaction derived from increasing amounts of a commodity increases at a decreasing rate as the total amount of the commodity already possessed increases. (see Figure 1) Note that the previously reviewed laws of psychophysical response (e.g., Fechner's Law and Stevens Psychophysical Law) possess this property of diminishing marginal utility.

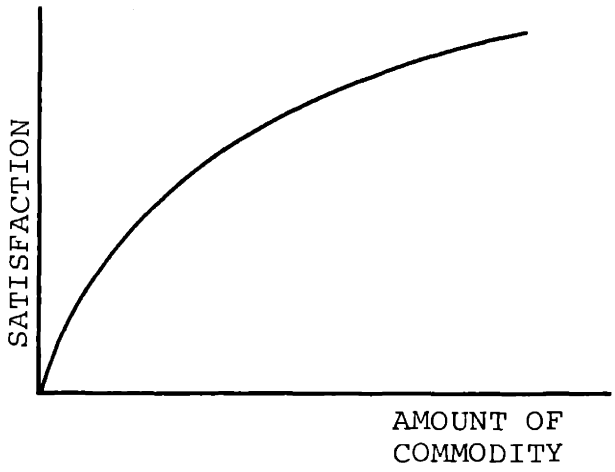


FIGURE 1: Satisfaction from an Economic Commodity

The first use of this principle to explain economic behavior appears to be in the 18th century with D. Bernoulli's discussion of the famous St. Petersburg Paradox (16). Bernoulli proposed that the paradox could be resolved by postulating that a person's marginal utility for money is expressed by

$$\frac{\Delta U}{\Delta M} = \frac{k}{M}$$

where M is the monetary stimulus and U is the utility response. Solving this expression for U yields Fechner's Law,

$$U = k\log M + a,$$

as previously discussed.

Bernoulli's discussion did not receive the widest circulation, and subsequent developments of utility theory in the mid-1800's were apparently independent of this previous work (14). Although a number of the prominent economists of the day tried to suggest that utility could be measured directly and that this cardinal utility could be used in economic analysis, the trend has been away from such strong assumptions. The basic promise which has been retained is that the individual, when confronted with a selection of goods and their prices, will choose to spend his limited income on that mixture of goods which will yield the greatest satisfaction. That is, he will maximize his utility.

By the beginning of the 1900's the idea of cardinal utility was losing favor among economists, most of whom were asserting that it is only necessary than an ordinal preference field exist. Pareto in particular may be noted for his extensive use of this assumption (15).

It should be noted that throughout this development, only one detail seemed to separate the views of economist and psychologist (5). This is the economist's interest in a theory describing what the rational man would do rather than what the actual man does. Derivation of such a theory often required assumption of the very things psychologists were trying to measure.

With the economist's growth away from strong assumptions of utility has come the psychologist's interest in economic behavior. For example, as previously mentioned, the psychophysical law has been found to be quite broadly applicable in economics. Galanter (12) in fact found that the utility of money in a gamble could be expressed as

$$U = 3.71\ M^{0.43}$$

when there are no losses involved. The work in the two fields has converged substantially to yield data in common areas, of use in a variety of social, political, and economic situations. It is this convergence which suggests that the user's response to service provided by systems of constructed facilities may be described in terms of utility.

3. Multi-Dimensionality and Attributes Space

It is apparent that one considers many factors in perceiving and judging a complex stimulus such as the service of a system of constructed facilities. There is thus a need to consider the possibility that utility might be defined on several scales, and how several scales might be related to yield a feeling of satisfaction of dissatisfaction.

There is much evidence to indicate that people making choices will, given time, recognize separate attributes of the alternatives (16). There also is evidence that to some extent people perceive different types of utility for these various attributes, but that on each individual attribute scale, the descriptions of utility presented earlier are realistic. Thus, it may be postulated that an alternative with multiple attributes will have a multi-dimensional utility space, where each attribute (stimulus) axis, ignoring the others, will have utility curves as discussed. If this is so, how do the different utility measures interact to give a basis for final judgement?

The earliest attempts to answer this question were made by economists trying to predict the mix of commodities on which a consumer would spend his fixed income. If there are N commodities, it was suggested that there were N utility functions u_i, and that the utility of the product mix was the sum of these functions (5)

$$U = u_1 + u_2 + \ldots + u_N$$

The consumer would try to maximize total utility. More of any commodity could be bought, thus increasing one utility quantity, but money was limited and marginal utility decreased with quantity. Hence, there was a problem to be solved in the maximization. The solution is the point at which the ratios of marginal utility for each commodity to its price are all equal (14).

A very popular use of such additive utility has been in management decision theory. In this application, the utility theory is intended to be prescriptive (17), that is, to tell the manager what the rational decision should be, given his basic set of values. Recognizing the uncertainty of future events, and hence the uncertainty of utility arising from the results of a decision, the measurement has been placed in a statistical context. One tries to maximize expected utility.

$$U = u_1p_1 + u_2p_2 + \ldots + u_Np_N$$

where u_i and p_i are the utility and probability of occurance of possible outcomes of a particular action. Quite a body of theoretical work has been developed to utilize this approach (see for example Von Neumann and Morgenstein (18) or Luce (19)) and for strictly monetary situations, experimental evidence suggests that expected utility may indeed be so maximized (20).

When one is dealing with multi-attributed choices, however, choices which cannot be reduced to purely monetary terms, the situation is different. It is the old matter of apples and oranges - if one simply wants fruit, the two may be added. If one considers the differences, choice is more difficult. In fact, the problem is one of substantial current interest in psychology and marketing analysis. It is now generally agreed that the individual scales of utility (as suggested in Figure 6) are not generally separable for the purposes of measurement, that the entire perceptual space must be considered at once.

The most effective approach to this problem seems to be the Coombsian model of attribute space. Coombs (21) suggests that the attributes of a commodity or choice alternative - the stimulus - form a multi-dimensional Euclidean space. Each alternative is represented as a point within this space. Each person has an ideal set of values of these attributes against which he compares the alternatives, and thus each person is represented in the attribute space by an ideal point. The ideal point is the combination of attributes which the individual would tend to prefer to all others. The particular alternative which an individual would choose is predicted as the one which maps into a point closest to the individual's ideal point in the attribute space.

In the symbolism of previous discussion, the model suggests that there is in fact an attribute space $[Z_i]$ within which a constructed facility might be represented as a point. The vector $\overline{Z}m$ would give such a point. The promising feature of the techniques developed, based upon this model, is that the space may be found directly from sample stimuli, without reference to the intervening variables $[y_j]$ or an immediate need to measure parameters $[x_k]$. It would perhaps be worthwhile to describe briefly in which an experimental determination of the attribute space is carried out.

A set of experimental objects are presented to a subject. In the cases where these techniques have been used, small market items such as toothpaste or magazines, it has been possible to show the actual alternatives directly to the user. For systems of constructed facilities, the use of photographs, architectural renderings, or full-scale models might be satisfactory. (See, for example, the preliminary work of Winkel (22).

The subject (user) may be asked to select the alternative which he likes best from among the set. He is then asked to compare the other items, a pair at a time, with this preferred alternative and to indicate which of the pair is not like the preferred item. The comparisons are made for all possible pairs of items. This procedure is termed the method of paired comparisons.

The set of mathematical inequalities which these judgements represent are then input to a computerized algorithm which will produce a multi-dimensional mathematical function. This function will reproduce the similarity judgements made by the subject, in terms of inter-point distance, and suggests the number of attributes being considered by the subject. Shepard, in 1962, made a significant breakthrough by producing the first computer algorithm to successfully produce such a function (23). Several other techniques have since been developed, and the connection of such procedures with statistical factor analysis has been demonstrated (24).

Examining the predictive function produced, one can find measurable qualities of the samples to serve as indicants of the attributes. That is, knowing the end points of the problem, the attribute space $[Z_i]$ and the actual samples, one can attempt to reconstruct the intermediate parts $[y_j]$, $B(y)$, $[x_k]$, $A(x)$. While these methods are still fairly novel, and certainly untried for cases such as a system of constructed facilities, they offer a promising alternative to the psychological scaling procedures reviewed previously.

4. Synthesis: The Typical Utility Function and Aspiration Levels

The preceding sections have, by means of historical review, laid the groundwork for developing the concept of serviceability presented earlier in this chapter. This section will present a synthesis first of the ideas of utility as a means for characterizing response and then two possible approaches to finding the information required for estimation of serviceability.

It has been shown that a fairly broad range of so-called stimuli may be investigated using a concept oı internal subjective response, or utility. It is suggested here that the user's perception and judgements of service provided by systems of constructed facilities may be characterized in this same fashion. This characterization provides a common basis for a diverse set of factors which the constructed facility should have.

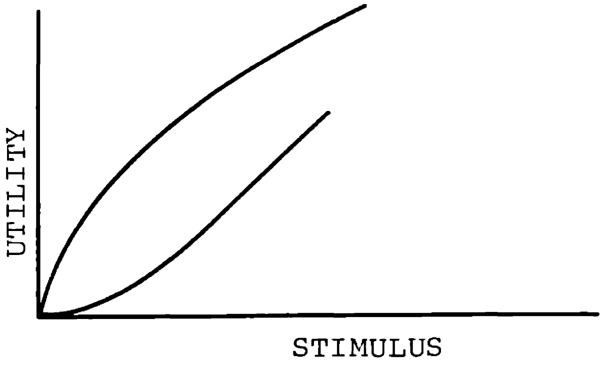


FIGURE 2: Utility "Laws"

While much of the work on various aspects of utility has suggested utility functions of the form shown in Figure 2, such functions are found primarily as a result of working in one direction from an assigned origin. It has been found (12, 26) that over a sufficiently broad range of stimulus, a point of influction will be found, and a more general shape for the utility function would be similar to that shown in Figure 3.

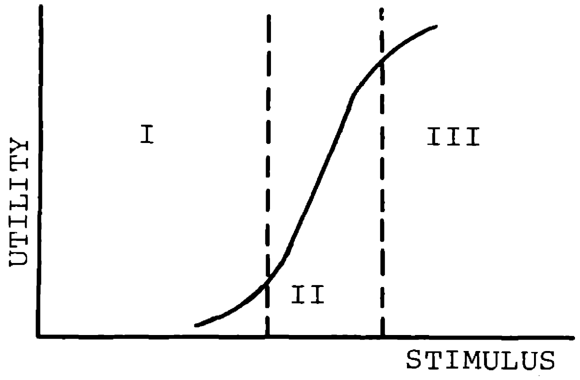


FIGURE 3: The Complete Utility Function

This typical s-curve may be discussed in terms of three regions. Region I is a range of small stimulus, in the neighborhood of the threshold of perception. Examples falling in this range might be small changes in the background noise level in a house, or the difference of a few pennies in the price of an expensive auto. That is, the subject perceives only fairly sizable absolute changes in the stimulus and utility rises quite slowly.

In region II, the subject is confronted with a stimulus which he can perceive and judge with a high degree of discrimination. Region II is an area of maximum sensitivity, influenced by physiological and psychological preconditioning, i.e., by what one is familiar with.

Region II may be thought of as containing a natural origin of judgement, a point about which valid judgements may be made. Such a point may then be viewed as the origin of the laws presented previously. Work by Galanter (12) is an example of utility measured in both directions from this middle ground.

In region III, the subject has reached saturation, an inability to consider relatively small changes in stimulus. For example, the desirability of two large sums of money, both beyond the familiarity of the subject, will yield small variations in utility. The overall utility or feelings of the subject may be characterized as relatively uniform.

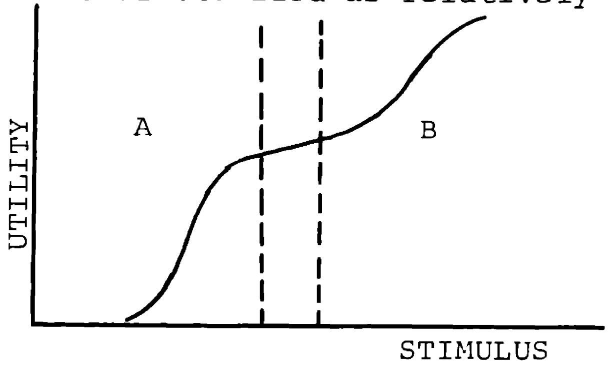


FIGURE 4: A Dual Range Function

In some cases, where several points of inflection are observed in a utility curve, as shown in Figure 4, a change in the nature of response is postulated (25). For example, if income is the stimulus, the curve might be the perceived utility of income for a poor man (Region A) and for that same man when he becomes rich (Region B). Income falling in the area between regions A and B is so high that the poor man cannot really judge; he is happy and feels diminishing marginal utility. Given the opportunity to enter Region B, however, the man's views change and he strives for higher income.

While the presentation of utility in the manner of Figure 4 suggests that actual measurements may be made over a broad range of stimulus, such measurements are difficult to obtain and have always been subject to questions of validity. In fact, it is felt by some people that, while the function pictured may exist, to suggest that it is measurable requires unwarranted assumptions (25).

(See Figures 5 and 6 on Page 94)

However, for the purposes of the serviceability measure which has been suggested, only one point is necessary--the aspiration level (27). (See Figure 5). The region of rapidly increasing utility (Region II in Figure 3) may be termed the critical region. It gives the range in which the subject is most sensitive to changes in stimulus. The subject is most able to make judgements in this area. The aspiration level is associated with stimulus level that is the upper bound on this rapid rise in utility. For a discrete valued utility function, this level of aspiration is located by the point at the upper bound (26). On a continuous function, the aspiration level is located at the point of maximum

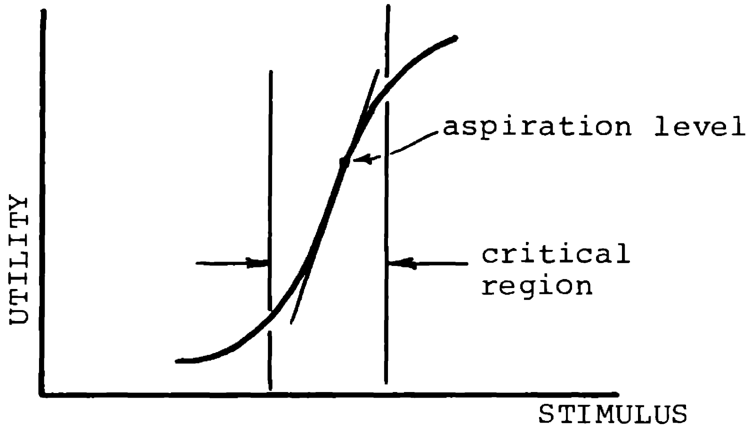


FIGURE 5: Critical region and aspiration level

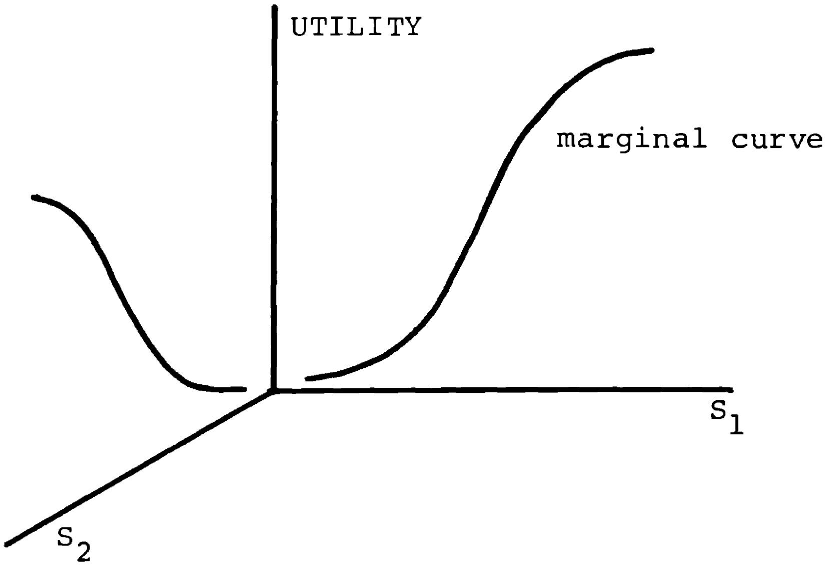


FIGURE 6: Multiple stimuli

slope (27).

The aspiration level is identified with the idea of achievement of a goal. It is found that the subject will be generally satisfied (please, complacent,...) with a level of stimulus above the aspiration level (27). This idea is useful as a decision criterion, and is the basis for the serviceability measure presented here.

It is observed that the aspiration level is variable, changing in the individual with time and experience. It rises with success and goes down with failure. Cases of this pehnomenon are familiar; one will tolerate more trouble with an old car than with a new one; students who have always done only average work do not strive for higher grades with the same intensity of those who are used to high scores.

In the case of problem solving, it is suggested that the aspiration level, the goal to be achieved, will lower as difficulties of finding any solution increase (28). It might be generalized that the aspiration level will depend upon part history and current expectations.

One may easily extend the psychological idea of aspiration level into the realm of economics by postulating that the aspiration level is reflected in the individual's demand curve by the decision to buy at the quoted price. Then the demand curve shifts when tastes change. Leibenstein (29) identifies several external effects on utility as changes in individual demands caused by the actions of other individuals.

These economic effects are a matter of the individual acquiring more of the stimulus commodity in reaction to a shifted aspiration level. On the other hand, the individual may shift his aspiration level when the goal cannot be achieved. Such behavior is commonly observed in response to the physical environment (30), where individuals will adapt to initially undesirable conditions. This effect is especially pertinent to discussions of the evaluation of slum housing conditions. It is suggested here that a particular individual, at a particular time, will judge the service provided by a system of constructed facilities to be satisfactory if he perceives this service to fall above his aspiration level. It is thus necessary to find only this aspiration level to ascertain user satisfaction.

This is satisfaction for an individual. For a group of users, it would be expected that there would be a distribution of aspiration levels. Discovery of the statistical characteristics of this distribution will yield and estimate of the serviceability function S, which is the probability that a user will judge $Z \geq Z_m{}^*$, and thus find the service to be adequate.

5. Synthesis: Determining the Serviceability Function

There are three primary pieces of information which must be found if the serviceability function is to be determined. First, one must know the dimensions of the users' judgement space Z_i. Then, one must have a means for characterizing a facility in terms of Z. Finally, one must find the users' reaction as related to Z and thus estimate $S(\overline{Z})$. There were two basic approaches to these problems reviewed or alluded to in previous sections. They are discussed below explicitly.

The first approach is to use the techniques based on the Coombsian model of attribute space. This approach, though untried in this area, is desirable because of its directness. One begins by applying an experimental technique such as the method of paired comparisons. At the same time, the subject would be asked whether the particular example under consideration is acceptable. Analysis of this data would yield the dimensionality of the attribute space, using the previously described computerized algorithms, and with the added question would permit one to plot a surface enclosing all of those points which were felt to be acceptable. This gives Z_m*.

Applying the technique to a representative group of users will permit computation of a norm, or average, of judgement (24) for each experimental item. This norm is the vector Z characterizing a facility. Then, from this distribution of ideal points and their associated surfaces of acceptability, one can find S(Z).

To fully implement the analysis, one must then find a set of parameters $[x_k]$ which will serve to predict Z. This may be done using statistical correlation techniques. It may be possible in this way to predict Z directly as a function of indicants,

$$Z = B[A(X)]$$

without making direct references to intervening variables $[y_j]$. For example, it is not necessary to know that temperature and humidity determine a quantity called effective temperature, which is correlated with comfort, if one can immediately predict serviceability as a function of the two basic parameters.

Proceeding in this fashion to some extent obviates the need to worry about interactions among parameters, which may effect users' judgements. That is, a facility which has service on one particular component in $[Z_i]$, will perhaps be considered satisfactory because of very high values on other components. The direct determination of $[Z_i]$ from examples takes account of such interactions.

The second approach, which does not have this last feature, is to try to describe $[Z_i]$ from other sources. In so doing, possibly appropriate parameters $[y_j]$ will be suggested. Measurements of S will then be made on individual scales $S_i(y_j)$, which give serviceability relative to the one component Z_i, with the others held constant. (Figure 6). In some cases, the serviceability function so obtained will be considered a good representation of the users' judgement; in other cases, it is simply the best available approximation.

(See Figure 6 on Page 94)

The identification of $[Z_i]$ is undertaken through reviews of literature, discussions with users, and introspective analyses of what one considers to be important. One will try to catalog all of the factors which apparently concern the user, and will then try to synthesize a set of component scales which adequately concern the user, and will then try to synthesize a set of component scales which adequately cover these multifarious factors.

The synthesis may be facilitated by computerized techniques for decomposition of data (for example, Alexander (31) or Milne (32); see Appendix A). The group of factors and their relations to one another are submitted to an algorithm which breaks the group into small subgroups of relatively highly related factors. In recombining the subgroups, one has an opportunity to see certain unifying characteristics which may suggest the desired components of serviceability. These characteristics are hidden from easy view in the myriad of discrete factors. This approach was used in an example of urban housing, to be discussed in an appendix.

When these components have been developed, serviceability subscales S_i may be found by using the scaling techniques discussed previously, linked with a question of whether the particular situation is satisfactory. This approach was used in the example of highway pavements discussed in a later chapter.

The principal advantages of this second approach is its relative simplicity. The amounts of experimental data and computation required are much smaller than in the previous case. At the same time, the implicit assumption of separability of the serviceability function appears to be not too inaccurate in many cases. For example, the structural integrity of a pavement, as long as it is high, will have little effect on rideability or on the user's opinions of whether rideability is acceptable.

With either approach, the desired output is a way to estimate serviceability as a function of characteristics of the system of constructed facilities. Whether this estimate is derived directly as $S(X)$, or via transformations $y = A(x)$ and $Z = B(y)$, the result is a prediction of the probability that a user will find the service of the facility to be acceptable. Alternatively, for a known population of users, this is a measure of the percentage of this population who would be satisfied. Some particular aspects of this function may now be explored.

D. Some Aspects of the Serviceability Function

1. Trade-offs Among Judgement Subscales

Serviceability has been described in terms of a multi-dimensional function of users' perceptions and judgements of physical behavior. That is, each subscale of the assessed servicability function provides an estimate of the probability of user satisfaction with that single aspect of service behavior, effectively disregarding judgements on other subscales. There is, however, no reason to suppose that in general no interaction of subscales occurs in judgement, that there are no tradeoffs among judgemental variables. One may try to explore the nature of these tradeoffs.

At one extreme, it might be assumed that the user considers no interaction - no tradeoff - among factors $[Z_i]$. It may then be postulated that the overall serviceability of a system will be equal to the serviceability on the lowest subscale. That is:

$$S = (S_i')_{min},$$

or a chain is as strong as its weakest link.

If there is any interaction, such that high ratings on some subscales will compensate for lower ratings on some sub-scales will compensate for lower ratings on others, this statement would seem to be valid as a limit:

$$S \geq (S_i')_{min}$$

That is, the overall serviceability of a system is at least as high as the lowest subscale. Hence, the indirect approach to developing the serviceability function gives at a worst a lower limit of the probability that the original problem of equations (4a) is solved.

(See Figure 7 on Page 103)

Figure 7 suggests visually what this result means, for the case of the two factors. The marginal functions $S_i'(y_i)$ are perhaps typical, but are not intended to suggest any general conclusions about the form of such component subscale functions. It may be concluded that the contour of the points (y_1, y_2) such that $S(Z) = b$ (assuming trade-off) lies entirely on or outside of the angular figure defined by $S = (S_i')_{min}$ and $(S_i')_{min} = b$.

This statement suggests that if one is trying to allocate resources, with no knowledge of interactions of the judgement factors, a suitable strategy would be to maximize the minimum: i.e., maximize $(S_i')_{min}$ as an objective function. This strategy is referred to as a "max-ranking"

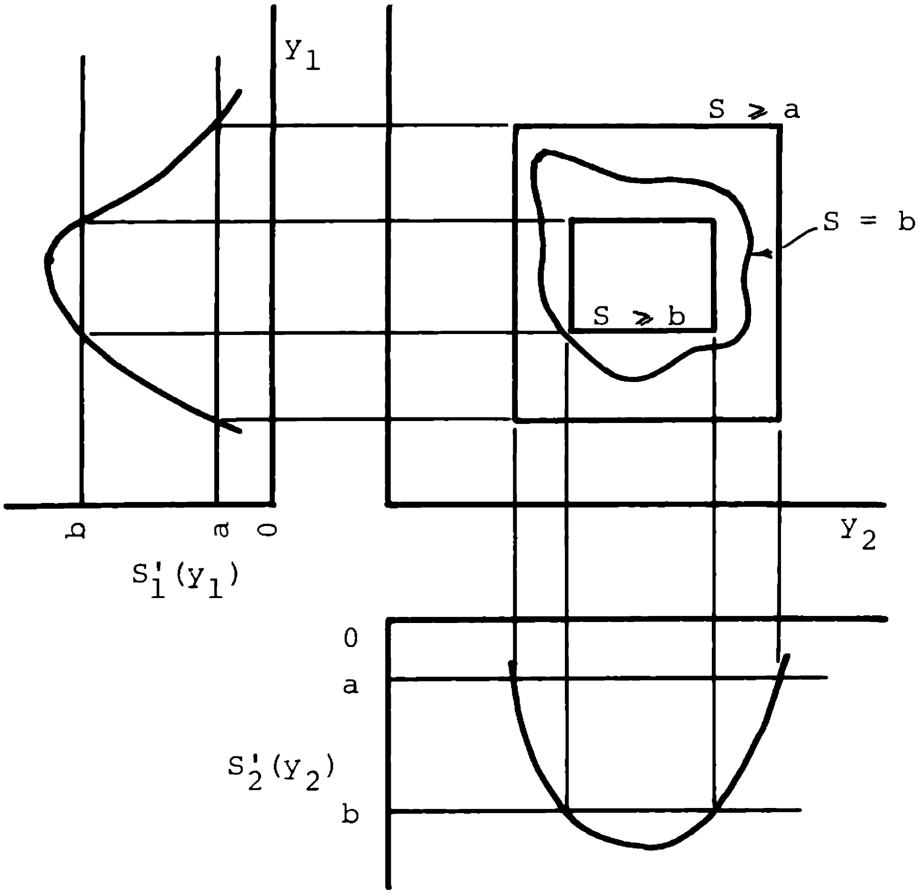


FIGURE 7: Limits of the serviceability function

criterion (33), and leads to a leveling of the ratings on individual subscales.

In some cases, one may feel that particular components of $[Z_i]$ are more important than others. That is, these components are considered to be dominant. In this case, higher values of serviceability will be desired on the dominant sub-scales S_i'. The first stage of resource allocation would then be to assure that the facility will have high enought values on these scales, after which the other scales could be optimized as before. This approach is analogous to what Simon terms "satisficing" (28), that is, finding an action which is at least good enough, if not the best.

The above is one extreme of assumptions. At the other extreme one may adopt Coombs' original view (24) that all individuals consider a perfect tradeoff among factors, and that this is reflected in decisions based upon inter-point distances. That is, a person's ideal point is given as, say, y_m^{**} and judgement made on the basis of the distance $|y-y_m^{**}|$.

Coombs' original assumptions, modified for individual differences of perception by Horan (34), suggest that it will always be possible to transform the rectangles of Figure 7 to squares, as in Figure 8. Then, overall serviceability is defined by the circle inscribing this square.

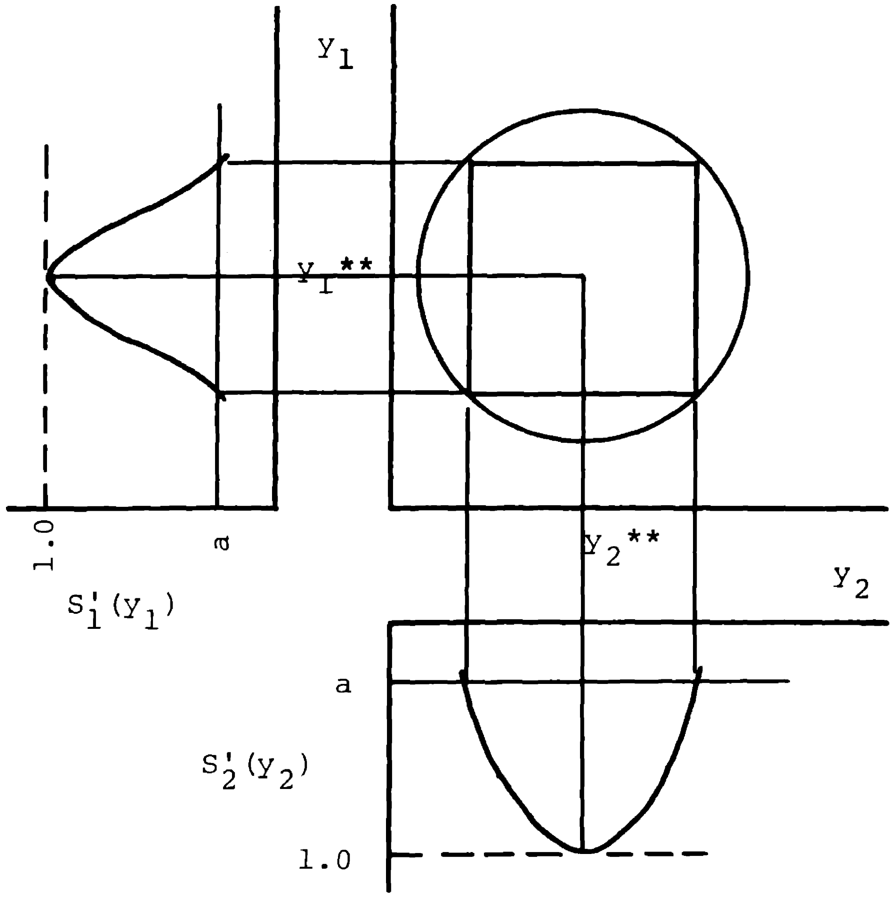


FIGURE 8 : The Coombs model

The impact of this assumption is that in the regions of $[z_i]$ where such perfect tradeoff is possible, it is possible to collapse the multi-dimensional serviceability function into a single dimension. This would simplify the entire evaluation and design decision problem by eliminating the difficulties associated with vector representation. For example, optimization of resource allocation would become an application of the well-studied techniques of quadratic programming (35).

The two cases presented are extremes, and it may be suggested that the actual situation in user response lies somewhere in between. For example, it seems reasonable to suspect that tradeoff might be acceptable within subsets of the space $[z_i]$, provided that behavior on all scales is within some particular range. This is an area of inquiry in need of much work and offering possibilities for contribution.

In view of the lack of knowledge in this area, it must be stressed that on of the major points to be made in this analysis is that the service behavior of systems of constructed facilities will require a number of dimensions in its description. To reduce a description of service to a single parameter is at this stage of understanding inappropriate, to say the least. To attempt to do so will only obscure the issues involved.

2. The Probability Basis of Serviceability

The principles upon which the measure of serviceability is based, and the interpretation of serviceability as a probabilistic variable may imply something about the behavior of the serviceability function. In effect, the user's utility response curve is replaced by a step function of acceptance, with the step at the aspiration level. If one is investigating an arbitrary group of users, it might be assumed that there is a small uniform probability of occurrance of this step in any single, correspondingly small, region of the judgement scale.

The number of users satisfied at a particular level of z_i will then be predicted as a binomial distribution of probability. If there are M users in the group, then the serviceability function becomes

$$S(Z) = \frac{\bar{n}_Z}{M}$$

where $\bar{n}_Z$ is the predicted number of satisfied at or below a given Z. For large M, the binomial distribution is approximated by a normal distribution. That is, in such cases it might be expected that serviceability will be predicted by a normally distributed random variable over z. It will be noted that the data on serviceability with respect to rideability in the case of highway pavements closely fits this conclusion (Chapter VI).

E. Problems of Application

1. Making Evaluations

This section presents a short discussion, in more explicit terms than previously used, of some of the operational problems involved and possibilities for solution. These problems may be viewed as falling roughly into two categories: There are general problems of developing the tools and techniques to permit better estimation of serviceability, and there are the problems associated with analysis of particular types of facilities.

On the general level, a number of experimental methods have been developed in the fiels of psychometrics and psychophysics which, as has been mentioned, might prove of value for the analysis of systems of constructed facilities. Indeed, when linked with separate structure-finding techniques, some of the more basic of these methods have been applied with reasonable success, for example in the evaluation of highway rideability (Chapter VI) and certain aspects of comfort in housing (Appendix D). A good deal more work will be required, however, before such techniques can be considered standard in this application.

Of special interest are the non-metric scaling techniques previously discussed. It is here recommended that a rather extensive investigation of the application of such techniques to systems of constructed facilities would be a highly worth while endeavor.

On the level of particular types of facilities, the problems of principal importance are those which hinder immediate estimation of serviceability (as contrasted with the more general problem of accuracy). These problems are primarily of presentation and measurement of samples for evaluation. In the case of the AASHO Road Test (Chapter VI), subjects were taken to a number of different locations to observe and judge highway pavements. Such a procedure is time-consuming, and in cases such as urban housing, where longer term impressions are relatively more important than in highways, of questionable validity. Other means of presenting the stimulus are desired. Work such as that of Winkel (22) involving pictorial display, may prove useful. This is an area in which a great deal of work may yet be done.

2. A Comparison with Consumer's Surplus

The relation of serviceability to demand has appeared as a point for discussion several times in preceeding pages. One more aspect of the comparison requires consideration. The idea of consumers' surplus as a measure of relative desireability of one alternative over another has gained attention (see Bhatt (36)) and might be related to the serviceability measure.

Consumers' surplus is based upon the idea that at a given price, there are people buying a product who would be willing to pay more, and are receiving a benefit because

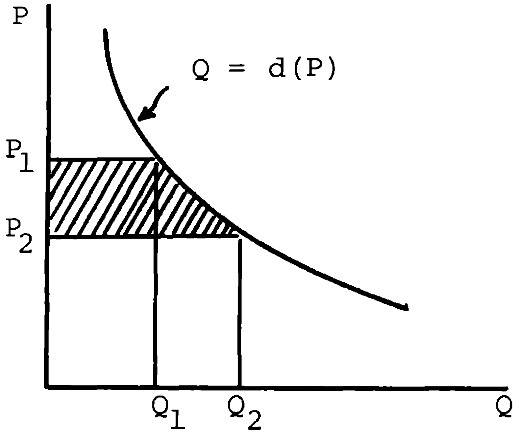


FIGURE 9: Changes in Consumers' Surplus with Price

they do not have to. Figure 9 shows a simple demand curve of price versus quantity desired. At a price P_1, Q_1 units are demanded. If the price is lowered to P_2, the first purchasers are saving $(P_1 - P_2)Q_1$, for they would have been willing to pay that much more for Q_1 units. This quantity is their gain in consumers' surplus. In addition, a larger quantity will be purchased, each unit being bought at a price lower than that which the purchaser would have been willing to pay, up to the last unit. At this last unit the price is just equal to that which the last purchaser is willing to pay. The total increase in consumers' surplus is then the shaded area indicated in the diagram.

In the more complex and realistic situation, the demand function depends upon a number of factors, so that consumers' surplus may change even if the actual price paid is held constant. As suggested in Chapter I and Appendix C, a change in physical service qualities could be equivalent to a change in price, in terms of having an effect on demand.

Raising the serviceability of a constructed facility would effectively move the demand curve to the right at any given price, a greater amount of service will be desired if

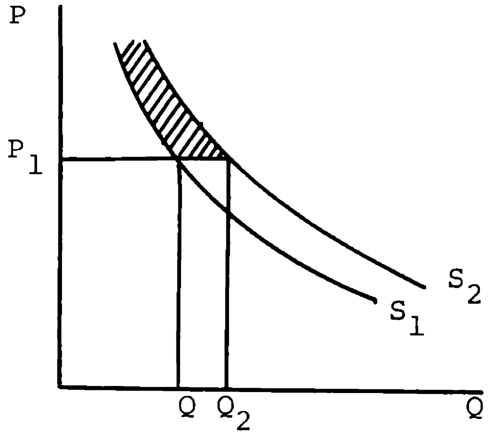


FIGURE 10: Consumers Surplus and Service Quality

quality is raised. Increased consumers' surplus is indicated by the area above the price line and between the two curves. In this view, an increase in serviceability will indicate an increase in consumers' surplus, other things being equal.

As developed herein, the serviceability function will give an estimate of how far to the right the demand curve moves as physical system parameters improve, given the range of psychological and economic factors which influence this function, including expectations and past experience about the system. It is impossible however, without specifying a serviceability function at every price level, to say exactly how far the movement of the curve might be. Such a specification would be equivalent to finding the complete demand function.

In the absence of such specification, the actual increase in consumers' surplus cannot be estimated unless the planning analyst is willing to make some assumptions. For example, if the shape of the demand curve is estimated to remain constant as serviceability changes, then a rough estimate of change in consumers' surplus is obtainable. Such assumptions cannot be recommended as a great deal of work will be needed to investigate this relation between planning and demand decisions.

F. Summary

This chapter has been devoted to the description of serviceability as a measure of effectiveness for constructed facilities and with ways in which this parameter may be predicted. Serviceability was defined as a measure of the degree to which satisfactory service is provided to the user, from the user's point of view. This probability is estimated as the fraction of users judging a facility's qualities of service to be adequate.

In practice, this parameter is estimated in terms of serviceability subscales, predicting fraction of users satisfied with respect to indicants of apparently independent aspects of service. For example, it will be shown in Chapter V that the serviceability of highway pavements may be evaluated with respect to quality of ride, safety, and structural integrity. Hence, serviceability emerges in practice as a multidimensional parameter for the evaluation of present physical service qualities of a facility.

Two approaches to the measurement of serviceability. The first approach begins with the identification of component subscales, followed by scaling along each of these subscales. It should be established, to the degree that it is possible to do so, that these subscales are independent. It was suggested that computerized algorithms for investigating problem structures will be helpful in identifying relatively independent subscales, and that relatively standard psychophysical and psychometric techniques are available for subsequent serviceability evaluation. The use of this approach is illustrated in Chapter V.

The second approach depends upon the application of newer experimental techniques referred to as non-metric scaling methods. Here, identification of components subscales and measurement of response are undertaken simultaneously, and the user is not required to make any numerical judgements about services. Here too the computer service as an aid, in this case quite important, in following this approach.

The discussion of serviceability tacitly assumes that the physical characteristics of service may be predicted, and that they are known with certainty. This is not, however, the case. The physical system's behavior is in fact highly uncertain, and predictions must be made in a probabilistic fashion. Reliability and maintainability are thus proposed as measures of effectiveness of the system of constructed facilities, and will be discussed in the next chapter.

CHAPTER IV

RELIABILITY AND MAINTAINABILITY: MEANING AND MEASUREMENT

A. Introduction

The previous chapter discussed the user's response to the service of a system of constructed facilities. Throughout the discussion, it was implicitly assumed that one could actually (and accurately) measure the characteristics $[\chi_k]$ of a system. In fact, it is not so easy to be certain of $\bar{\chi}$ for a particular facility at a given time. And even worse, the facility's behavior must be predicted for future times.

In short, the physical service behavior of a system of constructed facilities is essentially uncertain. Both system characteristics and environment can be predicted at best in only a stochastic manner, in terms of probabilities. It is suggested here that design decisions for systems of constructed facilities must be made with an awareness of these uncertainties. Reliability and maintainability will, as components of performance, reflect these needs.

Reliability is defined as a measure of the probability that a facility will not fail during its design life, that its physical service will remain adequate. This measure in effect preceeds serviceability, which takes the physical service as given. But, as has been shown, reliability also depends upon serviceability for its definition.

A system of constructed facilities generally has a fairly long service life. This aggrevates the problems of uncertainty because of the increased difficulties of making predictions over longer periods of time. Further reliance must generally be placed upon actions to be undertaken during the facility's service life - i.e., operating and maintenance activities. Maintainability is proposed as a measure of the degree to which the service behavior depends upon such continued effort throughout the design service life of the system of constructed facilities.

This chapter will discuss the problems of uncertainty and prediction, and the meaning and use of reliability and maintainability as measures of effectiveness. Section B will try to present a picture of the major sources of uncertainty. Section C will discuss the definition of reliability and two approaches to its computation including consideration of the problem of modeling service life behavior. Section D is devoted to maintainability. The summary in Section E serves as a quick review of these ideas and a comment upon using them.

B. Sources of Uncertainty

In the previous chapter, it was suggested that the serviceability of a system of constructed facilities depends ultimately upon characteristics of that system's physical service behavior, designated $[\chi_k]$. It is now suggested that these service characteristics will depend upon some interaction of the system with its service loads and environmental qualities, as proposed visually in Figure 1.

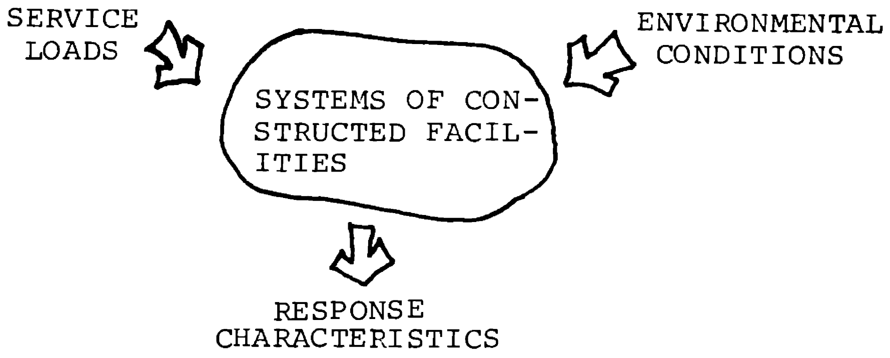


FIGURE 1: Service Behavior of Constructed Facilities

It is convenient to denote this interaction symbolically as

$$\bar{\chi} = T(c,e)$$

χ is, as before, the service description which is useful for predicting serviceability. c and e are descriptions of the characteristics of a system of constructed facilities (its service capabilities) and the environment (including all so-called service loads), respectively. The function T might be termed the technology of the system of constructed facilities, which predicts behavior as a function of loads and capabilities.

For example, χ for a highway pavement might include a measure of the amount of permanent deformation at the surface of the pavement. Then c would include such factors as materials' strengths and moduli and layer thicknesses. Similarly, e will perhaps include magnitude and configuration of vehicle loads, temperatures, and rainfall. The technology function T will comprise equations predicting deformation as a function of the chosen c and anticipated e, and would assume differing forms, dependent upon whether the pavement is considered to be flexible or rigid.

The sources of uncertainty in the system may be viewed in this context as residing in the three factors c, e, and t. That is, system capabilities c can be only imperfectly controlled and are subject to natural dispersion of their values. Loads e must be predicted for the future service of the facility, and are also dispersed. Finally, predictions of behavior are only as good as the models by which they are made, and the technology descriptions used in practice are always approximations or idealizations of the so-called real world.

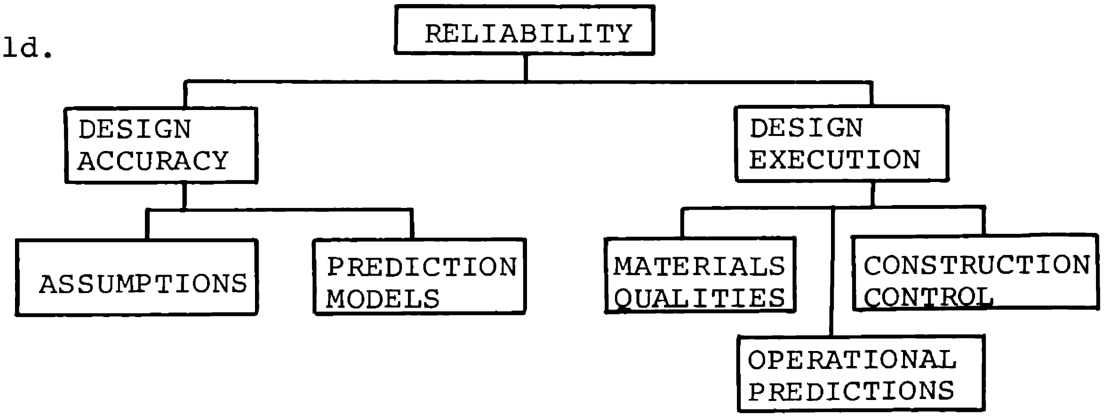


FIGURE 2: Components of Reliability

Figure 2 suggests the major aspects of these uncertainties, and also the basic separation between the uncertainties of T and those of (c,e). The uncertainties of T are controlled by the accuracy of both assumptions made in the analysis and of the models derived from these assumptions. Predictions will be only as good as the models used to make them. This source of uncertainty must be recognized.

However, this source of uncertainty is quite difficult, if not impossible, to evaluate without extensive experience. Often, the only estimate of a model's accuracy and validity will be personal judgement. This factor will be neglected in present discussion, for it involves many questions beyond the scope of this work. It will be assumed that if c and e are known with certainty, χ can be predicted with certainty. This assumption could be relaxed without changing the essential form of any of the following arguments, and is made primarily to limit the size of the discussion.

The uncertainties of c and e may be seen as occuring in the execution of a design, i.e., in the implementation and operation of the system. Uncertainties arise from natural variations in materials, from the ways in which actions are undertaken (not only in construction as it is usually understood, but also in maintenance), and from the possible variations in future operating conditions, in terms of loads and environment.

It is suggested that these uncertainties are best

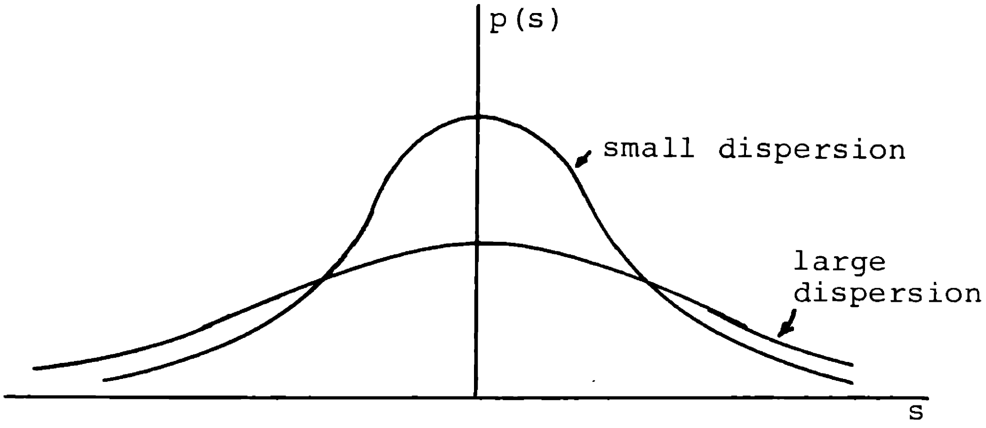


FIGURE 3: Portrayal of Uncertainty

considered in terms of the probabilities of occurance of the conditions in question. Greater uncertainties are reflected in greater dispersion of possible values, and thus in correspondingly lower probabilities over a given interval. (See Figure 3). The distributions of probability of occurance of c and e may be utilized to estimate the probable values of χ, and thus the reliability of the system of constructed facilities.

C. The Nature and Use of Reliability

1. Defining Failure

The first question which must be answered is "what constitutes failure for a system of constructed facilities"? The answer lies in the serviceability function and in the role which the constructed facility plays in its interactions with social, political, and economic subsystems.

For the individual user, failure occurs if $Z_m < Z_m*$. The serviceability of the system, S(Z), has been defined as the probability that this individual failure does not occur. It was further suggested that one will perhaps wish to raise S(Z) as high as possible, subject to constraints on scarce resources.

To review quickly, it is proposed that there will be, for any particular system of constructed facilities, a minimum level of achievement S_f which may be termed the failure level. That is, if the overall serviceability of a system of constructed facilities is such that

$$S(Z)<S_f$$

this system will have failed, as far as the analysis is concerned. S_f is the minimum acceptable probability of user satisfaction with a given set of service characteristics.

It was proposed in the last chapter that, given no particular knowledge about the form of S(Z), it could be said that

$$S\geq(S_i)_{min}$$

It may then be said that, as a limit, if serviceability on the lowest rated subscale falls below the failure level, system failure occurs. Applying this criterion to each of the i judgement scales in $[Z_i]$, one derives a statement of I _failure modes_, $S_f>S_i$, i = 1,2,...,I.

Each of these failure modes will imply certain conditions χ_i* (not necessarily single values), defining failure in the system in terms of behavioral characteristics. These conditions will in turn imply relationships between c and e, according to the technology description $\chi = T(c,e)$, which will determine failure. One may conceptually express these relationships in terms of demands placed on the system $D_i(e_n)$ and the ability of the system to resist these demands $R_i(c_m)$. The occurrence of χ_i*, indicating failure, is designated as

$$R_i(C_m) < D_i(e_n)$$

Failure is said to occur in the i'th mode.

For example, one aspect of serviceability of highway pavements may be termed rideability, referring to the quality of ride that pavement provides. The demands made upon the system may be characterized in terms of the accumulated total of equivalent wheel loads. The resistance of the system, for a given loss of rideability, may be predicted as a function of materials strengths and layer thicknesses. Then if the predicted demand loads exceed the number it is expected that the pavement can resist before that given loss of rideability occurs, failure may be expected.

In general one considers I failure modes in this fashion. Satisfactory service is rendered when the following inequalities hold (1):

$$R_1(c_1,c_2,\ldots,c_M) \geq D_1(e_1,e_2,\ldots,e_N)$$

$$R_2(c_1,c_2,\ldots,c_M) \geq D_2(e_1,e_2,\ldots,e_N)$$

$$\begin{array}{lcl} R_3(\qquad\qquad\qquad) & \geq & D_3(\qquad\qquad\qquad) \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ R_I(\qquad\qquad\qquad) & \geq & D_I(\qquad\qquad\qquad) \end{array} \qquad (1)$$

For convenience and completeness, the entire sets $[c_M]$ and $[e_N]$ are included in the I inequalities, although each expression will generally have its own subset of parameters upon which it depends.

Reliability is then defined as the probability that all of these inequalities hold. That is,

$$\begin{aligned} R &= P[R_i \geq D_i] \quad i = 1,\ldots,I \\ &= P[\text{no failure occurs}] \end{aligned} \qquad (2)$$

It should be pointed out that the failure level of serviceability need not be the same on all scales $[Z_i]$. It has been mentioned that there may be dominance relations among the subscales of serviceability, such that higher values are desired on some S_i' than others. This situation will not change the basic argument or the form of this definition of reliability.

One may note the similarity between the definitions of serviceability and reliability. In effect, these two measures are intended to assist conceptually in the estimation of the probability that the system will behave as desired, and then that the behavior is indeed adequate within the context of larger systems of which a system of constructed facilities is a part.

Satisfying the inequalities given in (1) is the traditional means by which systems of constructed facilities are designed. The basic difference between traditional approaches to the analysis of systems of constructed facilities and that suggested herein is that the traditional approach assumes that these inequalities can, with certainty, be satisfied, while here it is proposed that the outlook is less certain. Reliability reflects this uncertainty.

2. Computations of Reliability

Reliability has been defined as the probability that failure does not occur in any of the I possible failure modes. One is now faced with the difficulties of evaluating this function, of estimating the probability of success.

Two basic approaches to making these evaluations may be identified, which are pertinent to systems of constructed facilities: the first of these might be referred to as an analytical approach. In this case, one has definite mathematical statements relating demands to resistance, for each failure mode, as a function of appropriate $[c_m]$ and $[e_n]$. A major advantage of such an approach is that it may be feasible to develop functional relationships stating required $[c_m]$ for given $[e_n]$ and S_f, in a probabilistic framework.

Most standard design methods have been formulated through such relationships, traditionally used in a deterministic fashion.

The second approach might be termed an activities approach. One will try to describe the chain of events which occur, leading eventually to failure in a particular mode. This approach can often be taken when analytical models are not available. That is, it may not be necessary to know how something is happening, simply that it is occurring. An estimate of the probabilities of occurrance of these individual events than yield a probability of failure.

One may identify a third approach to reliability estimates, which is unlikely to be useful for the analysis of systems of constructed facilities. This is a straight statistical approach, using full scale models of the system (2). This technique has been useful in aerospace and electronics, especially in quality control of parts. Such testing may be of use for systems of constructed facilities as a means for estimating the behavior of components of the larger system.

The two principal approaches may be used in concert, either in series of parallel applications. For a series case, one may have a model of behavior which applied until a certain limit is reached. When this limit is reached, one describes subsequent events which could lead to failure, using the activities approach, because no model is available. The analytical model in effect was used to compute the probability of occurrence of the first step.

A parallel application might be useful when there are more than one identifiable physical phenomena leading to the occurrence of failure conditions χ^*_I. For example, loss of highway pavement rideability might occur due to the progressive deterioration caused by accumulated vehicle loads, or due to break-up of the pavement caused by the occurrence of cracks, followed by rains sufficiently heavy to cause loss of subgrade support. There are models available to predict the progression of the first phenomenon, but the second can be predicted only in terms of the chances that these events occur. A closer look at each of these two approaches will now be taken.

For the sake of discussion, an example will be assumed, such that there are only two failure modes. For the first failure mode, assume that there is a model available to permit analytical approach to be used. This model could be theoretical in its derivation or simply a statistical correlation which appears functionally meaningful to the analyst. The criteria for use of a model are that it should predict behavior as a function of system and environmental parameters, and that it should predict as accurately as possible.

A model is selected which, hopefully, meets these criteria. Consider, for example, that the system is a highway pavement, and that the first mode is loss of rideability. There are than models available to predict the number of equivalent 18 kip loads required to cause a given loss of rideability for a pavement of given properties. The loss of rideability which is considered to be failure is determined from the minimum serviceability level set by higher level considerations. The pavement properties will be represented in terms of their probabilities of occurrence, which depend on such factors as construction control, design decisions, temperature, etc. This distribution of $P(c_1,c_2,...)$ may be put into the model chosen, to generate the probability of capabilities $P(R_1)$. This generation may be done analytically (in the mathematical sense) or numerically (through simulation).

Similarly, the probable demands are estimated. For example, if exponential growth from an initial daily traffic is assumed, total loads will depend upon the distributions of initial traffic and growth rate. Hence, $P(D_1)$ is estimated.

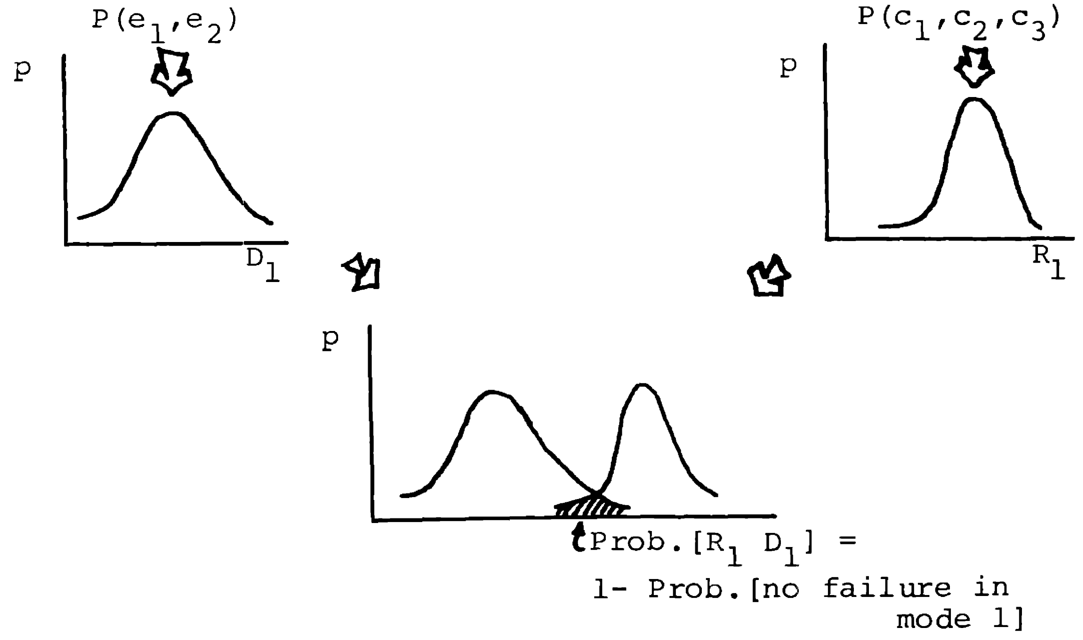


FIGURE 4: Analytical Approach to Reliability

Figure 4 illustrates the procedure. Idealized distributions for (e_1,e_2) and (c_1,c_2,c_3) are programmed into the computer. For example, it might be assumed that a normal distribution holds for each individual variable. The computer then takes "sample" for each distribution, using a random number generation technique. With an adequate number of samples, the assumed distributions of (e_1,e_2) and (c_1,c_2,c_3) will be reproduced.

Each time a sample is drawn from each distribution, values of D_1 and R_1 may be computed. In principle, given adequate numbers of samples, the expected distributions of demands and resistance can be described. With these descriptions, one is in a position to compute $P[D_1 > R_1]$, the probability that failure will occur in this first mode.

For the second possible mode of failure, assume that there is no good analytical model. An activities approach must be used. One must define a string of discrete events which will occur, culminating in the occurrence of failure. Such a string of events may be termed a possible lifetime of the system (3), and there will generally be several possible lifetimes associated with any single failure mode.

The concept of a lifetime may be formulated mathematically. Let $j\ell i$ be the j'th possible lifetime leading to failure in the i'th mode. $j\ell i$ may be described as a composite event comprising the union of an ordered set of elemental events $j^{\{f}t^{\}}i$. These elemental events are defined to be independent and described by unconditional probabilities of occurrence pt. For example, the occurrence of breakdown of the heating system and the onset of cold weather could be called independent for an analysis of housing and would then be elemental events whose union would comprise a failure through loss of comfort, a system lifetime. On the other hand, because cracking in concrete is dependent upon moisture conditions, structural failure defined by fracture might have only one elemental event, consisting of a combination of humidity and micro-cracking.

The failure lifetime is defined as a union of elemental events

$$_j\ell_i = [f_1 U f_2 U \ldots U f_t],\ f_1 \varepsilon\, _j\{f_t\}_i$$

where the set $_j\{f_t\}_i$ is the set of those events which are included in this particular lifetime. This statement may be represented pictorially as a tree diagram (see Figure 5).

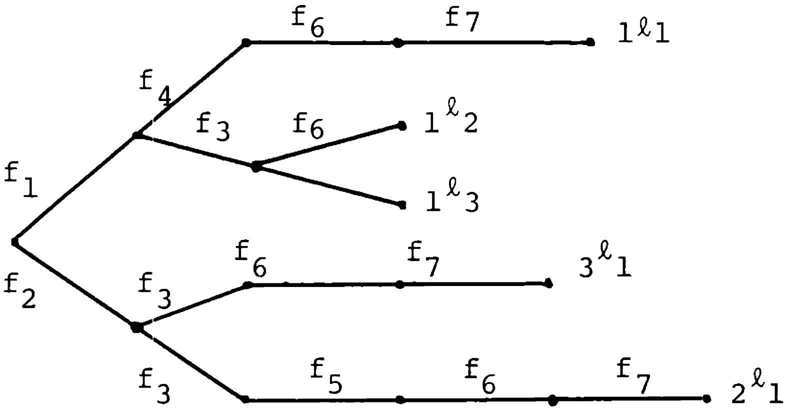


FIGURE 5: Possible Lifetimes of a System

Each branch in the tree, each path from the initial point to the terminal point, is a lifetime. One should notice that some elemental events are included in more than one lifetime. For example, heavy rains could play a role in the occurrence of subgrade subsidence or in frost heaving under a highway pavement, where either subsidence or heaving could cause failure through loss of rideability. Hence it is necessary to identify the set of events $j\{f_t\}_i$ in the above definition of $j^{\ell}i$.

The probability that a failure mode will occur may be expressed as the probability of occurrence of any of the several lifetimes which cause that failure. Then for the particular failure mode i,

$$P(\text{failure in mode } i) = P(\text{any } j^{\ell}i) \quad j = 1,\ldots,J,$$

with J possible lifetimes associated with the mode.

That is, to adopt previous form,

$$P(R_i < D_i) = P(\text{any } j^{\ell}i) \qquad (3)$$

This statement gives a slightly different view than that of the analytical approach. In the analytical approach, it is possible to say that there is a distribution of combinations of D_i and R_i, as suggested in Figure 4. With an activities approach, one feels that it cannot be said precisely what values the functions D_i and R_i to assume, but that if one of the group of possible lifetimes occurs, these values are such that $D_i > R_i$. One cannot say if the failure will occur because of relatively high D_i or low R_i.

The elemental events f_t were defined to be independent are predicted by probabilities p_t. So, one may say that

$$P(j^{\ell}i) = p_1 p_2 \ldots p\tau$$

where the probabilities p_t are associated with the τ events $_j\{f_t\}_i$ in that lifetime.

For example, returning to the illustration, the second failure mode might be loss of structural integrity of the highway pavement, where this refers to the pavement's ability to serve a heavy vehicle. That is, if the conditions of the pavement are such that the occurrence of a heavy load will result in a definite failure, then structural integrity is lost even though the loading does not in fact occur, because the pavement was intended to serve this load.

One possible lifetime might be described as the use of fine grained material in the base course, followed by the occurrence of major cracking in the surface layer, followed by the occurrence of heavy rains, followed by washing of the base material, ending with a loss of subgrade support. As stated, there are five elemental events, although the last two -- washing and loss of subgrade support -- might be considered to be synonymous. Assuming this is the only possible lifetime, $P(D_2 > R_2) = p_1p_2p_3p_4p_5$ where these p_t refer to the events named.

It is sometimes relatively easy to define events f_t and much more difficult to estimate the associated independent probabilities p_t. In this case one might prefer to define a <u>partial lifetime</u> as the occurrence of a subset of events which form a part of one or more lifetimes. In Figure 5, the composite event (f_1,f_3) is a partial lifetime, which might be designated λ^a. The probability of occurrence of the partial lifetime λ^a in this case would be given as

$$P(\lambda^a) = p_1p_3$$

The advantage of defining a partial lifetime is that one may be able to estimate its probability of occurrence directly, much more easily than one could find the needed p_t's. Computationally, any particular λ is really just another elemental event. Philosophically, it is recognized to contain a number of "more elemental" events.

If a partial lifetime λ^a has been defined as the first n events in the lifetime $j^{\ell}i$, then the remainder of that lifetime is another partial lifetime λ^b, containing the remaining $(\tau-n)$ events of the τ in $_j\{f_t\}_i$. Then $j^{\ell}i = \lambda^a + \lambda^b$. One may compute,

$$\begin{aligned} P_j(j^{\ell}i) &= P(\lambda^a)P(\lambda^b) \\ & P(\lambda^a)P(j^{\ell}i - \lambda^a) \end{aligned}$$

Obviously, elemental events can be combined to give a variety of possible partial lifetimes, so that this equation is a rather general statement. Further, a single partial lifetime may be contained in more than one lifetime. Equation (3) may then be rewritten as

$$P(R_i < D_i) = \sum_a P(\lambda^a)P(j^{\ell}i - \lambda^a) \tag{4}$$

This form is most convenient when there are several possible lifetimes for which a large number of the initial events are identical. In Figure 5, the sequence (f_2, f_3) is found in two lifetimes and might thus be conveniently handled as one partial lifetime.

In practice, it will often be found that the system may be characterized by a description of its current state, which is taken to indicate that a particular partial lifetime has occurred. The future behavior may be predicted from this state condition without reference to the events within the partial lifetime. For instance, the observation of some amount of cracking in a highway pavement may be viewed as an implicit definition of a partial lifetime. If maintenance is ordered, failure will not occur. Then the probability of failure may be viewed as dependent only upon the observation of cracking, and not on how it came about.

With this approach, it will be argued later that in many cases for systems of constructed facilities, equation (4) may be conveniently replaced by a Markov process. It will be suggested that a Markov process will in many cases prove to be a reasonable first approximation of reality, and because of its computational characteristics, a desireable approximation also.

As a result of applying an analytical approach or an activities approach, or some combination of the two, the probabilities of occurrence of failure in each of the I possible failure modes is estimated. It may be the case that because of physical interdependencies among the processes causing failure, it is necessary to estimate joint probabilities of occurrence of more than one mode at a time. Thus, it is impossible to expand in any general way upon the expression defining reliability as

$$R = \text{Prob.}\ (R_i \geq D_i) \quad i = 1,2,\ldots,I$$

If the I failure modes are stochastically independent (4), then

$$R = \sum_{i=1}^{I} P(R_i \geq D_i)$$

Stochastic independence is a requirement distinct from the independence of judgement variables $[z_i]$ which led to the original statement of inequalities in (1).

In the two mode example discussed throughout this section, the occurrence of the second mode is probably dependent upon the non-occurrence of the first. The level of detail given in the example was insufficient to determine stochastic independence or lack thereof. But in a later chapter, a substantially expanded version of this example will be used to illustrate the overall analysis of reliability, anticipating this analysis, it is suggested that in this case reliability would be given as

$$R = P(R_1 \geq D_1) + P(R_2 \geq D_2) \; [1 - P(R_1 \geq D_1)]$$

rather than as the simple sum $P(R_1 \geq D_1) + P(R_2 \geq D_2)$.

3. Lifetime Modeling

While it has been only indirectly pointed out in the discussion of the activities approach to reliability prediction, the physical behavior of a constructed facility is highly time dependent. This dependence is inherent both in the physical phenomena - such as aging and service wear - which comprise the observable aspects of service behavior and in the evalutation of behavior relative to a specified design service life.

To compute reliability, one must first construct a model of the service life behavior of the constructed facility. It is suggested here that a generally useful way of representing the system is in terms of a state space. A state is a description of the condition of the system in terms of appropriate characteristics (for example $[X_k]$), along with such historical data as may be needed to make predictions. A partial lifetime is a possible example of a state description. The elemental events included comprise the historical data which is important in predicting subsequent failure.

It was pointed out that practical experience with systems of constructed facilities indicate that one may often represent the system in terms of only the current values of appropriate $[x_k]$ without regard for how this condition came about. That is, the actual partial lifetime which resulted in a particular condition x, of the many possible lifetimes, will have no impact in such a case upon predictions of the system's future behavior.

Because of this situation, in many cases one may represent the service life behavior of a system of constructed facilities in terms of a Markov process. This special type of stochastic process has a number of convenient computational advantages.

A Markov process may be defined as follows (6): For any integer $n>1$, if $t_1<\ldots<t_n$ are parameter values, the conditional X_{t_n} probabilities relative to $X_{t_1}, X_{t_2}, \ldots, X_{t_{n-1}}$ are the same as those relative to $X_{t_{n-1}}$; that is, for each λ,

$$P[X_{t_n} < \lambda | X_{t_1}, X_{t_2}, \ldots, X_{t_{n-1}}] = P[X_{t_n} < \lambda | X_{t_{n-1}}]$$

A process which possesses this Markov property has no memory - i.e., its predicted future behavior depends only upon the current state of the process. Future development is independent of the way in which the current state was reached.

From the definition of the Markov process, one may derive a general statement, the Chapman-Kolmogorov Equation:

$$p_{st}(m,n) = \sum_k p_{sk}(m,r)\ p_{kt}(r,n) \quad m<r<n$$

That is, the probability of going from state s to state f in the time between m and n is the sum of possible chances of going from s to any other state k in the period m to r, and then from k to t in the remaining time. This equation may be compared with the previously derived statement;

$$P[R_i < D_i] = \sum_a \{P[\lambda^a] . P[{}_j\ell_i - \lambda^a]\} \tag{4}$$

State s in the Chapman-Kolmogorov equation will correspond to the initial point of a service lifetime, while state t is the terminal occurrence of failure in mode i. Each intermediate state k represents a set of conditions occuring at the end of any particular partial lifetime λ^a. When the partial lifetimes can be represented by x only, one need not distinguish among the many possible ways that x occurred, and a single state replaces many partial lifetimes. The Markov process thus becomes a very compact way of modeling system service behavior.

It is possible to define a Markov process in terms of discrete or continuous time and discrete or continuous states, and the reader is referred to the cited references for discussions of these variations. The full range of desireable computational features are realized in the discrete state, discrete time process (see Drake (7)), and it is suggested here that such a process will often provide a useful first approximation of many aspects of systems of constructed facilities.

This model is particularly desireable as a means of investigating maintenance policies. Often, maintenance actions are undertaken contingent upon the observation of a particular set of conditions. Further, inspections and other normal maintenance activities are often periodic, carried out at regular time intervals.

It of course cannot be shown that the service behavior of a system of constructed facilities may be generally represented as a Markov process. In fact, if such a process is useful at all, it will often be in conjunction with other stochastic models. Such use will be illustrated in later pages. It is here simply suggested that the Markov process will often be applicable, and that where it can be applied, it will prove useful. This application, for particular types of constructed facilities, is an area in which further work should prove useful.

D. Uncertainties of Future Actions -- the Uses of Maintain ability

1. Definitions and Rationale

Throughout the service life of a system of constructed facilities, there are actions undertaken upon which the behavior of the facility will depend. A major class of these actions, termed maintenance, are to some degree planned for, to slow or prevent deterioration of the facility's service, or to repair deterioration which has occurred (8). Reliability as a component of performance permits an estimate to be made of the levels of uncertainty in any proposed alternative facility. But this estimate will depend upon the proper execution of the maintenance actions, and is thus subject to the uncertainties of human influence throughout the service life.

It is then proposed that design decision should be made with considerations of the sensitivity of a plan to this particular brand of uncertainty. To what extent is effort required during the design service life to assure that adequate service will be provided? A measure of this sensitivity will be suggested, and will be termed maintainability.

Maintenance actions may be classified roughly into two categories: normal maintenance is the regularly scheduled day-to-day activity required to keep reliability at high levels. Normal maintenance is preventive in nature. It is intended to assure that partial lifetimes are not completed, or to tighten the distributions of system characteristics. In many cases, the failure of a system will hinge upon poor execution of normal maintenance actions, and this poor execution will be represented as elemental events in a service lifetime.

Repair maintenance is required when failure has actually occurred or is felt to be undesireably close (i.e., loss of reliability), prior to the end of the design service life. Repair maintenance actions are intended to restore the system to an adequate level of service. These actions might be viewed as associated with the events which occur with probability (1 - R).

Maintainability is described as a measure of the degree to which continued effort is required during a facility's design service life. High maintainability is achieved through minimization of the number of maintenance actions needed and the time required to complete these actions, relative to the design service life of the facility. It may immediately be seen that maintainability will be closely linked with reliability through the scheduling of maintenance activities.

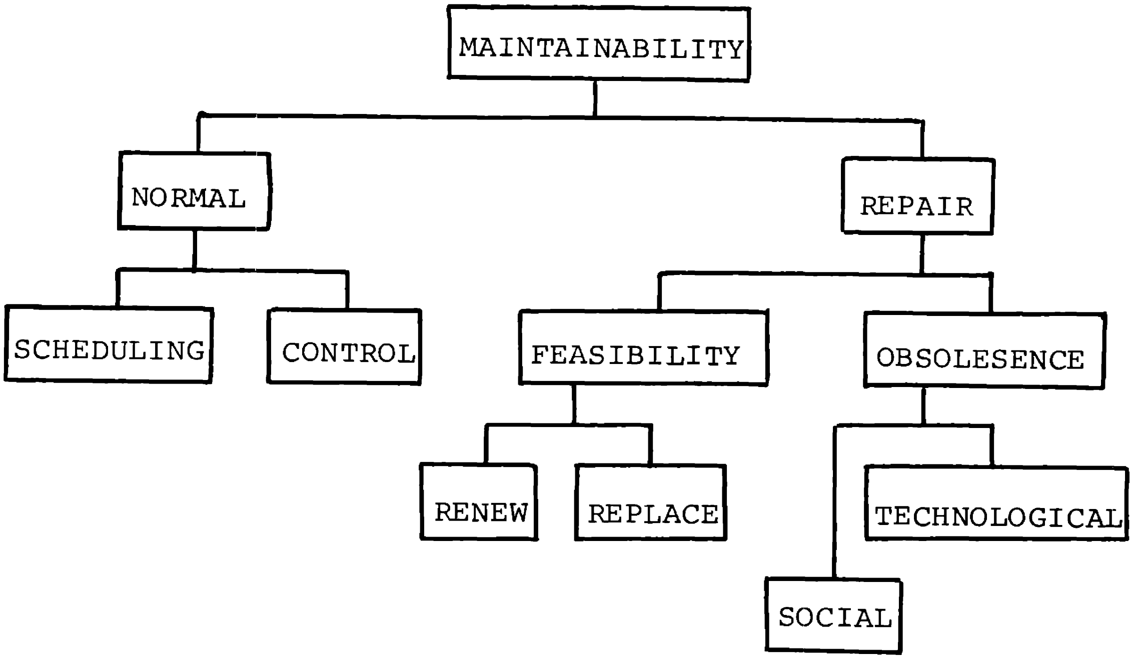


FIGURE 6: Components of Maintainability

In practice, maintainability will depend on a range of factors from the degree to which it is feasible to rapair a particular type of failure to the efficiency of the maintenance organization in detecting and acting upon failures and normal maintenance needs (Figure 6). Scheduling and control of materials and parts inventories and activities may become critical. The decision whether a part of the constructed facility is to be repaired (for example, patching or plaster walls) or replaced (versus use of plaster board) enters into planning and design. Finally, questions of obsolesence of the facility will sometimes complicate evaluation of the benefits of maintenance.

This then is a qualitative view of what maintainability will mean in decision. A more quantitative view is now of interest, from which functional measures of maintainability may be derived.

2. Measures of Maintainability

A system of constructed facilities is intended to provide service throughout a particular design service life. Overall serviceability will generally start at some high level, and will deteriorate with age and service usage of the facility until it reaches the failure level, defining a failure age. Generally speaking, this failure age must, at a given level of reliability, equal or exceed the design service life if the facility is to be considered satisfactory. Figure 7 illustrated the trend.

To the extent that normal maintenance is effective, its neglect would be expected to lead to earlier failure of the system. The manner of occurrence of the failure is suggested by the second trend line in Figure 7 to be through more rapid deterioration. A certain amount of the service life would thus be lost. This amount will be designated T_n for later discussion.

An unexpected failure and its repair may be viewed in a similar fashion. Some time T_r will be required to return a facility to satisfactory service given that the failure occurs. This lost time, analogous to T_n, will be a characteristic of the mode of system failure and the feasible repair actions. It may be pointed out here that the prob-

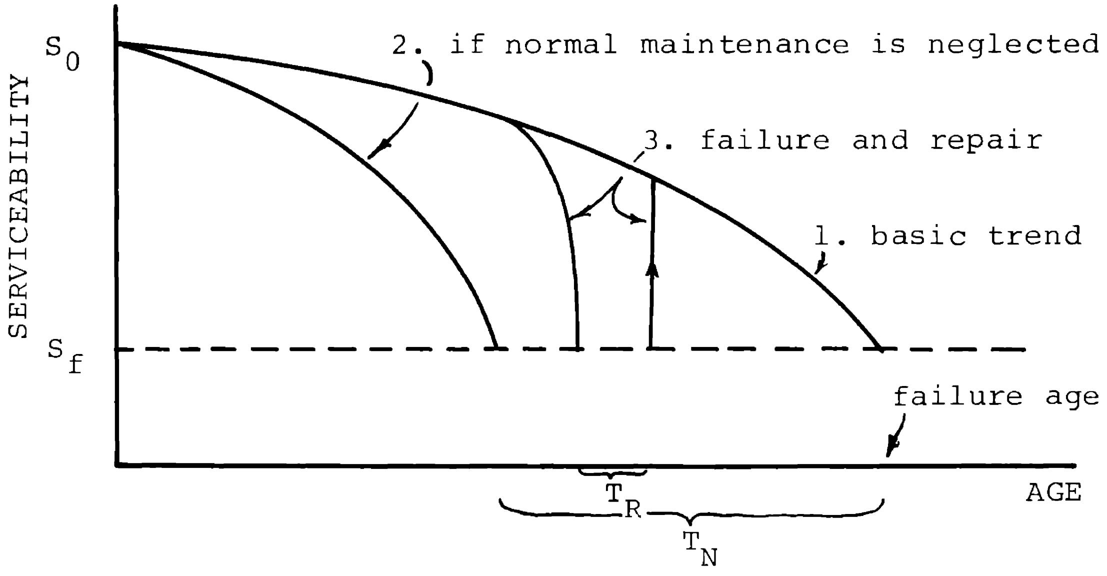


FIGURE 7: Consequences of maintenance

ability of occurrence of any such trend as number 3 in Figure 7 is estimated to be (1 - R).

The expected values of the parameters T_n and T_r will be characteristics of the particular system of constructed facilities. Larger values of T_n and T_r will indicate a greater dependence of the system upon maintenance activity. A ratio of either of these parameters to the total design life will indicate the service availability of the system, with respect to losses associated with maintenance (9). This indication of availability provides a useful basis for estimating maintainability.

In particular, a coefficient of maintainability with respect to repair maintenance may be stated as

$$M_r = \frac{T_D}{T_r}$$

where T_D is the design life and T_r is expected repair time lost, as described above. A high value of this ratio indicates a low sensitivity of the system to repair maintenance activities, i.e., high maintainability. It is the inverse of the expected fraction of service life lost if repair is needed, and may intuitively be viewed as an estimate of the number of times failure could occur before the service life is exhausted. Ease of repair and reduction of the need for same will increase this coefficient of repair maintainability.

In similar fashion, a coefficient of maintainability with respect to normal maintenance is proposed. In this case, T_n/T_D is the fraction of the service life which might be lost if normal maintenance is neglected.

This normal maintenance fraction is associated with the events that normal maintenance is neglected and failure occurs. It is possible in this case that failure would have occurred anyway, that it was not the neglect of normal maintenance which is at fault. To obtain a true estimate of maintainability with respect to normal maintenance, an adjustment must be made for this possibility.

If the following probabilities are defined:

$1-R$ = probability of failure

$P[NM]$ = probability that normal maintenance will be carried out

$P[F|NM]$ = the probability that failure will occur, given that normal maintenance is carried out,

$P[NM|F]$ = the probability that normal maintenance was carried out given that failure occurred;

Then from the definition of conditional probabilities one may state that

$$\frac{1-R}{P[NM]} = \frac{P[F|NM]}{P[NM|F]}$$

To obtain an estimate of the chance that a failure, and thus any time which may be lost, is due to something other than normal maintenance, assume that $P[NM|F] \to 1.0$; i.e., it is certain that normal maintenance was adequate even though failure occurred. Then one finds

$$P'[F|NM] = \frac{1-R}{P[NM]}$$

$P'[F|NM]$ is an estimate of $P[F|NM]$ on the basis of the assumed adequacy of normal maintenance.

The fraction of service life lost in the event of failure, which would have been avoided if normal failure were carried out, is now estimated as $(T_n/T_D)(1 - \frac{1-R}{P[NM]})$*. A coefficient of normal maintainability may now be given as

$$M_n = \frac{T_D}{T_n} \left(\frac{P[NM]}{P[NM] - (1-R)}\right)$$

As with M_r, this parameter increases with ease of maintenance and decreases with sensitivity of the facility to maintenance caused service losses.

It may be seen that the meaningful range of values for both of these coefficients is $1<M<\infty$. It makes little sense to consider M 1, for while this is theoretically possible, it indicates a loss of time in excess of the design service life. A value of M=1 indicates that normal maintenance is absolutely necessary or that a failure cannot be repaired.

* The fraction $\frac{1-R}{P[NM]}$ is constrained by "rationality" to be less than unity. If the probability of failure is large and P[NM] is relatively small, the assumption of $P[NM|F] \rightarrow 1.0$ becomes meaningless.

At the other extreme, a facility which is planned to have no normal maintenance will exhibit $M_n \to \infty$ as the expected time lost due to neglected maintenance, T_n, approaches zero. One would expect that M_r will always be a finite value, because repair time cannot be eliminated (unless there is a duplicate system ready for immediate use -- an unlikely situation for systems of constructed facilities).

E. Summary

It has been suggested that reliability and maintainability be used as measures of effectiveness in the analysis of systems of constructed facilities. These measures are intended to reflect the uncertainties and risks involved in a particular system. The measures will depend upon the inherent variabilities of system and environment, operations and maintenance management policies, accuracy and detail of prediction models, and other such factors.

It is proposed that the service life of the constructed facility is best modeled, for purposes of prediction, as a stochastic process. The occurrence of failures may be predicted through use of models representing the conditions at failure, or through analysis of the activities which lead up to failure. The practical aspects of the behavior of systems of constructed facilities will often make it worthwhile to consider the use of Markov processes to model certain aspects of behavior. Reliability and maintainability are then predicted using these models of service behavior.

Expected trends of serviceability may also be predicted.

The use of such a probabilistic approach to prediction makes greater demands upon the analyst than does the more traditional deterministic approach. The greater demands appear primarily in the form of a need for more extensive data on the characteristics of system and environment. In general, it requires several pieces of information for example a description of the probability distribution type, a mean value, and a standard deviation, to describe probabilistically a variable which is predicted deterministically by a single number.

These increased requirements for data handling can and undoubtedly will be accomodated through applications of the computer. It is suggested here that in fact the entire framework of models for predicting service life behavior for a particular type of systems of constructed facilities and a particular design decision problem, and ba computerized, and that doing so will be value by permitting the decision-maker to explode a substantially broader range of solutions in his problem solving.

CHAPTER V

THE ANALYSIS OF HIGHWAY PAVEMENTS

A. Analyzing the Pavement as Part of the Transportation System

In the highway pavement one has a fairly clear example of a constructed facility which is intended to serve as part of a larger system. There is certainly little reason for the existence of the long band of asphalt or concrete except for the role it plays in a transportation network.

Properly, the constructed facility in highway transportation will include not only the pavement as it is generally defined, including its substructure and foundations, but also bridges, interchanges, cut and embankment slopes, lighting, traffic signals, etc. Certain associated components, such as right-of-way and air rights structures, might be included on occasion, as they effect the transportation function of the highway, but these components are not strictly part of the constructed facility. In this example however, the discussion will be restricted to the pavement structure, and it is suggested that in many cases (in this particular example, inter-urban roads) this restriction will still yield an analysis useful for investigating the interactions of physical, social, political, and economic systems. In the present case then, bridges and similar structures will be neglected, for simplicity's sake, and the discussion will concentrate upon highway pavements.

Having roughly identified the system, one might next examine the role of the facility as part of a transportation system, and investigate the nature of the users. In this case, there is a substantial amount of literature devoted to the analysis of transportation as a factor in urban and regional growth and development (1,2). At another level of discussion, the highway as an influencial agent in social patterns and political institutions is becoming apparent (3).

The transportation planner, who makes the decision that a particular transportation link will be a highway, will generally consider factors of comfort, convenience, safety, speed, and cost as the measure of transport systems effectiveness (4). Satisfactory service by the highway pavement is considered to be rendered when the pavement does not interfere with the desired levels of these five measures. That is, the physical behavior must be such that the users - direct, indirect, and subsidiary - do not find the overall transportation system impared by the highway pavement. With this criterion in mind, one may attempt to structure the analysis of the system of constructed facilities.

More specifically, the planning decision begins with an equilibration of supply and demand, assuming relatively standard features of the highway link - i.e., features which will not produce a significant alteration of demand. It then falls to design to determine the possible technologies which will meet these equilibrium requirements, and thus to permit the planner to evaluate alternative courses of action, each alternative including a particular physical system as a component of the whole. This chapter will examine the analysis of highway pavements in this light.

B. Pavement Serviceability

1. The Components

The first step in the analysis is to identify the component subscales of serviceability. Figure 1 suggests that serviceability for highway pavements may be represented by three components: rideability, safety, and structural integrity. These components were suggested through application of the previously discussed techniques for analysis of problem structure, through hierarchical decomposition. (See also Appendix A).

Rideability is a descriptive term for the quality of ride which the pavement system gives. This factor is readily apparent to the direct user, whose comfort is effected by his response to the physical stimuli of the pavement. It might also be expected that comfort will be correlated with liklihood of damage of goods shipped over the road, and would thus be of some interest to the indirect user. Travel speed is of course of importance to both of these groups, as is cost. Vehicle operating costs are a principal factor in this consideration. Rideability as a whole will effect the demand for the facility, or rather the willingness to use the facility, and is thus of importance to the subsidiary user.

Actually, rideability is governed by a complex interaction of vehicle and pavement characteristics (Figure 2).

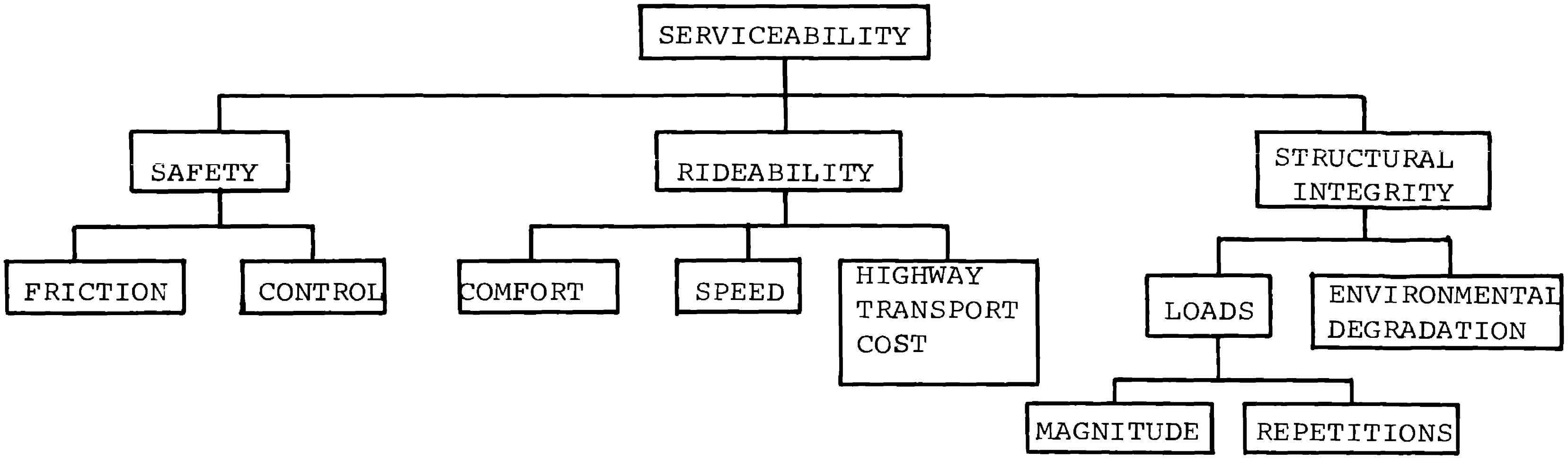


Figure V -1 : Highway pavement serviceability

The perceived quality of ride will be a variable depending on vehicle characteristics, surface qualities, etc. That is, evaluation of serviceability with respect to rideability will be very much a function of current technology and the transportation role of the pavement. The same pavement will be evaluated differently if it is to serve as an inter-urban expressway or as a rural secondary road.

Safety is a similarly complex component of serviceability. Its evaluation is especially difficult because of the ethical and moral questions involved in placing values on human injury and death.

The highway pavement will effect safety in two ways. Friction characteristics of the system will influence the various forms of sliding and skidding which may lead to losses. This area of safety has received substantial attention over a number of years. Another source of unsafe behavior, one which has not gained such attention, is the hazards which may cause loss of control of the vehicle. Large irregularities, lack of lane markings, narrow pavement, are examples of such control factors.

Structural integrity is a concern directed toward future behavior. Can the pavement support the loads for which it is planned, throughout the remainder of the design life, without losses of rideability or safety? An example of how structural integrity might be lost without immediate loss of rideability or safety might be found in the phenomenon of pumping. A substantial void under a rigid pavement would not be noticed by vehicles as long as the pavement slab does not crack and exhibit displacement. Such cracking might not occur under light auto loads, but would be virtually assured under a single heavy truck. This is a loss of structural integrity.

Structural integrity will depend upon vehicle loads -- their magnitude and number of repetitions -- and upon the environment -- moisture, temperature, chemicals, etc. Environmental degradation, for example, asphalt embrittlement, concrete-sulfate reactions, loss of joint filler due to high temperatures, can cause failures, even without substantial loading action.

The serviceability of highway pavements can be characterized in terms of rideability, safety, and structural integrity. The question now is how to measure these parameters.

2. Serviceability with Respect to Rideability

Rideability, particularly with respect to comfort, is a fairly subjective parameter. There have been attempts to circumvent subjectivity by viewing human response mechanistically, for example by constructing mass-spring-dashpot models (5). On the other end of the scale, purely subjective rating techniques have been used (CGRA (6) and AASHO (7)). Quite good reviews of these various approaches have been published; the interested reader is referred to these (for

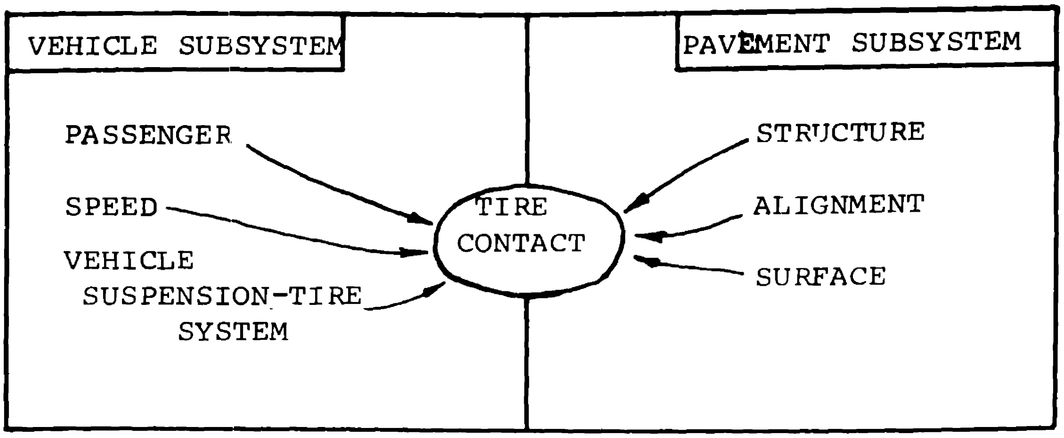
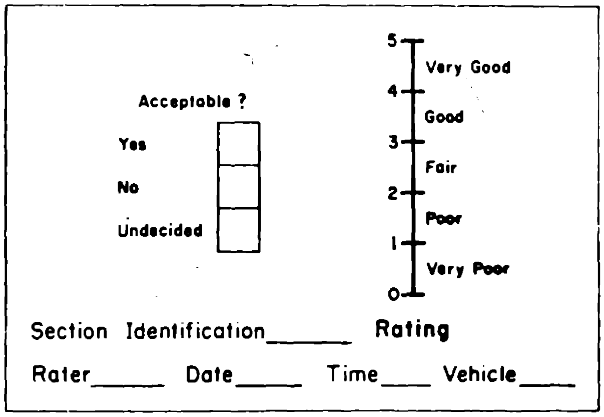


Figure V -2 : Vehicle-pavement interactions



Acceptable ?

Yes □
No □
Undecided □

5
Very Good
4
Good
3
Fair
2
Poor
1
Very Poor
0

Section Identification______ Rating

Rater_____ Date_____ Time____ Vehicle____

Figure 1-F. Individual present serviceability rating form.

Figure V -3 : Individual rating form for rideability (from AASHO (7))

example, Hutchinson (8), or Holbrook (9)).

For this example, data from the AASHO Road Test is used. The manner in which ratings were obtained illustrates quite well the techniques suggested for estimating serviceability by psychometric scaling. Further, it is felt that the data obtained is among the best available and provides a good example of the steps of identifying indicants and operationally useful scales.

There is one point requiring attention before proceeding. In the AASHO Road Test, the term serviceability was used to refer to pavement surface qualities which are here viewed as indicants of serviceability with respect to rideability only. That is, the AASHO concept of serviceability does not display the full breadth of meaning argued in the present work. Hence, in the following discussion, the term rideability has been inserted where AASHO might have used serviceability.

In the AASHO Road Test, a panel including "highway designers, highway maintenance men, highway administrators, men with materials interests, trucking interests, automobile manufacturing interests, and others" (7) were asked to rate 138 sections of highway pavement. Ratings were made on a scale of 0 to 5 of the pavement's ability to serve traffic at the time of rating. Figure 3 shows the rating form used. Panel members were told to use whatever criteria they wished, and were asked to judge whether the section was satisfactory or not.

(See Figure 3 on Page 157)

In effect then, each individual was being asked to rate his utility value for the pavement and to indicate the relative location of his critical region. The panel might have been of more general value had it included a representative selection of highway users rather than so-called experts. This biasing of the scale could lead to a lower individual rating and higher aspiration level relative to the average user of a given pavement section. The halo effect (see Chapter III) and the influence of common, specialized background among the panel members would be apparent in their possible tendencies to judge the pavement more critically. Subsequent studies reinforce this idea (10).

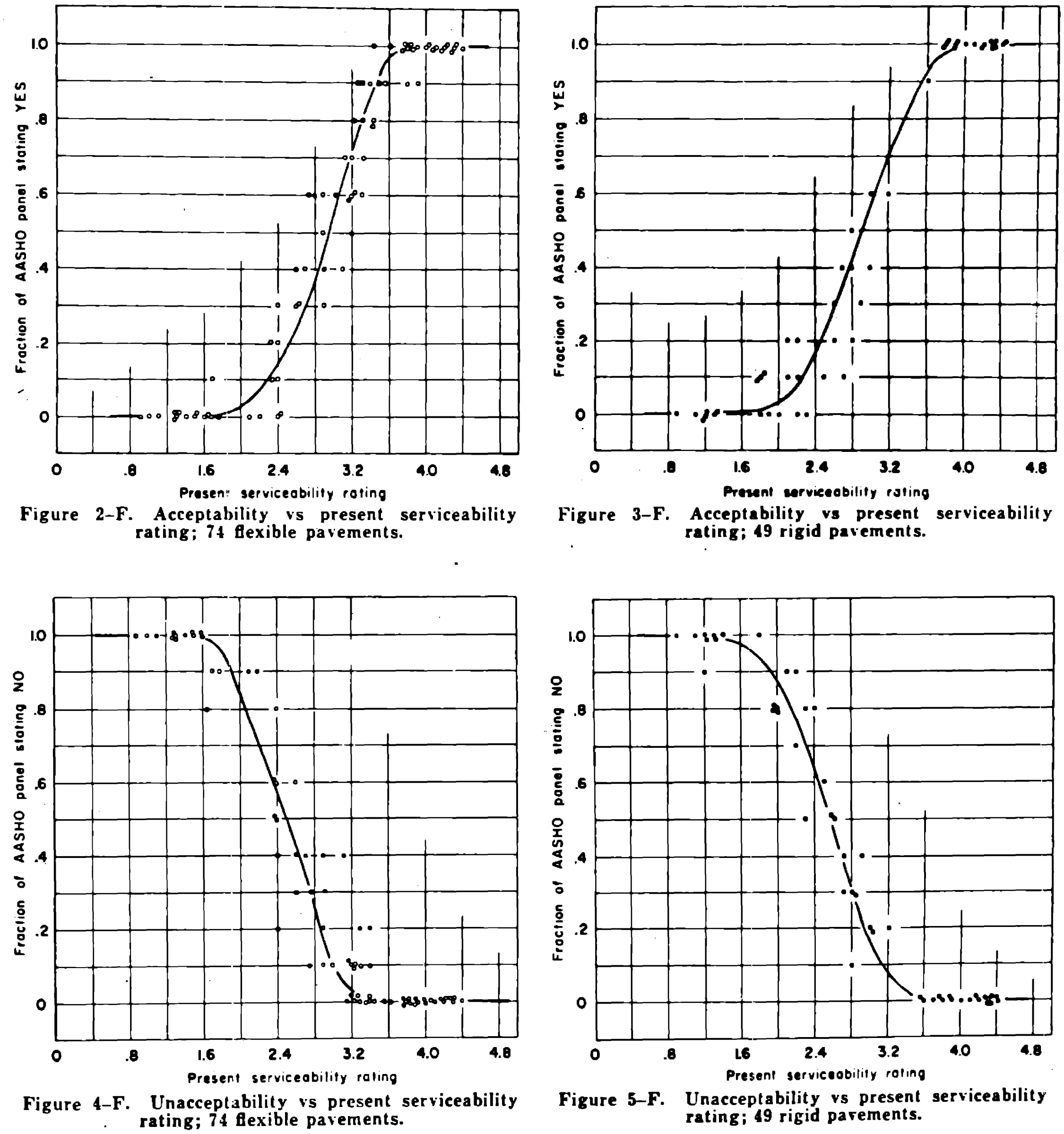


Figure 2-F. Acceptability vs present serviceability rating; 74 flexible pavements.

Figure 3-F. Acceptability vs present serviceability rating; 49 rigid pavements.

Figure 4-F. Unacceptability vs present serviceability rating; 74 flexible pavements.

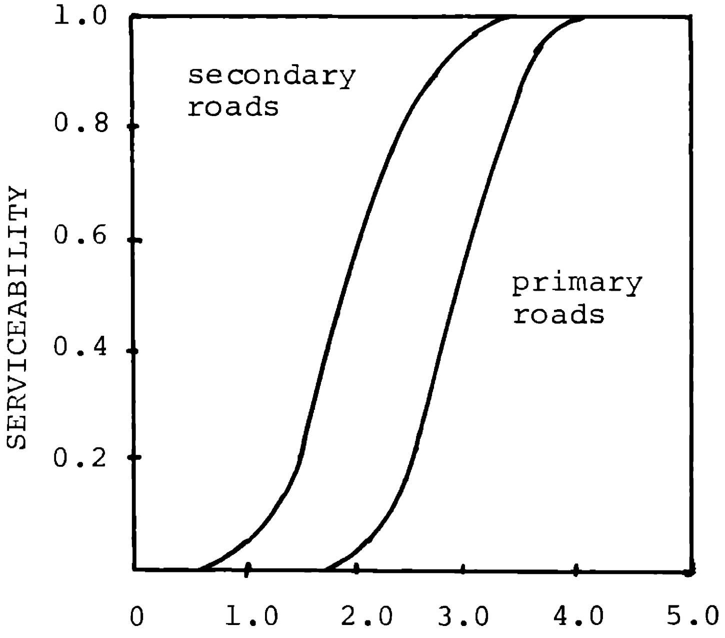
Figure 5-F. Unacceptability vs present serviceability rating; 49 rigid pavements.

Figure V -4 : AASHO rideability ratings (from 7)

From the individual ratings and the statements of acceptability, it was possible to derive directly a measure of rideability as discussed here. Figure 4 shows the AASHO plots of the fraction of the panel adjudging the pavement satisfactory versus the ratings given. The rideability scale values are the means of the individual ratings.

These ratings were derived for pavements serving high speed mixed truck and passenger traffic. It would be expected, from previous discussion, that for secondary roads the individual's aspiration level would be lower than for these high speed primary roads. As a result of this, assuming that the same general factors are influential in the choice process, the serviceability value should be higher on a secondary road than on a primary road with the same rideability rating. Data from Nakamura (10) and Purdue (11) would seem to support this expectation, although some additional work would be required to test the hypothesis thoroughly. Figure 5 shows serviceability versus rideability rating for primary and secondary roads.

(See Figure 5 on Page 162)



COEFFICIENT OF RIDEABILITY

FIGURE 5: Serviceability with Respect to Rideability

The rideability rating derived here is still a subjective measure of performance. The next step required is to find suitable indicants of this measure such that complete panel ratings are not required whenever a pavement section is to be evaluated. Numerous instruments have been derived to measure various aspects of the pavement's physical characteristics, and there is little agreement among highway personnel as to what is correct.

The AASHO analysis identified longitudinal roughness, cracking, patching, and in flexible pavements, lateral rutting, as the principal physical characteristics of the pavement which determines rideability (6). Lateral and longitudinal roughness are found to be by far the most important factors (11). These indicants were statistically fitted to the subjective evaluations to yield regression equations of the general form

$$r = A_o + A_1F + A_2\sqrt{C+P} \quad \text{Rigid Pavement}$$

$$r = A_o + A_1F + A_2\sqrt{C+P} + A_3(\overline{RD})^2 \quad \text{Flexible Pavement}$$

in which r = rideability rating;
F = roughness measure;
C = cracking;
P = patching;
$\overline{RD}$ = rut depth;
A_o, A_1, A_2, A_3 = constants.

The physical factors F, C, P, and $\overline{RD}$ are at present defined in terms of the techniques used to measure them. For example, F, roughness, may be measured using a roughometer, in which case the value in inches per mile is inserted directly into the equation. Other instruments measure similarly, but may require some transformation before insertion into the equation, as in the case of slope variance measured with a profilometer. Similarly, rut depth, ($\overline{RD}$), is defined as a particular number of measurements taken transversely to the pavement centerline along a straight-edge resting on the pavement. For a complete discussion of these measurement techniques and the instrument used, including suitable values of the coefficients, one may refer to such authors as Holbrook (9), Phillips and Swift (12), or Yoder and Milhous (11).

It is suggested that the function r -- it will be referred to as a coefficient of rideability -- is a suitable indicant of serviceability with respect to rideability. The measure is of course restricted to the area of high type pavements, for which it was formulated. The current technology is also implicit in this parameter. If, for example, new vehicles are developed which travel at much higher speeds, but are more sensitive to pavement surface characteristics for comfort and operating costs, then it would be expected that the ratings shown in Figure 5 would be shifted to the right. But for the purposes of this example, the measure is adequate.

3. Serviceability with Respect to Safety

The evaluation of serviceability with respect to safety is an especially difficult problem. The entire area of highway safety has been marked with lack of understanding and responsibility, and with general confusion due to the ethical questions involved. Traditional efforts to reduce safety considerations to monetary measures run into trouble when attempts are made to put values on human life and limb. The values are seldom justifiable except as attempts to obtain results in the face of extreme difficulty.

An additional problem is that safety is not generally directly apparent to the direct user of the highway. In most situations, the user is unaware of the chances that he will be involved in an accident. Only when a highway section is notoriously unsafe does safety become a conscious judgmental attribute.

It is felt that this situation is primarily one of education. If the user knew that one road was likely to be safer than another, he would prefer the safe one (other things held equal). One might then argue that it becomes obligatory that the decision-maker either educate the user or assume a responsibility for maintaining safety in such a way that the user would find satisfactory.

Following this line of thought, one might suggest that a suitable judgemental parameter for safety would be the relative probability of occurance of accidents. That is, if the user knew what the average and extreme values of accident statistics are, and he were given the probability of accident occurance on a particular road, he could make some judgement of that road. One could then experimentally attempt to find the function of serviceability versus the probability of accidents. The experiment might be conducted along the lines of the previously described investigation of rideability, perhaps with the assistance of visual aids such as the new computer simulation of highway driving conditions.

Unfortunately, no such work has been done. In order to illustrate the results which might be expected, some assumptions will have to be made and the function developed as an exercise in logic. First, the role of obstructions or other unusual features which might spur accidents will be neglected, as there is little data available in this area. Second, the step to an operational level will be made for the investigation of skidding behavior. That is, the function will be constructed immediately in terms of a physical parameter, analogous to the coefficient of rideability as a predictor of serviceability with respect to rideability.

Skidding occurs through the complex interaction of a set of parameters including pavement surface, vehicle speed, tire characteristics, and the depth of water on the pavement. At an extreme, skidding occurs due to a hydroplaning effect in which the vehicle actually rides up on a thin film of water.

The critical parameter is the breaking force coefficient or effective skid resistance (13). For a dry pavement at slow speeds, this parameter is the coefficient of friction as normally defined. But under other conditions, it reflects the net effect of all factors on the system's skidding characteristics. The coefficient will generally fall between 0.0 and 0.7 (14,15).

Accident studies indicate that the probability of accidents increases quite sharply as the braking force coefficient falls below 0.30 (16). In California 0.25 is considered to be the minimum acceptable value of coefficient of friction before remedial action is required (17). Other studies have indicated that the average value for pavements in reasonably good condition is about 0.5 (18).

Now assumptions must be made to postulate the users' response to safety features. It is suggested that these values of the braking force coefficient are representative of the current state of highway technology, and that the individual user will on the average be satisfied with a level of risk commensurate with this current state. Figure 6 illustrates a curve of serviceability versus peak braking force coefficient, which might be derived in this manner.

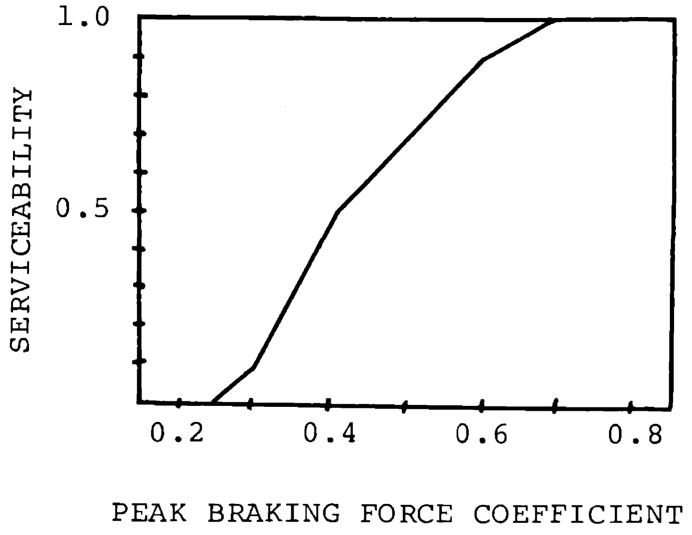


FIGURE 6: Serviceability with Respect to Safety

Referring to Figure 6, five distinct points are indicated on the curve. A value of the braking force coefficient F of 2.5 is considered complete failure by any standards. At 3.0 performance is still poor, but approaching a reasonable value. British standards rate this level slightly below satisfactory. 4.0 represents an acceptable value, about equal to that encountered on high quality, high speed pavements when they are wet (moderate rain, slide value). So the average user would be as likely to find it acceptable as to reject it. A value of 0.6 is high, about equal to the peak value expected on a high quality, high speed pavement. A value of 0.7 is considered safe by any standard.

While a great deal of work is still needed to verify the type of judgements made in deriving this scale, it is felt to be fairly representative of the conditions of safety with respect to pavement frictional characteristics. In using the braking force coefficient as an indicant, considerations of surface, drainage, and vehicle speed are incorporated. And, as verification that risk does vary as suggested, studies of pavement grooving have shown that improvement of the braking force coefficient from .25 to .35 can give a reduction in the number of accidents on the order of 75%, where the pavement at 0.25 was recognized as low in safety (18).

4. Serviceability with Respect to Structural Integrity

Structural integrity is not a directly perceptible quality of the highway pavement. Rather, it refers to qualities of the pavement which will effect the pavement's ability to provide adequate service in the future. Structural integrity depends upon the vehicle loads to be applied to the pavement - their magnitude and frequency of repetition - and upon environmental loads and degradation - water, temperature, and chemical effects - as well as upon the characteristics of the pavement system.

Often, loss of structural integrity will be associated with other modes of failure. For example, cracking in a rigid pavement effects rideability as well as structural integrity. But loss of structural integrity may occur alone. If pumping of base course material leads to sufficient loss of support, a breakup of the pavement will be likely if heavy loads are applied. However, behavior under lighter loads may be apparently satisfactory, even though structural integrity has in some degree been lost.

The pavement is designed to withstand a projected traffic load in providing its transportation service. This load will be stated in terms of vehicle weights and numbers. Using a concept of equivalency of loads (19), it is possible to convert this projection to a total number of applications of a single magnitude of load. The 18 kip axle load is a popular choice. It may be suggested then that the pavement which can resist the total number of predicted equivalent loads will be satisfactory. A pavement which has lost this capability - through poor design, excessive accumulation of damage, etc. is unsatisfactory.

The individual user will judge the system in terms of whether he can be served adequately. The load-carrying capability of a pavement may be extended by prohibiting vehicles whose load is greater than some particular quantity. This prohibition reduces the total of equivalent loads to be resisted. The individual who wishes to use a heavier vehicle will, however, consider the pavement a failure.

It is then suggested that serviceability with respect to structural integrity may be suitably expressed as a function of the maximum vehicle load which the pavement can serve without premature losses of safety and rideability. Determination of this function depends upon the particular traffic pattern for which the pavement has been planned. Figure 7 shows a curve derived for typical traffic composition.

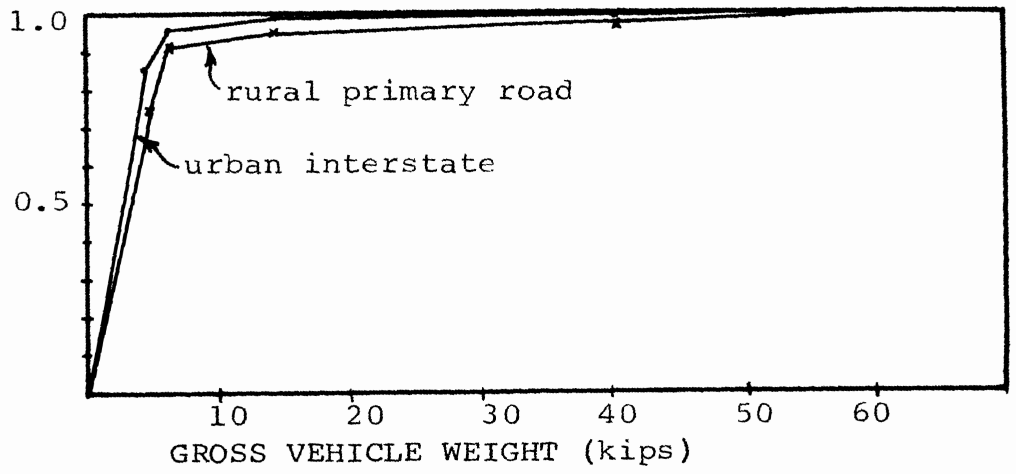


Figure V: -7 : Serviceability with respect to structural integrity

Using data from California (1966), traffic is represented by five principal weight classes (20). These curves are plotted in terms of gross vehicle weight, but could just as well be converted to axle loads, wheel loads, or any other pertinent measure. As illustration of variation in serviceability requirements as a result of the basic transportation function, two curves are shown, one for urban interstate and one for rural primary roads. Passenger autos comprise a larger percentage of traffic in the urban situation, and so a pavement which will resist all of the projected passenger vehicle loads has a better chance of being acceptable to the individual user.

C. Pavement Reliability

Figures 5,6, and 7 suggest measures of serviceability with respect to rideability, safety, and structural integrity. These measures have been stated in terms of physical system parameters which can be measured or predicted for any particular pavement system configuration. Thus it is possible to evaluate lifetime serviceability trends when this particular configuration is to be subjected to particular service demands.

The analysis of pavement reliability will begin with an identification of failure modes, which in turn begins with statement of minimum acceptable levels of serviceability. Such a statement requires reference to the overall transportation system. The highway as a transportation link will have been planned to serve a certain community of users and to meet certain economic criteria. The linkage between these higher level concerns and the constructed facility will determine failure criteria.

For example, if the highway is to serve half of the potential traffic between two cities, a serviceability of 0.5 might be required as minimum on all scales. If, in addition, the road is to serve large trucks and is thus an important freight channel, serviceability with respect to structural integrity might be set higher, say at 0.97. Failure has then been specified as the occurrence of one or more of three conditions: coefficient of rideability less than 2.9, braking force coefficient less than 0.41, or inability to support predicted future traffic loads up to the 40 kip range.

These criteria form the first stage of a description of the failure modes for the pavement system. These statements are then translated into physical system qualities. Figure 8 is a suggestion of the considerations that will be made. This figure shows possible causes of a number of forms of pavement distress, for rigid and flexible pavements (21). Each type of distress will, if severe enough, lead to losses of serviceability. One might view each of the types of distress as specifying one of the failure mode inequalities discussed in Chapter IV.

It will be noted that the distress types listed will

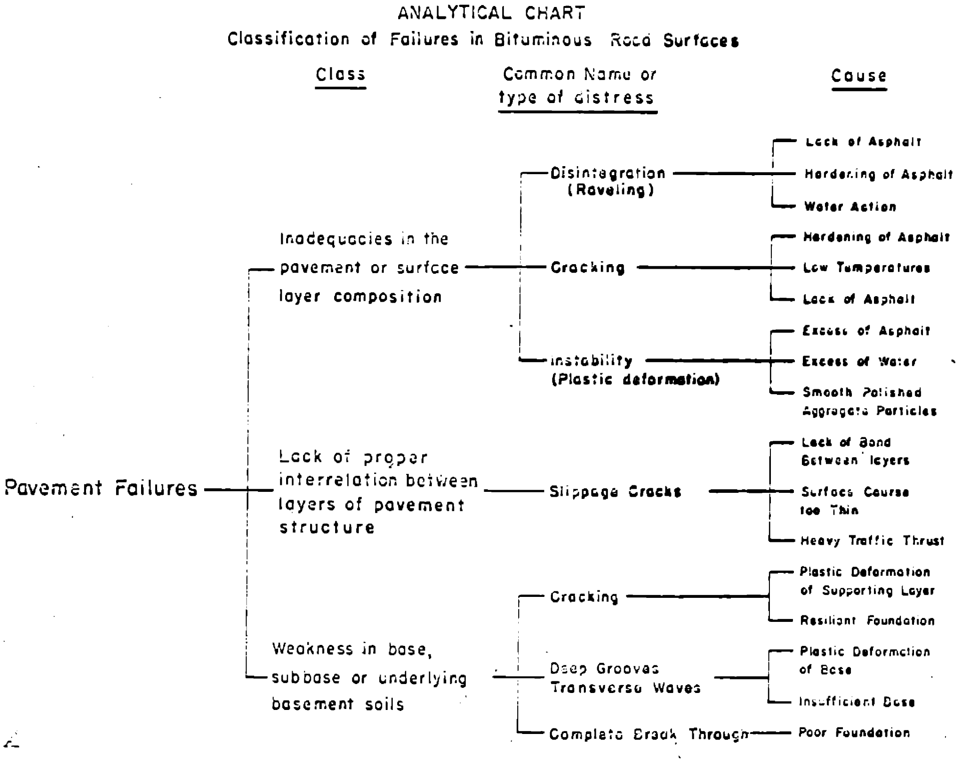

ANALYTICAL CHART
Classification of Failures in Bituminous Road Surfaces
Class
Common Name or type of distress
Cause
Pavement Failures
Inadequacies in the pavement or surface layer composition
Disintegration (Raveling)
Lack of Asphalt
Hardening of Asphalt
Water Action
Cracking
Hardening of Asphalt
Low Temperatures
Lack of Asphalt
Instability (Plastic deformation)
Excess of Asphalt
Excess of Water
Smooth Polished Aggregate Particles
Lack of proper interrelation between layers of pavement structure
Slippage Cracks
Lack of Bond Between layers
Surface Course too Thin
Heavy Traffic Thrust
Weakness in base, subbase or underlying basement soils
Cracking
Plastic Deformation of Supporting Layer
Resilient Foundation
Deep Grooves Transverse Waves
Plastic Deformation of Base
Insufficient Base
Complete Break Through
Poor Foundation

Chart A

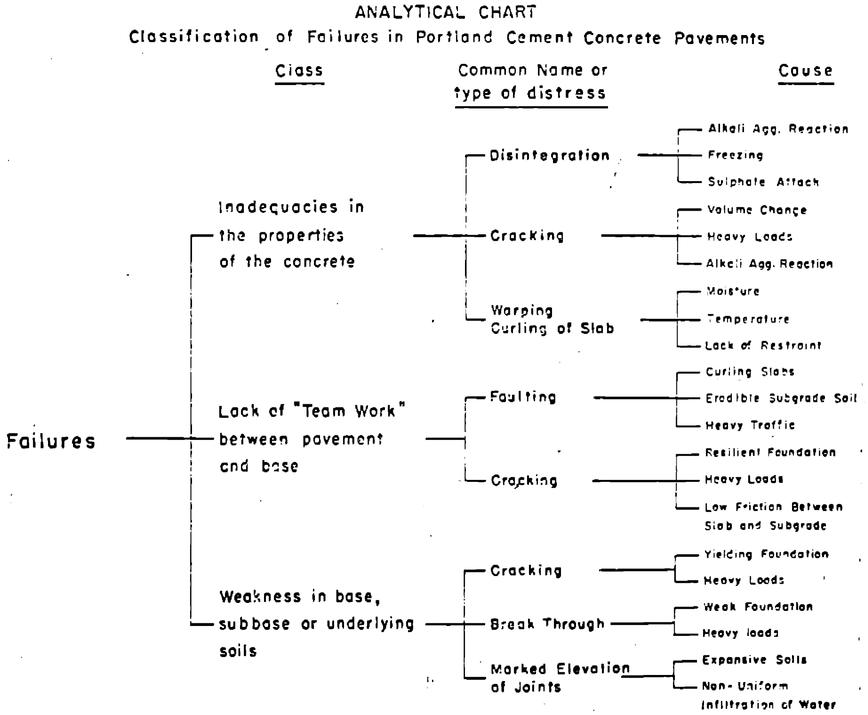


Chart B

from Hveem ( )

Figure V -8 : Typical failure modes

affect primarily the pavement's serviceability with respect to rideability and structural integrity. A complete description of the failure modes possible would lead to inclusion of "distress types" such as loss of surface friction due to bleeding of excess oil in asphalt or polishing of aggregate in concrete, or accumulation of excess water (leading to hydroplaning) due to interaction of heavy rains and microroughness properties.

In general, one would check each type of distress in proposing a design alternative, to assure that it has at least been accounted for. Of course, there is some probability of occurrence for each, which would be assessed in the reliability analysis. In the case of highway pavements, there are mathematical equations which allow one to check a number of possible failure modes simultaneously. In particular, a number of recent design methods have been based upon assuring a minimum level of what is here called coefficient of rideability. In so doing, a number of types of distress are implicitly accounted for. Longitudinal roughness, cracking, and transverse roughness (rutting), as a response to vehicle loads and base materials, are included. As will be discussed more fully in the numerical examples, one might choose to lump the failure modes thus accounted for together under the title of normal failure, suggesting that a design alternative is proposed with these modes in mind. Further, because there are explicit mathematical models involved, reliability analysis for these modes will follow an analytical path.

For the other failure modes, those not considered analytically, one will pursue an activities approach. The Markov models discussed previously then become a useful first approximation of behavior. For example, referring back to Figure 8, consider the case of "lack of 'team work'" in a Portland cement concrete pavement. Cracking and faulting are the two modes of distress suggested. The factors which influence cracking are included in the design method used in the forthcoming examples. This is then a facet of what will be termed normal failure, because it is what the pavement system is proposed to resist in normal service life behavior.

Faulting however is not included. The presence of an erodable soil, a precondition for failure in this mode, is not by itself sufficient to cause failure. A general sequence of events leading up to the occurrence of serious faulting might be as follows: an opening occurs in the pavement surface; water enters the opening and reaches the erodable soil; heavy vehicle loads are applied in sufficient number to cause erosion; the resulting loss of support becomes sufficient to cause the faulting. Note that warping of slabs is one way the initial opening might occur. Loss of joint filling or cracking could also serve this purpose.

These steps might then be represented in a state space of partial lifetime, as has been discussed. One then may make the reliability analysis by computing the conditional probabilities of occurrence of cracking and faulting - normal or abnormal failure - and then the probability of "lack of 'team work'".

Ideally, of course, there should be no "abnormal" modes of failure. All possible modes would be understood well and explicitly accounted for in analytical search procedure. The activities approach to reliability would then be used exclusively for investigation of operation and maintenance policies. Of course, the final analysis combines the various failure modes and the ways in which they might occur to yield an estimate of system reliability, the probability that serviceability will be adequate throughout the design life.

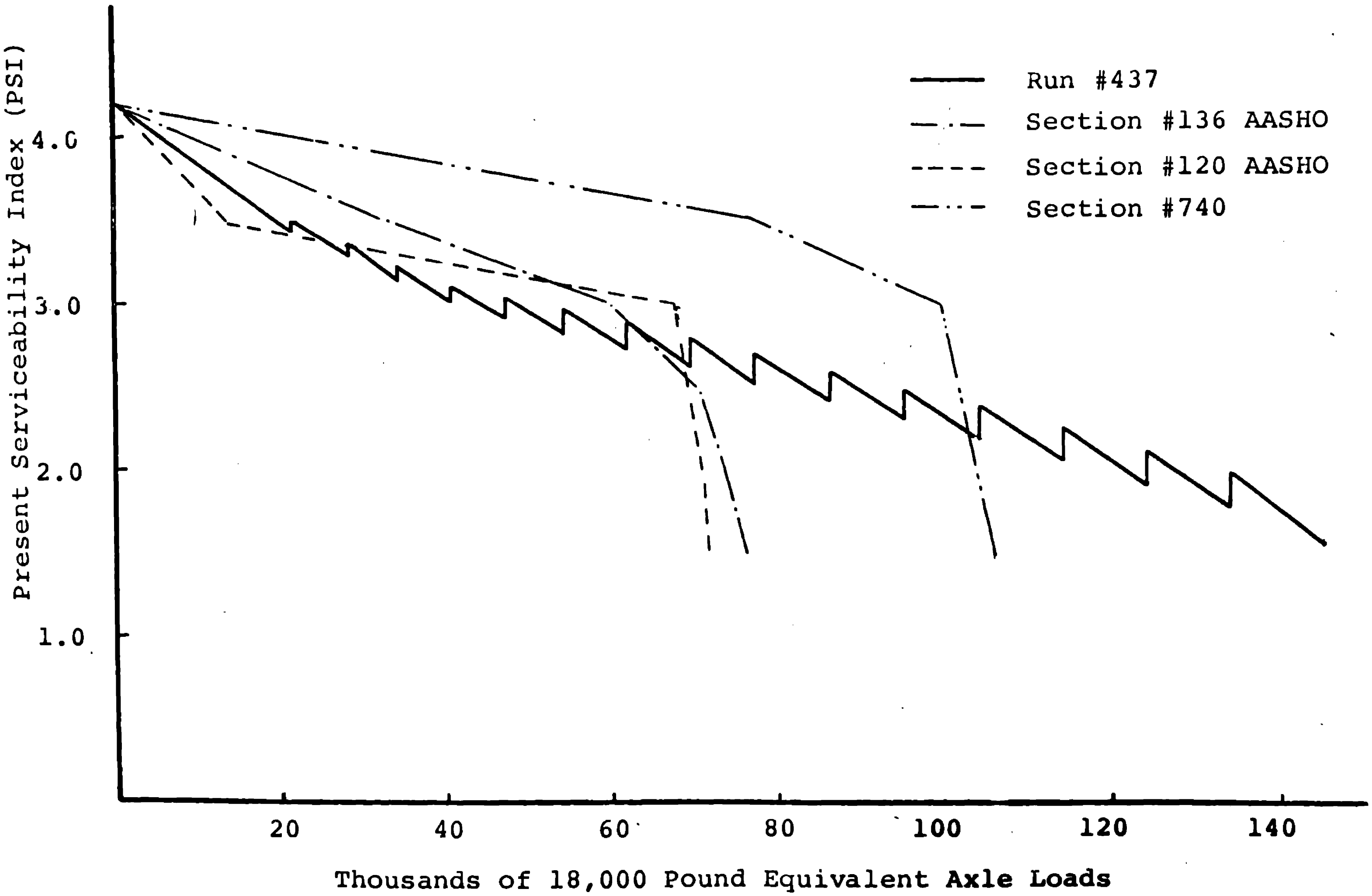
D. Highway Pavement Maintainability

It was suggested in previous discussion that maintainability for constructed facilities may be investigated with the aid of computed coefficients of maintainability, with respect to normal and repair maintenance operations. These coefficients provide a rough measure of the sensitivity of the service life behavior of the facility to maintainence efforts scheduled to be made during the service life. It may be noted that both measures vary inversely with the fraction of service life that might be lost if the scheduled operations are neglected. Stated another way, if a system is proposed to require no maintenance effort throughout the design life, then maintainability will tend to become very high, subject to reductions caused by possible associated losses of reliability. That is, it is expected that little or no effort will be required during the design service life to keep service at adequate levels.

Viewed in these terms, it may be suggested that the maintainability of current highway pavement systems is quite high. The pavement system, as it is viewed in much of current practice, emerges from the initial design decision without consideration of future maintenance needs. That the consideration of maintenance can be rationally included in the early stages of the decision-making process has been amply shown by Alexander (22).

Alexander's work comprises a valuable review of highway maintenance and its role in system costs and service behavior. The interested reader is referred to this work for a fuller presentation of maintenance than is desirable here. It will be worthwhile to illustrate here how Alexander's work fits within the present framework for analysis of highway pavement systems.

Figure 9 is reproduced from Alexander's work, and shows the behavior of several pavement systems. The three AASHO sections have essentially no maintenance activity, while the saw-tooth shape of the PSI trend (the AASHO serviceability definition - here the coefficient of rideability) for run #437 is due to the execution of maintenance actions. All systems were proposed to meet the same normal failure criteria.



PSI vs. Applications of 18 kip Equivalent Axle Loads

Figure V -9 : Pavement ageing trends (from Alexander (22))

For the purposes of illustration, suppose that these four curves are predictions of the expected value of coefficient of rideability versus axle load applications. In fact, this supposition is true for run #437, while the AASHO sections are historical data. If the pavements were proposed to resist 86,000 load applications (the number chosen as that at which run #437 has a rideability value of 2.5), then some conclusions on reliability and maintainability may be drawn.

Reliability for run #437 is defined to be 0.50, as the expected value (i.e., the mean) of 2.5 occurs at 86,000; e.g., it was so set up in the preceeding paragraph. It is apparent that the reliability of section #740 must be much higher than 0.5, as the expected failure age is slightly above 100,000 repetitions. Similarly, sections #136 and #120 have expected failure ages of roughly 70,000 repetitions, and would thus have reliability lower than 0.5.

On the other hand, the three AASHO sections have coefficients of maintainability, with respect to normal maintenance, which may be quite large. The coefficient is given as

$$M_N = \frac{T_D}{T_N} \left[\frac{P(\text{Normal Maintenance})}{P(\text{Normal Maintenance}) - (1-R)}\right]$$

where;

T_D = design service life,

T_N = expected service life lost if normal maintenance is neglected

P(N.M) = probability that normal maintenance will be executed

R = reliability.

The ratio of T_D/T_N is undefined, becoming infinitely large as T_N goes to zero, the case if no normal maintenance is scheduled. It might be argued that M_N is higher for Section #740 than for the other two AASHO sections, because its higher reliability results in a lower value of 1-R, thus sending the product $T_N(1-R)$ to zero faster. This argument is of only passing interest here.

Of interest here is the estimation of maintainability with respect to normal maintenance for run #437. Figure 10 projects the results of neglect of normal maintenance. Failure would be expected to occur at 53,000 applications. The ratio T_D/T_N then assumes a value of 86,000/(86,000-53,000), or 2.6. If it is assumed that the maintenance organization is good, so that P(normal maintenance) tends toward 1.0, - i.e., the scheduled maintenance will certainly be executed - then

$$M_N = \frac{86}{33}\left(\frac{1.0}{1-0.5}\right)$$
$$= 5.2$$

This number if of value to the analyst as a means of comparison with other possible alternatives.

The aim of the above discussion was to demonstrate the evaluation of maintainability and to illustrate the linkage between the present work and Alexander's work, which is part

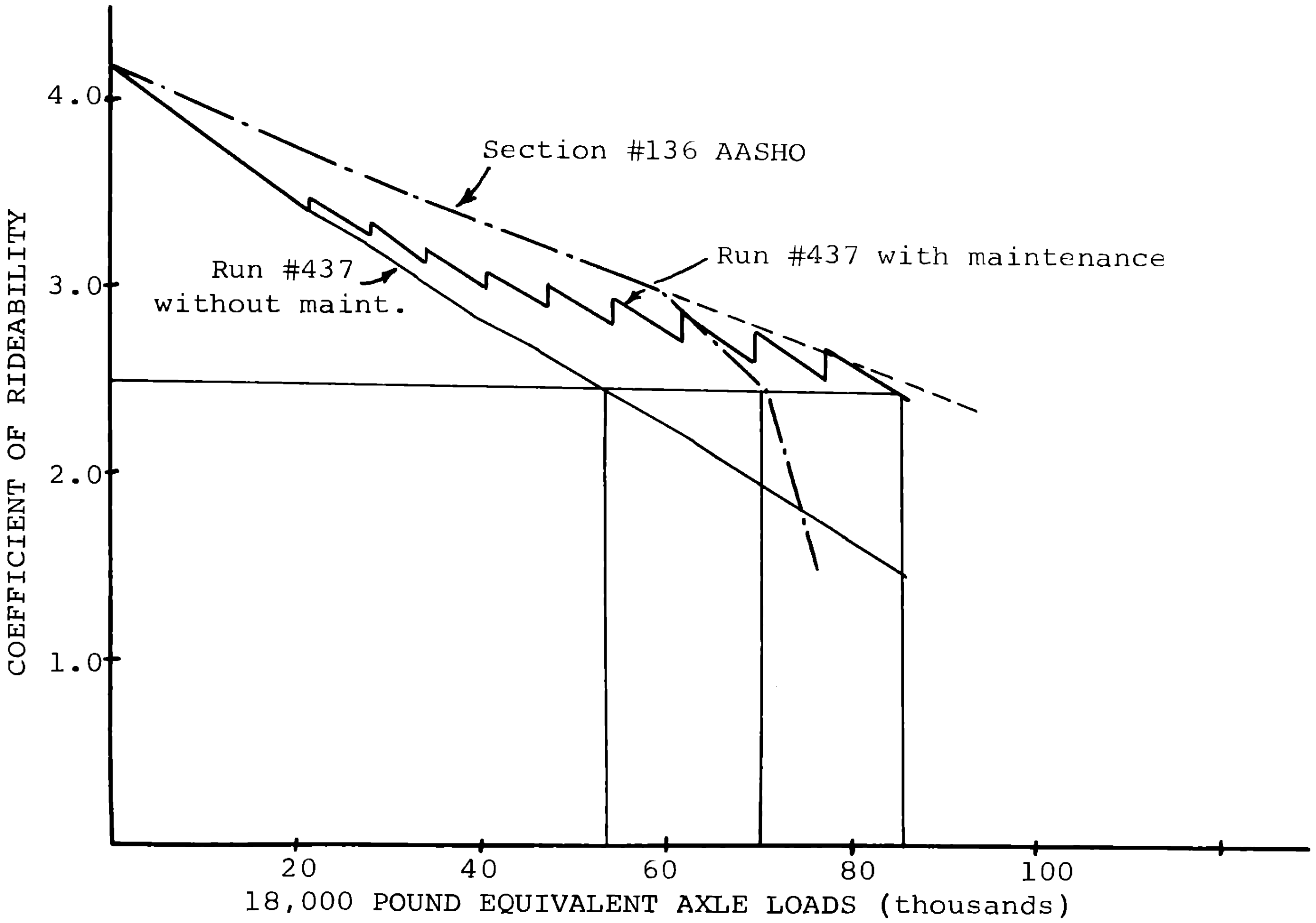


Figure V -10 : The impact of maintenance on rideability (adapted from Alexander (22))

of a larger set of models for estimating highway costs. These models can form a valuable link with the present work, as a means for integrating design and planning activities.

E. Examples

To illustrate the application of ideas presented here, the discussion will now proceed in terms of a numerical example. Assume that a rural interstate road is to be designed to handle light traffic for a 15 year design service life. The design life will have been determined as desireable by analysis on the broader scale of the transportation system. The overall transportation analysis would also provide traffic projections and the minimum acceptable serviceability levels, as discussed previously.

Figure 11 shows the traffic growth pattern predicted. From an initial demand prediction, a constant percentage growth rate is projected to compute a total number of vehicles, or in this figure, equivalent 18 kip axle loads. Assume that a legal load limit is set at 30 kips gross weight, implying a serviceability requirement with respect to structural integrity of 0.97. (Refer to Figure 7).

As was discussed, it might be decided that serviceability should be kept at or above the 0.50 level in order to assure the economic justification of the project. Very roughly, this requirement might be understood as saying that, of the potential demand generated, 50% of the users will consider this highway to be at least an acceptable trans-

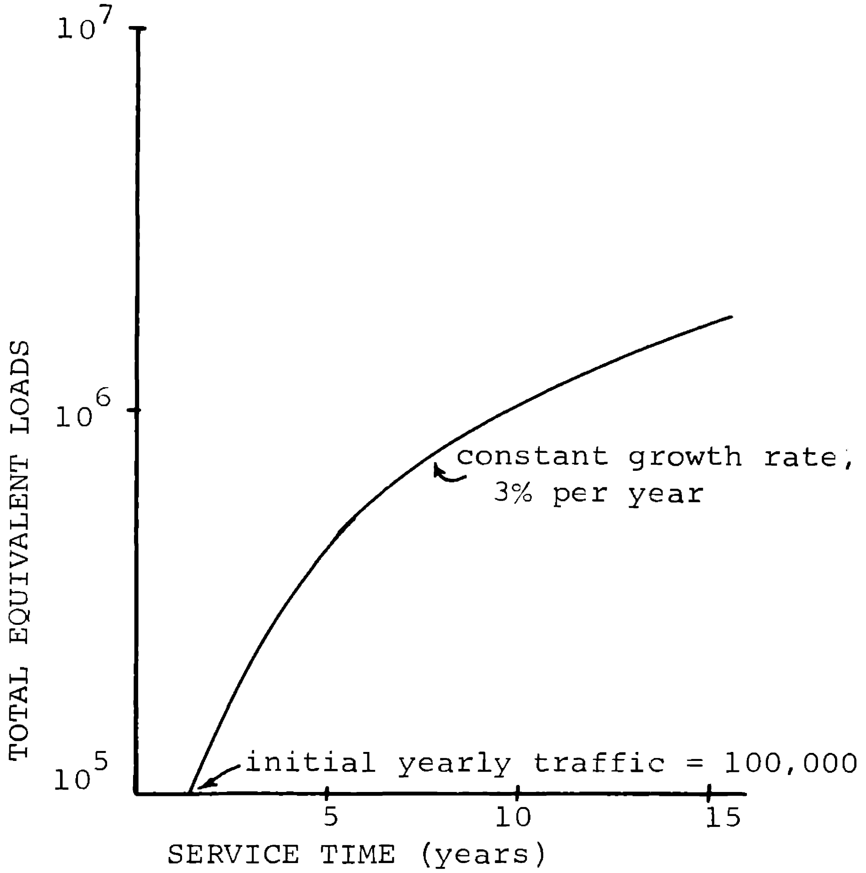


Figure V -11 : Traffic projections for pavement example

portation alternative. Of course, this argument applies primarily to safety and rideability, as a structural integrity failure level has been set (in a dominance relationship) at 0.97.

This then defines the service requirements for the pavement. The next step is to proceed with a search for possible solutions.

The initial search approach is by means of standard design methods. It may be assumed that the decision maker immediately restricts himself to standard rigid or high-type flexible pavements. If the traffic volumes were lower, this might be the case.

The standard methods to be used here are derived from the AASHO Road Test and subsequent studies of that sort. Thus, initial search is conducted in terms of rideability, and the concept of load equivalency, mentioned earlier, is based upon the number of load applications required to bring about a given loss of surface riding quality. This assumption is necessary before Figure 11 can be derived from initial traffic studies.

The formulas to be suggested are used for illustration cnly. Other suitable models are available and could have been used.

For rigid pavement, the extended AASHO equation (23) will be used. This model is as follows:

$$\text{Log } \Sigma W = -9.483 - 3.837 \text{ Log}[\frac{J}{S_x D} (1 - \frac{2.61a}{Z^{\frac{1}{4}} D^{\frac{3}{4}}})] + \frac{G}{\beta}$$

where;

ΣW = number of equivalent 18 kip loads to failure

J = a coefficient depending upon slab continuity

S_x = 28 day concrete modulus of rupture (psi)

D = pavement scab thickness (inches)

Z = E/K

E = concrete modulus of elasticity (psi)

K = modulus of subgrade reaction (psi/inch)

a = radius of loaded area = 7.15 in

G = $\log (\frac{P_o - P_t}{P_o - 1.5})$

P_o = initial rideability index

P_t = terminal (failure) rideability index

β = $1 + \frac{1.64 \text{ x } 107}{(D+1)^{8.46}}$

For flexible pavements, an extension of the Asphalt Institute method (24) will be used. This model is as follows:

$$\text{Log } \Sigma W = \frac{1}{3.97} (6.13 + T_A (CBR)^{0.4}) + \text{Log } \gamma$$

where;

Log W = equivalent 18 kip loads to failure

T_A = total equivalent thickness of asphalt (in.)

= $a_1D_1 + a_2D_2 + a_3D_3$

a_1, a_2, a_3 = layer strength equivalencies

D_1, D_2, D_3 = layer thicknesses (in.)

CBR = California Bearing Ratio $\gamma = \frac{\text{Log } P_o - \text{Log } P_t}{\text{Log } P_o - 0.4}$

P_o = initial rideability index
P_t = terminal (failure) rideability index

The extension is discussed in Appendix B.

These two models predict the total number of loads to failure (a predefined level of rideability) for a given design (pavement thickness). The implication of using these models is that loss of rideability will now be referred to as normal failure, as this is the expected mode. Both equations may be inverted to give a thickness of pavement required to resist the desired number of loads.

To arrive at alternatives possessing a reasonable level of reliability, one might initially design pavements for expected traffic higher than the actual projection. Since the actual behavior of the pavement is dispersed about this mean, reliability relative to the actual projection of load will then be computed. Thus, the initial guesses for standard rigid and flexible pavements might be made by assuming a 5% growth rate, leading to roughly 2.1 million load applications over 15 years, rather than the 3% projection which defines actual service requirements.

Assume that soils explorations indicate preliminary subgrade design values of CBR = 3 (or k = 100 psi/in.), and that other necessary values of system parameters are chosen by the designer from past experience. See Figure 12.

DESIGN CONDITIONS

Initial Traffic - 200 vpd (18 kip equivalent)
Traffic growth rate - 3% per year
Design life - 15 years
Soil (subgrade) - CBR = 3, k = 100 pci

I. Portland Cement Concrete

9" s_x = 400 psi

II. Asphalt Concrete

6" Type IV asphalt
12" granular base

FIGURE 12: Basic Design Alternatives

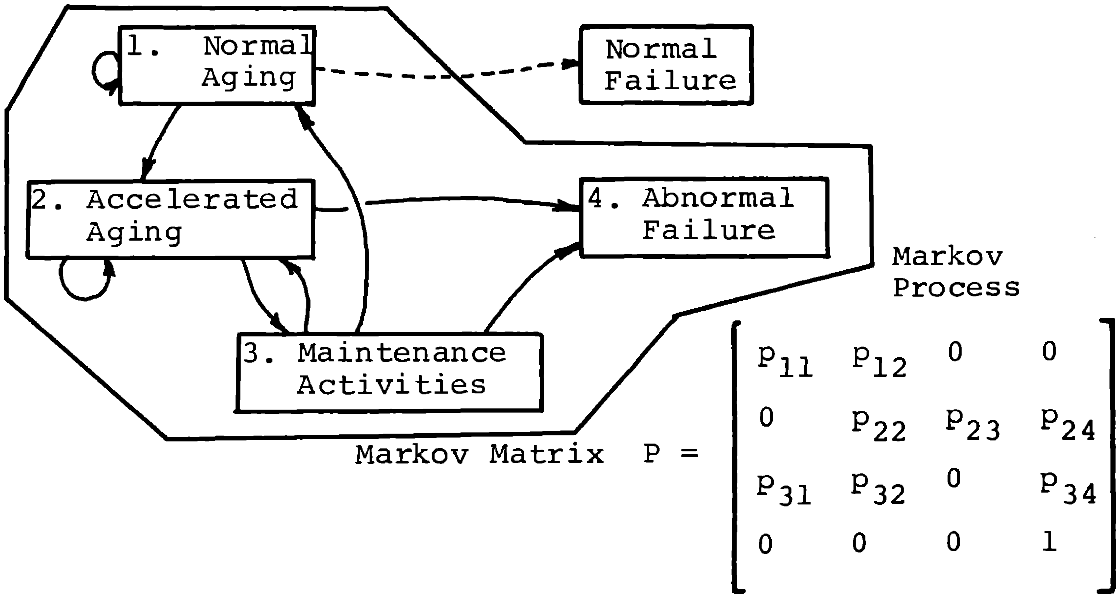
Trial designs of a 9 inch rigid pavement and a 12 inch (total equivalent asphalt thickness) flexible pavement are derived. It is suggested that these are approximately the configurations that a pavement designer, proceeding in standard fashion, might have proposed (25,26). The above are preliminary designs. They must be constructed and will be subject to problems of construction control and uncertainties of materials and environment. This opens up the possibility of various strategies in implementation, some of which are suggested in Table 1. A Monte Carlo simulation was undertaken to compute probability estimates for normal failure

STRATEGY	COEFF. OF VARIATION	RELATIVE COST	PROB. OF NORMAL FAILURE
Asphalt	5 %	1.30	0.24
	10 %	1.10	0.36
	15 %	1.00	0.44
Concrete	5 %	1.69	0.06
	10 %	1.43	0.22
	15 %	1.30	0.30

Table 1 : Quality control strategies

under each strategy.

This analysis gives the probabilities of failure for several partial designs, with respect to losses to rideability. The distributions derived in simulation give the probability of loss of minimum acceptable serviceability as a function of total number of load applications. But the designs must be completed to include the operation and maintenance stages and the other possible modes of failure. Because the alternatives being developed so far were proposed in reference to one particular mode of failure, these other modes are abnormal failures -- that is, as far as the design algorithm is concerned, these modes cannot occur or do not exist. (Note that so far at least six alternatives are potentially suggested: standard flexible and rigid pavements with three levels of construction quality).



Markov Matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & 0 & 0 \\ 0 & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & 0 & p_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

FIGURE 13: Service Life Markov Sub-Model

Figure 13 suggests an extended but quite simple model for a highway pavement, presented in terms of a state space. The service behavior within this system may be modeled in part as a Markov process. An explanation of the states is in order.

The first state, normal aging, refers to the expected behavior of progressive deterioration of rideability over the period of 15 years. Based upon traffic growth and materials quality, there is a certain probability that the system will go to the normal failure state, as was computed by Monte Carlo methods above.

Given that the process has not gone into the normal failure state, it will be in the Markov process which includes "other" factors. The second state is termed accelerated aging. If some event occurs, such as cracking of loss of joint filling, which sharply changes system characteristics such as the stresses in the pavement or the access of water to the base, then the damage or deterioration of the pavement resulting from each passage of a vehicle is likely to be greater than normal. The deterioration rate increases and with it the possibility of premature failure. The probability of passing into this accelerated aging state will depend upon construction quality, operating control, environmental factors and the normal maintenance policies.

Dependant upon normal maintenance or inspection policies, there is some probability that the process will enter this third state, in which repair maintenance procedure are begun to stop the accelerated aging. These repair maintenance actions, depending upon their quality and adequacy, will return the system to its initial normal aging state or to the accelerated aging state. Or, if repair maintenance is inadequate (which includes the possibility that it is not initiated in time) the system will possibly experience abnormal failure, the fourth state.

Two features of the model are artifacts of the modeling process. First, because probabilities are figured with respect to a particular period of time, the possibility that the system might remain in a particular state is admitted. In this case a six month period is used. Second, this model is intended to represent gross consideration of operating and maintenance policies, for example at this initial planning stage. The states are thus highly aggregated. For more detailed investigations, i.e., in maintenance planning, the states could be broken down into more refined descriptions and the Markov matrix expanded accordingly. On the other hand, some of the data used in this example was more detailed than desirable and had to be reduced to appropriate values for the model. Model size and refinement are a function of decision level and computational resources.

Table 2 reviews four alternative operating-maintenance policies in terms of their Markov transition matrix representations. These matrices represent probabilities on a basis

DESCRIPTION	P MATRIX	REL. COST
I. Standard operating policies	$\begin{bmatrix} .95 & .05 & 0 & 0 \\ 0 & .40 & .20 & .40 \\ .60 & .30 & 0 & .10 \\ 0 & 0 & 0 & 1.0 \end{bmatrix}$	1.00
II. High maintenance activity, standard quality	$\begin{bmatrix} .95 & .05 & 0 & 0 \\ 0 & .40 & .50 & .10 \\ .60 & .30 & 0 & .10 \\ 0 & 0 & 0 & 1.0 \end{bmatrix}$	1.10
III. Standard maintenance activity, high quality	$\begin{bmatrix} .95 & .05 & 0 & 0 \\ 0 & .40 & .20 & .40 \\ .80 & .10 & 0 & .10 \\ 0 & 0 & 0 & 1.0 \end{bmatrix}$	1.05
IV. High maintenance activity, high quality	$\begin{bmatrix} .95 & .05 & 0 & 0 \\ 0 & .40 & .50 & .10 \\ .80 & .10 & 0 & .10 \\ 0 & 0 & 0 & 1.0 \end{bmatrix}$	1.15

Table 2 : Operating and maintenance policies as a Markov sub-model.

of discrete stops over a period of six months. This is, each entry P_{ij} is the estimated probability that the system, given that it is in state i when inspected now, will be in state j when inspected again in six months. Associated with the changes in probabilities would be changes in costs.

There are now a total of 24 possible alternatives under consideration: standard flexible and rigid pavements, each with three levels of construction quality and four operating-maintenance policies. Table 3 summarizes these alternatives and gives the reliability for each, as computed through Monte Carlo simulation of normal failure linked with the Markov model of abnormal failure possibilities.

Coefficients of maintainability may also be computed from the Markov sub-model. By the way in which the model was constructed, repair maintenance has been excluded. That is, it has been assumed, for the computation of reliability, that once a failure state occurs, no repair is made. Hence, no coefficient of maintainability for repair is defined. For normal maintenance policies, it is possible to compare the four strategies in Table 2.

One finds from the Markov sub-model that the expected time to failure under strategy I is 20 periods. Strategy III gives a value of 21 periods, while II and IV have expected times in excess of the 30 period design service life. Coefficients of maintainability are defined relative to Strategy I, which is the minimum activity level. If the

DESIGN	CONST. QUAL.	OPER. POL.	RELIABILITY	RELATIVE COST	COEFF. OF NOR. MAINT.
Asphalt	5 %	I	0.30	1.30	—
		II	0.57	1.43	8.5
		III	0.31	1.62	8.7
		IV	0.60	1.65	3.8
	10 %	I	0.25	1.10	--
		II	0.48	1.21	2.9
		III	0.26	1.16	8.1
		IV	0.50	1.27	3.0
	15 %	I	0.24	1.00	--
		II	0.46	1.10	2.8
		III	0.25	1.05	8.0
		IV	0.47	1.15	2.9
Concrete	5 %	I	0.35	1.69	--
		II	0.67	1.85	4.5
		III	0.37	1.77	9.5
		IV	0.71	1.94	5.2
	10 %	I	0.34	1.43	--
		II	0.66	1.57	4.4
		III	0.36	1.74	9.4
		IV	0.70	1.64	5.0
	15 %	I	0.31	1.30	--
		II	0.61	1.43	3.8
		III	0.33	1.62	9.0
		IV	0.64	1.65	4.2

Table 3 : Summary of initial alternatives

increased activities of Strategy II are neglected, and abnormal failure occurs, the expected service time lost will be on the order of 10 periods.

Coefficients are then computed relative to Strategy I, and are summarized in Table 3. It is interesting to note that Strategy III, which has activity only slightly above that of the base level, yields coefficients of the same order of magnitude as II, which features substantial dependence on maintenance. This result is due to the correspondingly lower values of reliability and probability that maintenance will be carried out.

Table 3 is a summary of the service behavior analysis. One more variable which could be given is the expected serviceability at the end of the 15 year design service life. This was not computed for these 24 alternatives because they were all proposed to have the same aging characteristics. Any differences in expected failure age are due to the interactions of probability distributions. That is, a greater spread in, for example, concrete strength gives a slightly higher probability of high strength values, and thus possibly a rise in the mean number of loads to failure. Had a series of alternatives been proposed with stronger materials or thicker layers, higher expected serviceability at the end might have been found, with associated higher reliability, and of course higher costs. In this example, one might assume that there is no value attached to this form of effec-

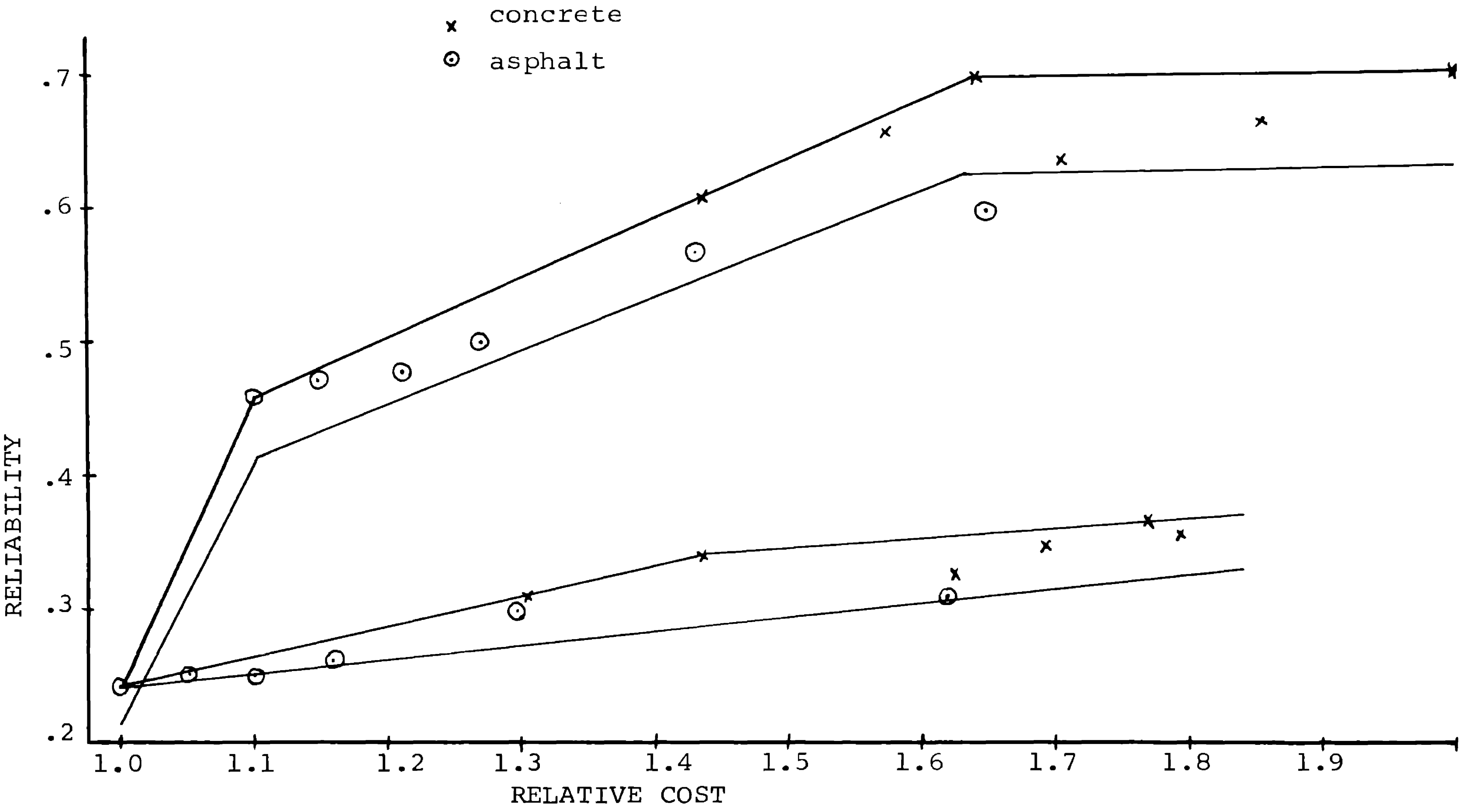


Figure VI-14 : Reliability as a decision variable

tive overdesign.

Another factor which in this example has been neglected somewhat more than it might be in a real case is that of cost. The relative cost factors estimated here might be viewed as statements of total present values. It is reasonable and in fact desirable to view these total costs in terms of their components of construction, operation-maintenance, and user costs. That is, decision will consider not simply the magnitude of costs, but also the distribution. Some further discussion of this will be undertaken in Appendix C, but a full treatment is beyond the scope of this investigation. It is hopefully apparent that these factors can be computed, and the reader is referred to work of Manheim et al. (27) or Stafford et al. (19) for further discussion.

F. Trade-offs and Decision

The analysis has now progressed to the point that a series of alternatives have been proposed, and have been checked or refined to yield at least acceptable performance. One must now consider economic efficiency to define production functions.

Figure 14 presents graphically some of the data of Table 3. All 24 alternatives are plotted to show reliability versus relative cost. This display is one view of economic efficiency, in that it suggests those alternatives which possess highest reliability at a given relative cost. Five alternatives form an efficiency frontier or production function.

Clearly, from this viewpoint, roughly half of the alternatives are inefficient. These alternatives are in fact those having high maintainability, i.e., low need for maintenance, suggesting that a trade-off of maintainability for reliability is advantageous in an overall cost framework. That is, it would seem that maintenance effort is a worthwhile investment in terms of improved performance.

If, as has been suggested in previous discussion, reliability is a principal concern for the subsidiary user, a group which includes the planner-decision-maker, then Figure 14 might be a first step in the decision process. A desire for high reliability would lead to immediate elimination of the alternatives not at or near the efficiency frontier. Recognizing the inaccuracies of prediction which enter an analysis using such aggregated models as those used here, one might choose a range of, for example, 10% below the efficiency frontier to define a cut-off for further consideration of alternatives. Twelve alternatives survive application of this criterion.

One might follow a similar approach for maintainability, preparing Figure 15 for the eleven alternatives for which the coefficient of maintainability is defined. Using the same 10% screening rationale as above, 7 alternatives are acceptable.

With this step, one finds a total of 8 alternatives,

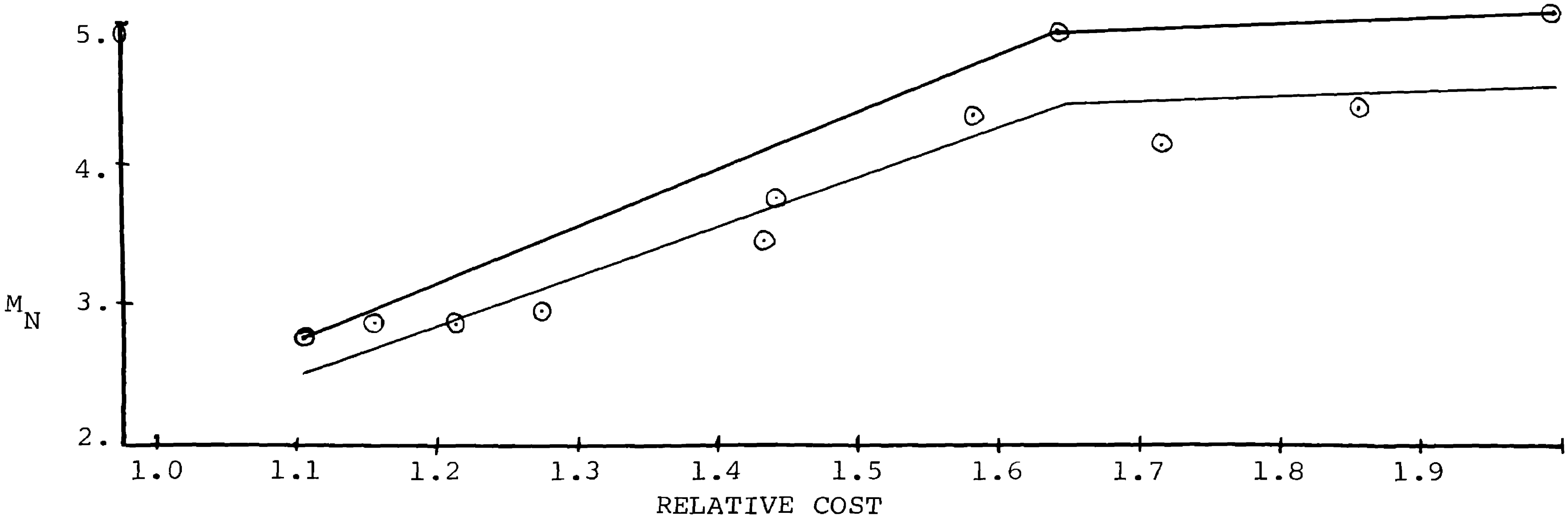


__Figure V -15__ : Maintainability as a decision variable

or a third of the original 24, which would appear to satisfy the goal set for this system of constructed facilities. These are summarized in Table 4. These alternatives exhibit the feature that both reliability and maintainability increase with relative cost. This is in large part due to the common bases from which the alternatives were derived, a factor which also manifests itself in the clustering of bituminous alternatives at the low end of the cost scale versus concrete at the upper end. In general, one would not expect these aspects of performance to be so well correlated. Decision would of course present a much more difficult problem.

As it is, the decision to be made in this example is cast as one of how much to spend for how much service. Although they have been skirted in this example, there would also generally be questions (as has been suggested) of distribution of cost. For example, the third entry in Table 4 has a higher reliability than the first, and a higher cost. However, the difference in cost is due to scheduled maintenance activities, and will occur after construction. Hence, lack of currently available funds may prove no deterrent to the selection of the higher alternatives, if there is hope that the additional funds will be available. On the other hand, fear of dependance on the quality of future maintenance activities, as features in all but the first of these alternatives, could lead one to continue searching for possible solutions.

ALTERNATIVE	REL. COST	RELIABILITY	COEFF. OF MAINTAINABILITY
A15 I	1.00	0.24	--
A15II	1.10	0.46	2.8
A15IV	1.15	0.47	2.9
A10II	1.21	0.48	2.9
C15II	1.43	0.61	3.8
C10II	1.57	0.66	4.4
C10IV	1.64	0.70	5.0
C 5IV	1.99	0.71	5.2

Table 4 : Relatively efficient alternatives

Now consider the use of the analytical model during the service life to control service in an efficient, dynamic manner. Assume that it was decided to construct a concrete pavement with low-quality control and a policy of frequent maintenance inspection (i.e., the 5th alternative in Table 4).

Now, some time after service has begun, data is collected and compared with predictions. (Note: this discussion could rapidly be turned into a demonstration of statistical decision theory, which is beyond the scope of this work. Therefore, the following is just a presentation of possibilities). Suppose, for example, that at the end of five years, it is found that traffic has grown at a 6% rate, at the expense of a rapid transit system. Hence, the regional planners would like to slow growth on traffic, rather than accomodate total demand.

A possible means of cutting demand is to permit some deterioration in rideability, increasing user discomfort and effective user costs. This policy could be undertaken through judicious neglect of maintenance activities. The result of this action should be to encourage some users to take other transportation modes.

First, maintenance effort could be reduced. In effect, the probability of a transition to accelerate aging would be increased. Some savings would be made in the maintenance accounts, at the expense (in this case desirable) of operating costs. Linked with this action could be another dealing with repair actions. The maintenance policies would be modified to decrease the probability that the system would enter the abnormal failure state, while raising the probabilities of transition to maintenance and back again to rapid aging. These actions would keep reliability at the same or a higher level while lowering the expected serviceability.

Examples could be presented and expanded at infinitum, but hopefully the puspose of illustration has been achieved. It is felt that the application of the principles of analysis which have been presented in the area of highway pavements is desirable and feasible. Further, such applications are immediately practical as substantial quantities of useful data are being collected in the highway field, although substantial effort will be required to reduce it to useful forms.

CHAPTER VI

SUMMARY AND EVALUATION

A. Summary

The objective of this study has been to present a framework for analysis of systems of constructed facilities, a framework which will be useful in resource allocation at the level of design decision. This analysis is intended to provide information to decision-makers about the activities of design, implementation, operation, and maintenance through which a facility is realized, and about the relation of resources required in these activities to the services which a facility delivers.

Central to the framework proposed here is the view that a constructed facility is intended to provide service to users. The manner in which a facility provides these services is termed its performance, and is evaluated over the facility's design service life in terms of serviceability, reliability, and maintainability. Together, these three components of performance estimate the present adequacy of service and the likelihood of continued adequacy.

The physical behavior of a constructed facility is predicted in probabilistic terms. Stochastic models of behavior are utilized, based upon the processes involved in aging and deterioration or upon observed events leading up to and accompanying failure. In the latter case, a Markov process will often provide a useful representation of service behavior.

The predicted physical service qualities are then evaluated with respect to users' needs and desires. Such evaluation depends upon factors of subjective response, for which techniques based upon psychological scaling procedures are useful.

Comparison of a number of possible alternative constructed facilities leads to the development of a "production function" or frontier of efficiency of performance with respect to resource usage. This frontier represents the output of the design analysis, a set of alternatives exhibiting qualities of adequate service throughout the facility's design service life, in a relatively efficient manner. Selection of any single point on this frontier for implementation must be made within a context of planning considerations.

The use of this framework and tools for its application are illustrated for the case of highway pavement. The possibility for tradeoff among activities in construction, operation, and maintenance is brought out through the assessment of the effects of such activities on performance and cost. It is shown not only that the present framework may be used as it is intended but also how it compares with more traditional approaches to analysis.

B. An Evaluation

Any evaluation of the ideas presented here must contend with two principal questions: First, to what extent does this approach actually foster interchange of information between planning and design levels of decision. That is, what role does this analysis play within a larger scheme of things? Second, can the analysis in fact be used, or are the data requirements and difficulties of application excessive for most systems of constructed facilities?

The first question may be approached through a brief look at work being done at the level of planning decision, with a view toward how the present work at the design level might be integrated with these efforts. This approach will elaborate in specific fashion upon points examined more generally throughout this thesis.

For example, Kilbridge, et al. (1), in developing their views on urban analysis, use housing as an illustrative case. They focus their discussion on the financial sector, and examine factors influencing the profitability of housing. Their motivation apparently lies in the view that increased profitability will serve to spur increased production and thus to alleviate a currently perceived shortage in housing. They concentrate upon the interaction of political and economic factors which will effect the physical system.

The present work may serve as input to their analysis through description of current and possible future production technology. Another close link is established through the influence of the physical system upon demands placed upon the facility in the consumption sector, and vice versa. And finally, conclusions they have reached such as the possibility that shorter service lives for housing would stimulate investment by effectively increasing tax shelters, have obvious direct impact upon the alternatives found in the analysis of the system of constructed facilities.

A similar situation is encountered in Forrester's work in urban dynamics (2). In this case, the view of the urban system is considerably more aggregated than that of Kilbridge, et al. Yet such factors as the aging rate of housing and industry are explicitly considered, and opportunities to investigate variations in demand are apparent.

In the field of transportation analysis, the concept of level of service as a determinant of demand has received increasing attention. The role of this analysis in that field has been discussed at several points throughout the work. Indeed, this field has served as a background for much of the thought behind this work.

Slightly further afield are the quite interesting possibilities for pursuing these ideas in the evaluation and selection of projects for international and regional development. An idea which has only recently begun to achieve recognition is that there may be certain factors inherent in the technology of a project which makes it more or less likely to be successful in the environment in which it is undertaken. (See Hirschman as an example (3)). The term successful in this context of course refers to the contribution the project makes to the development. For example, projects of low maintainability may prove more useful than otherwise equivalent projects of high maintainability. The former type, because of the deferral of resource consumption implied in low maintainability, will not only allow for, but may actually encourage the development of new technologies within the developing country. The long range effects of allowing people to learn from the maintenance activities may be significant. This is an area which may prove of considerable interest.

Hopefully, the ability of the ideas described here to be applied effectively to a broad range of facility types has been affirmed in previous discussion. A basic criterion for the selection of tools and procedures to be presented here was that they be useable. However, it must be recognized that substantial data in the form of experiment and observation will be required to apply and verify this analysis in any new area.

In particular, establishing the serviceability function will depend on experimental data gathered from users. While this data may to some extent be found indirectly, for example through literature reviews (as done here for serviceability with respect to safety of highway pavements), it is desireable to refer directly to the population of potential users, and the larger the sample, the better. This brings the investigation into the realm of the behavioral scientist, and one is referred to this literature for discussion of the problems of such sampling.

A second difficult area is in the structuring and calibration of stochastic models of a facility's behavior. Cases where sufficient observation has been made to permit a statistical approach to this problem are difficult to find. It is more likely, as previously suggested, that an approach using subjectively estimated probabilities in a context of statistical decision theory will be desireable, enabling a progressive learning from experience. This does require significant data collection and manipulation throughout a facility's service life.

CHAPTER VII

EXTENSIONS AND VERIFICATIONS: SUGGESTIONS FOR FUTURE WORK

A. Introduction

Quite often, the suggestions made for future work are primarily concerned with obtaining better estimates of particular parameters or more efficient ways of doing particular tests. In the present work, such particular objectives are of lesser interest, as there are some rather sizeable problems, the solution of which could prove useful.

These problems, of which a few will be briefly discussed here, fall roughly under three headings: applications, verifications, and extensions. There will of course be overlap among the three. "Applications" is intended to include concerns for the discovery and development of tools and techniques to permit a fuller application of the ideas presented herein to decision-making for systems of constructed facilities. "Verifications" signifies activities concerned with checking the logic and validity of decisions made with the help of this approach to analysis. "Extensions" will include activities to broaden capabilities for analysis and to integrate more fully the physical analysis with associated social, political, and economic concerns.

The following paragraphs will investigate each of these three areas. The specific questions presented are suggestive of work which should prove most productive in relation to the present effort, but are certainly not inclusive of all possibilities for valuable contributions.

B. Applications

While it has hopefully been illustrated that the ideas presented here can be applied to current problems with currently available techniques, there are several areas in which additional work would prove useful. The first and most obvious area is in analysis of particular types of systems of constructed facilities. The development of so-called performance-based building codes and design methods in engineering practice are one manifestation of such work. The objection to current practice in this area is that setting single-valued limits in codes might represent elimination of one important latitude in producing better service with greater efficiency. It is recommended that fuller recognition be given to the differences among users.

A second area has been discussed in previous pages, but is worthy of being mentioned again. This area is the development of applications of non-metric scaling procedures to systems of constructed facilities. These procedures, because they require a minimum of prejudgement on the part of the analyst, could prove of great value in discovering the dimensions of serviceability for particular systems.

The final area to be suggested is somewhat less specific than the two preceeding ones. It would be useful to explore means for extending the capabilities of the analyst for predicting the consequences of selection of particular facility configurations. The possibilities for a facility to become obsolete, thus, in effect, causing failure, are recognized but at this time are virtually impossible to predict.

In summary, three suggestions for additional work have been made:

1. Apply this approach to analysis to particular types of facilities. This will involve assessment of serviceability and should investigate construction of predictive stochastic models of service behavior.
2. Investigate the application of non-metric scaling techniques, as developed in marketing research, for the assessment of serviceability of systems of constructed facilities.
3. Explore means for improving the analyst's predictive ability, with special reference to the problems of obsolescence.

C. Verifications

When one proposes a way of doing things which is different, and hopefully better, than other ways, the problem of verification is important. Will analysis undertaken as described here in fact lead to greater understanding and better decision? One may approach this question by observing facilities in service and analysing for the consistancy of these observations with what the analysis might have predicted, or one can undertake specific experiments, that is, attempt to influence the observations. These possibilities are best explained in terms of specific examples.

For direct observation, housing could prove to be a useful example. There is clearly a correlation of some sort between the physical qualities of a dwelling unit and the rent which can be charged. An experienced developer can predict what a fireplace or a porch is worth in terms of increased value. And it is observed that as the age of a unit increases, and if good maintenance practices are not followed, the users' evaluation of service will decrease. If a full description of serviceability for housing were developed, it should be possible to observe this correlation in more than a qualitative fashion by investigating a large number of dwelling units within a city.

Similarly, one could observe an assortment of public housing projects. An evaluation of serviceability might be related to overall evaluations of whether the project was "successful" or not. One might expect that projects such as Columbia Point in Boston would prove to have low serviceability.

An experimental approach could be taken with a highway. One would expect that lower rideability should lead to a reduction in average vehicle speed, as a trade-off among components of serviceability. If one were, through adjustment of maintenance practices, to allow a section of well-traveled road to deteriorate more than an adjacent section, such a varying of speed might be directly observable. This principle is no doubt involved in the use of rough strips across a road on the approach to toll boothes, to encourage drivers to slow down well in advance of the stopping point.

Two more suggestions for future work are then given:

4. Make observations of facilities in service, and examine these observations in light of ex post facto analysis.
5. Consciously vary service, through judicious use of operation and maintenance, and compare with model predictions.

D. Extensions

An extension of substantial interest would be to pursue investigation of the links between social, political, and economic systems, and the system of constructed facilities. As briefly touched upon in previous discussion, there would seem to be a number of cases in which ideas used in this work and those used in other fields, such as social psychology and economics, have common points in their bases or derivation. An exploration of these common points could lead to a better understanding of how the physical system effects the non-physical systems, and thus to an improved ability to achieve desireable effects. This work would require a truly interdisciplinary approach, and an ability to see the implications for several fields, of a finding in any one.

Another area of endeavor which might be classified as an extension is the development of computer programs to assist the analysis. Work such as that of Alexander (1), which makes it possible to investigate a range of alternative actions rather quickly with the aid of the computer, will be quite useful in the search for solutions to the design decision problem. The principal area for development of such models probably will be in the prediction of lifetime service trends.

Thus, two possible areas for extensions of this work are suggested:

6. Explore the interactions of social, political, and economic systems, with the system of constructed facilities, and thus facilitate the ability to predict how the physical system will effect these other systems.
7. Develop automated means for using the analytical tools and techniques proposed herein.

REFERANCES

Chapter I:

1. Novick, D., ed., Program Budgeting : Program Analysis and the Federal Government, Harvard Univ. Press, Cambridge, 1966.

2. Bauer, R., ed., Social Indicators, M. I. T. Press, Cambridge, 1968.

3. Manheim, M. L., et al., Search and Choice in Transport Systems Planning : Summary Report, Report R68-40, Dept. of Civil Eng., M. I. T., Cambridge, 1968.

4. Hirschman, A. O., Development Projects Observed, Brookings Institution, Washington, 1967.

5. Stigler, G. J., The Theory of Price, Macmillan, Toronto, 1966.

6. Hirschman, A. O., The Strategy of Economic Development, Yale Univ. Press, New Haven, 1958.

7. Haikalis, G., Economic Analysis of Roadway Improvements, Chicago Area Transportation Study, Chicago, 1962.

8. Simon, H. A., Models of Man, Wiley, New York, 1957.

9. Grant, E., and W. Ireson, Principles of Engineering Economy, Roland Press, New York, 1964.

10. Solomon, M. J., Analysis of Projects for Economic Growth, Praeger, New York, 1970.

Chapter II:

1. Soberman, R. M., Transport Technology for Developing Regions : A Study of Road Transportation in Venezuela, M. I. T. Press, Cambridge, 1966.

2. Lago, A. M., "Cost Functions and Optimum Technology for Intercity Highway Transportation Systems in Developing Countries," Traffic Quarterly, Oct. 1968.

3. Manheim, M. L., et al., Search and Choice in Transport Systems Planning : Summary Report, Report R68-40, Dept. of Civil Eng., M. I. T., Cambridge, 1968.

REFERANCES (cont.)

4. Stigler, G. J., *The Theory of Price*, Macmillan, Toronto, 1966.

5. Bangs, R. B., *Financing Economic Development*, Univ. of Chicago Press, Chicago, 1968.

6. Palda, K. S., *Economic Analysis for Marketing Decisions*, Prentice-Hall, Englewood Cliffs, 1969.

7. Plourde, R., *Development of a Behavioral Model of Travel Mode Choice*, Ph.D. thesis, Dept. of Civil Eng., M. I. T., Cambridge, 1971.

8. Alexander, J. A., *Highway Maintenance*, Ph.D. thesis, Dept. of Civil Eng., M. I. T., Cambridge, 1970.

9. Guenther, K. W., *Predictive Models for Vehicle Operating Consequences*, Report R69-2, Dept. of Civil Eng., M. I. M. I. T., Cambridge, 1969.

Chapter III:

1. Arrow, K. J., *Social Choice and Individual Values*, Wiley, New York, 1964.

2. Fechner, G., *Elements of Psychophysics*, vol. I, trans., by H. E. Adler, Holt, Rinehart, and Winston, New York, 1966.

3. Guilford, J. P., *Psychometric Methods*, McGraw-Hill, New York, 1936.

4. Boring, E. G., "The Beginning and Growth of Measurement in Psychology," in *Quantification*, ed. by H. Woolf, Bobbs-Merrill, New York, 1961.

5. Miller, G. A., *Psychology: The Science of Mental Life*, Harper and Row, New York, 1962.

6. Thurstone, L. L., *Measurement of Values*, Univ. of Chicago Press, Chicago, 1959.

7. Stevens, S. S., "Measurement, Psychophysics, and Utility ity," in *Measurement: Definitions and Theories*, ed. by Churchman and Ratoosh, Wiley, New York, 1959.

8. ———, "A Metric for Social Consensus," *Science*, no. 151, 1966.

REFERANCES (cont.)

9. ———, "On the Psychophysical Law," *Psychological Review*, vol. 64, 1957.

10. Gulliksen, H., "Linear and Multidimensional Scaling," *Psychometrika*, vol. 26, no. 1, March 1961.

11. Winkler, R. L., "The Quantification of Judgement: Some Methodological Suggestions," *Jour. of the Amer. Statistical Assoc.*, vol 62, no. 320, Dec. 1967.

12. Galanter, E., "Direct Measurement of Utility and Subjective Probability," *Amer. Jour. of Psychology*, vol. 75, 1962.

13. Samuelson, P. A., *Foundations of Economic Analysis*, Athenium, New York, 1965.

14. Stigler, G. J., *The Theory of Price*, Macmillan, Toronto, 1966.

15. Schumpeter, J. A., *History of Economic Analysis* , Oxford Univ. Press, New York, 1954.

16. Simon, H. A., "A Behavioral Model of Rational Choice," *Quart. Jour. of Economics*, vol. 69, Feb. 1955.

17. Fishburn, P., "Utility Theory," *Management Science*, vol. 14, no. 5, Jan. 1968.

18. von Neumann, J., and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton Univ. Press, Princeton, 1947.

19. Luce, R, D., *Individual Choice Behavior*, Wiley, New York, 1959.

20. Katona, G., *Psychological Analysis of Economic Behavior*, Mcgraw-Hill, New York, 1957.

21. Coombs, C. H., *A Theory of Data*, Wiley, New York, 1964.

22. Winkel, G. H., "Community Response to the Design Features of Roads — A Technique for Measurement," presented at the Highway Research Board 49th Annual Meeting, Washington, Jan. 1970.

23. Green, P. E., Carmone, F. J., and Robinson, P. J., *Analysis of Market Behavior Using Non-Metric Scaling Techniques*, Marketing Science Inst. reprot, March 1968.

REFERANCES (cont.)

24. Coombs, C. H., and R. C. Kao, "On a Connection Between Factor Analysis and Multidimensional Unfolding," in *Psychological Scaling: Theory and Applications*, ed. by Gilliksen and Messick, Wiley, New York, 1960.

25. Alchian, A., "The Meaning of Utility Measurement," *Amer. Economic Rev.*, March 1953.

26. Becker, S. W., "Utility and Level of Aspiration," *Amer. Jour. of Psychology*, 1962.

27. Siegel, "Level of Aspiration and Decision Making," *Psychological Review*, vol. 64, 1957.

28. Simon, H. A., *Models of Man*, Wiley, New York, 1957.

29. Leibenstein, H., "Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand," *Quart. Jour. of Econ.*, May, 1950.

30. Wohlwill, J. F., "The Physical Environment: A Problem for a Psychology of Stimulation," *Jour. of Social Issues*, vol. XXII, no. 4, 1966.

31. Alexander, C., *Notes on the Synthesis of Form*, Harvard Univ. Press, Cambridge, 1964.

32. Milne, M. A., "CLSTR: A Structure Finding Algorithm," presented at first Design Methods Group Conf., Cambridge, June 1968.

33. Gall, D. A., "A Practical Multifactor Optimization Criterion," in *Recent Advances in Optimization Techniques*, ed. by Lavi and Vogl, Wiley, New York, 1966.

34. Horan, C. B., "Multidimensional Scaling: Combining Observations When Individuals Have Different Perceptual Structures," *Psychometrika*, vol. 34, no. 2, June 1969.

35. Wagner, H. M., *Principles of Operations Research*, Prentice-Hall, Englewood Cliffs, 1969.

36. Bhatt, K., *Fundamental Explorations in the Comparative Analysis of Transportation Technologies*, Ph.D. thesis, Dept. of Civil Eng., M. I. T., Cambridge, 1971.

REFERANCES (cont.)

Chapter IV:

1. Herrmann, C. R., and C. E. Ingram, *The Analytical Approach and Physics-of-Failure Technique for Large Solid Rocket Reliability*, TEMPO, General Electric, Santa Barbara, 1961.

2. Lloyd, D. K., and M. Lipow, *Reliability: Management, Methods, and Mathematics*, Prentice-Hall, Englewood Cliffs, 1962.

3. Leve, H. L., "A Reliability Framework for Structural Design," technical report, Douglas Aircraft Co., Santa Monica, 1964.

4. Feller, W., *An Introduction to Probability and its Applications*, Wiley, New York, 1968.

5. Bartlett, M. S., *An Introduction to Stochastic Processes*, Cambridge Univ. Press, Cambridge (England), 1966.

6. Doob, J. L., *Stochastic Processes*, Wiley, New York, 1953.

7. Drake, A., *Fundamentals of Applied Probability Theory*, McGraw-Hill, New York, 1967.

8. Alexander, J. A., *Highway Maintenance*, Ph.D. thesis, Dept. of Civil Engineering, M. I. T., Cambridge, 1970.

9. Wright, T. O., "Maintainability, the Measure of Availability," *1969 Annals of Assurance Sciences*, Gordon and Breach, New York, 1969.

Chapter V:

1. Isard, W., *Location and Space-Economy*, M. I. T. Press, Cambridge, 1956.

2. Lösch, A., *The Economics of Location*, Wiley, New York, 1952.

3. Thompson, W. R., *A Preface to Urban Economics*, Johns Hopkins Press, Baltimore, 1965.

4. Martin, B. V., and C. B. Warden, *Transportation Planning in Developing Countries*, Brookings Inst., Washington, 1965.

5. Coerman, R. R., "The Passive Dynamic Mechanical Proper-

REFERANCES (cont.)

ties of the Human Thorax-Abdomen System and the Whole Body System," Aerospace Medicine, vol. 31, 1960.

6. Canadian Good Roads Association, "Pavement Evaluation Studies in Canada," First Int. Conf. on Structural Design of Asphalt Pavement, Ann Arbor, 1962.

7. Amer. Assoc. of State Highway Officials, The AASHO Road Test, report 5, Highway Research Board, Washington, 1962.

8. Hutchinson, B. G., The Evaluation of Pavement Structural Performance, Ph.D. thesis, Waterloo Univ., Waterloo, 1965.

9. Holbrook, L. F., "Prediction of Subjective Response to Road Roughness by Use of the Rapid Travel Profilometer," Highway Research Record, no. 291, 1969.

10. Nakamura, V. F., Serviceability Rating of Highway Pavements, MSCE thesis, Purdue, Lafayette, 1962.

11. Yoder, E. J., and R. T. Milhous, Comparison of Different Methods of Measuring Pavement Condition, NCHRP report 7, Highway Research Board, Washington, 1964.

12. Phillips, M. B., and G. Swift, "A Comparison of Four Roughness Measuring Systems," Highway Research Record, no. 291, 1969.

13. Csathy, T. I., W. C. Burnett, and M. D. Armstrong, State of the Art of Skid Resistance Research, Special Report 95, Highway Research Board, Washington, 1968.

14. Keen, H. M., " Design for Safety," Highway Research Record, no. 214, 1968.

15. Horne, W. B., "Tire Hydroplaning and its Effects on Tire Traction," Highway Research Record, no. 214, 1968.

16. Beaton, J. L., E. Zube, and J. Skog, "Reduction of Accidents by Pavement Grooving," report no Mand R 633126, California DPW, Division of Highways, 1968.

17. Tamburri, T. N., and R. N. Smith, "The Safety Index," presented at HRB annual meeting, Washington, Jan. 1970.

18. Mills, J. A., "A Skid Resistance Study in Four Western States," Special report 101, HRB, Washington, 1969.

REFERANCES (cont.)

19. Stafford, J. H., et al., Highway Design Study, for Int. Bank for Reconstr. and Dev., CLM/Systems, Cambridge, 1970.

20. Winfrey, R., Economic Analysis for Highways, International Textbook, Scranton, 1969.

21. Hveem, F. N., "Types and Causes of Failure in Highway Pavements," Bulletin 187, HRB, Washington, 1958.

22. Alexander, J. A., Highway Maintenance, Ph.D. thesis, Dept. of Civil Eng., M. I. T., Cambridge, 1970.

23. Hudson, W. R., and B. F. McCullough, "An Extension of Rigid Pavement Design Methods," Highway Research Record, no. 60, Washington, 1964.

24. Maner, A. W., "Progress in Asphalt Thickness Design," Civil Engineering, March 1970.

25. Asphalt Inst., Thickness Design, manual MS-1, College Park, 1963.

26. Portland Cement Assoc., Thickness Design for Concrete Pavements, manual HB35, Chicago, 1966.

27. Manheim, M. L., et al., Search and Choice in Transport Systems Planning: Summary Report, report R68-40, Dept. of Civil Eng., M. I. T., Cambridge, 1968.

Chapter VI:

1. Kilbridge, M. D., et al., Urban Analysis, Grad. School of Business Admin., Harvard Univ., Boston, 1970.

2. Forrester, J. W., Urban Dynamics, M. I. T. Press, Cambridge, 1970.

3. Hirschman, A. O., Development Projects Observed, Brookings Inst., Washington, 1967.

Chapter VII:

1. Alexander, J. A., Highway Maintenance, Ph.D. thesis, Dept. of Civil Eng., M. I. T., Cambridge, 1970.

REFERANCES (cont.)

Appendix C:

1. Schodek, D. L., *A Methodology for Evaluating the Technical Performance of Housing Systems*, report R70-32, Dept. of Civil Eng., M. I. T., Cambridge, 1970.

2. Forrester, J. W., *Urban Dynamics*, M. I. T. Press, Cambridge, 1970.

3. Peters, P. D., *Residential Sectors and Urban Expansion*, M. R. P. thesis, Cornell, Ithaca, 1964.

4. Alexander, C., and M. L. Manheim, *HIDECS 2: A Computer Program for Hierarchical Decomposition of a Set with an Associated Linear Graph*, report R62-2, Dept. of Civil Eng., M. I. T., Cambridge, 1962.

5. Univ. of California, *Analysis of Refuse Collection*, Tech, Bulletin 8, Sanitary Engineering Research Project, Berkeley, 1952.

6. Grigsby, W., *Housing Markets and Public Policy*, Univ. of Penn. Press, Philadelphia, 1967.

7. Hester, J., *Systems Models of Urban Growth and Development*, M. I. T. Urban Systems Lab., Cambridge, 1969.

8. Kilbridge, M. D., et al., *Urban Analysis*, Grad. School of Business Admin., Harvard Univ., Boston, 1970.

BIOGRAPHICAL NOTE

Andrew Charles Lemer was born on December 25, 1944. He came to M.I.T., in 1963, after graduating from the Lovett School in Atlanta, Georgia, and received his S.B. and M.S. degrees in Civil Engineering in 1967 and 1968, respectively.

As part of his graduate work at M.I.T., Mr. Lemer has been involved in diversified research areas, including new communities planning, highway transportation in developing countries, and assessment of research needs for highway pavements. He has been involved with the planning and teaching of courses in highway technology, concrete technology, and engineering materials, as well as seminars on such topics as use of rubbish as a building material.

Mr. Lemer has had several publications, which are listed below:

> "An Integrated Approach to the Analysis and Design of Highway Pavement", Highway Research Record, No. 291, Highway Research Board, 1969.
>
> "Analysis of Highway Pavement Systems", Highway Research Record, No. 337, Highway Research Board, 1971.
>
> "Reliability of Highway Pavements", Presented at the 50th Annual Meeting, Highway Research Board, 1971, (in Press).

These papers were all coauthored by Professor F. Moavenzadeh.

APPENDIX A

HIERARCHIAL DECOMPOSITION*

*This approach is stated well by C. Alexander, in his Notes on the Synthesis of Form. The computer algorithm used was HIDECS 2 of Alexander and Manheim.

The computer has a tremendous capability to store and manipulate information. The growth of this capability has led people to devise ways of using the computer as an aid in understanding a problem and finding its structure. One such approach is through hierarchial decomposition.

One views a problem as a mass of interconnected factors or requirements. One can list all of the qualities which a solution should have, and can then state that each of these qualities is related to certain others. One can see intuitively that there will be clumps of qualities, within which individual qualities are more closely related to one another than to those in other clumps.

Clumps will in turn cluster together by closeness of qualities within. At one end of the hierarchy, there are individual listed qualities. At the other end, one sees the whole complex, inter-related cluster. Through decomposition of the problem (the large cluster) one hopes to learn something about its internal structure. Searching for a solution might then be a matter of proposing solutions for subsidiary clusters and building, up the hierarchy, to a complete solution.

Hierarchial decomposition with the computer is based upon the analysis of the structure of a linear graph. Such a graph has two elements: vertices and links connecting pairs of vertices. These links are non-directional. The process of decomposition consists of partitioning this graph, the set of vertices, into two or more subsets, each a graph of verticles and links. The partitioning is repeated successively, decomposing the original set into smaller and smaller subsets, until the limit of a complete decomposition into constituent vertices is reached. The result of this process is a tree diagram.

The computer program used for the decomposition uses algorithms of three types:

1. Criterion - The computation of the measure by which the value or "strength" of a partition is assessed.
2. Sampling - The selection of possible partitions to be evaluated.
3. Control - Storage of partition results; decisions about sequences to be followed in partitioning, when to stop, and printing of output.

The criterion for partitioning is based upon the representation of links as statements of statistical correlation between variables associated with the end point vertices. It is desired to obtain the partition which produces two subsets which have the least possible information transmitted across the partition. This is achieved by representing the links by random variables equal to 1, indicating linkage of the endpoints or 0, indicating no link. The probability of either possibility is 1/2, and a correlation coefficient is defined to describe the information transmitted between sets. The program samples possible partitions and computes this coefficient. The minimal coefficient defines the best partition.

The graph is represented by a matrix of the link variables. This matrix will be symmetrical. A random start, hill climbing type optimization search procedure is used to find an optimal partition. Thus there is a certain degree of randomness in the decomposition, and two successive decompositions of the same graph could give different tree structures.

As used in this investigation, successive runs were made and the various diagrams compared. Differences were seldom extraordinary. From these runs, a better understanding of the goals was derived, and goal fabrics such as those shown in the example problems were formulated.

APPENDIX B

EXTENSION OF THE ASPHALT INSTITUTE EQUATION

Based upon data gathered from various sources, the AASHO Road Test and others, the Asphalt Institute derived and presented an equation to be used in determining the design parameters for a given loading condition. This equation was

$$T_A = \frac{9.19 + 3.97 \text{ Log DTN}}{(\text{CBR})^{0.4}} \tag{1}$$

where,

T_A = total equivalent thickness of asphalt required (inches);

DTN = design traffic number, the daily equivalent 18 KIP loadings for a 20 year life;

CBR = California Bearing Ratio

In turn,

$$T_A = a_1D_1 + a_2D_2 + a_3D_3 \tag{2}$$

where,

a_1, a_2, a_3 = coefficients of substitution of other materials for good quality, type IV asphalt;

D_1, D_2, D_3 = depth of surface, base, and sub-base respectively

For standard materials, $a_1 = 1$, $a_2 = .5$, and $a_3 = .37$ were recommended values.

For the analysis to be undertaken, it was desireable to have a somewhat more general equation. This one was restricted to the traffic-counting assumptions and the defined initial and terminal rideability levels chosen by the Asphalt Institute as a basis for a design method.

The modification began with traffic. By definition:

$$\mathrm{DTN} = \frac{\Sigma W_{18}}{7300}$$

where ΣW_{18} was the total of equivalent 18 kip axle loads expected during the (assumed) 20 year life. Substitution gave

$$\mathrm{Log}\ \Sigma W_{18} = \mathrm{Log}\ 7300 + \frac{1}{3.97}\ [T_A\ (\mathrm{CBR})^{0.4} - 9.19] \qquad (3)$$

This equation predicts how many loads are required to reduce the rideability of a given pavement system (T_A and CBR) from an initial value of 4.5 to a terminal value of 2.5.

Reference to the AASHO Road Test Data and subsequent analyses showed that pavement deterioration tended to follow exponential curves. It was decided that this would not be a totally unreasonable assumption, so that it was stated that $S = S_o e^{-\alpha W}$.

At failure, $S = 2.5$, so $\alpha = (\mathrm{Ln}\ \frac{S_o}{2.5})\ \frac{1}{W_{2.5}}$. Over the narrow (relatively) range at which the approximation was desireable, a new failure level S_f would be caused by a different load application W_f, thus

$$S_f = S_o e^{-(\mathrm{Ln}\ \frac{S_o}{2.5})\ \frac{W_f}{W_{2.5}}}$$

Solving for the failure load W_f, one finds

$$W_f = W_{2.5} \left[\frac{\mathrm{Ln}\ S_o - \mathrm{Ln}\ S_f}{\mathrm{Ln}\ S_o - \mathrm{Ln}\ 2.5} \right]$$

The transition to common logarithms was made, and a factor defined.

$$\beta = \frac{\mathrm{Log}\ S_o - \mathrm{Log}\ S_f}{\mathrm{Log}\ S_o - 0.40}$$

Then

$$W_f = \beta\ W_{2.5}$$

$W_{2.5}$ is now the same as ΣW_{18} from equation (3). So, by substitution

$$\mathrm{Log}\ W_f = \frac{1}{3.97} [15.32 + T_A (\mathrm{CBR})^{0.4} - 9.19] + \mathrm{Log}\ \beta$$

Thus, the final form used in reliability analysis was given as

$$\mathrm{Log}\ W_f = \frac{1}{3.97} [6.13 + T_A (\mathrm{CBR})^{0.4}] + \mathrm{Log}\ \beta$$

with

$$\beta = \frac{\mathrm{Log}\ S_o - \mathrm{Log}\ S_f}{\mathrm{Log}\ S_o - 0.40}$$

and

$$T_A = a_1\ D_1 + a_2\ D_2 + a_3\ D_3,$$

with all terms previously defined.

APPENDIX C

URBAN HOUSING

The area of urban housing has been for a number of years recognized as beset by major problems. Present housing is judged to be inadequate in quantity and quality for a substantial portion of the population. While much research has been done in what is needed and how it might be achieved, relatively little result is observable. This situation would appear to stem from the uncoordinated planning of much of the research and from a failure to recognize for whom the housing is intended.

To fully implement a generally applicable analytical model for urban housing, using the approach described in this thesis, will involve a major effort. It would seem that in many cases the data required is simply not available, and must thus be gathered as part of the work. The few attempts at any sort of a comprehensive view have had to simplify the situation a good bit to make any progress. Schodek (1), for example, while looking at the broad range of factors describing service, was forced to use the idea of single number absolute failure levels, as is done in current building codes. Within the present framework, these statements must be viewed as extensions of the analyst's own individual aspiration levels. They give no recognition to the normally expected variations in response within a population of users, or to the role these variations play in the social, political, and economical systems which housing serves.

The purpose of this section is to present some preliminary work which has been done, and to suggest how this work might be extended to produce a fully implemented analytic tool for urban housing. Work of people such as Forrester (2) and Peters (3) are examples of the many views of the role of housing within the total urban system, and suggest the potential for benefits deriving from a fully implemented integrated and rational analysis of urban housing.

A. A Serviceability Function

The basic nature of the service provided by housing lies in providing shelter from a sometimes hostile environment. This concept of service may be expanded in terms of environmental qualities which the user experiences and of the structural characteristics associated with the integrity of the system. Figure 1 illustrates the possible form of this expansion.

Figure 1, which suggests components of a housing serviceability function, was derived using the hierarchical decomposition methods of Alexander and Manheim (4). (See Appendix A). A fairly extensive review of literature was made to yield a list of statements of qualities good housing should have. This list was broken down, and the breakdown used to suggest the components displayed in Figure 1.

There are then proposed to be two major classifications of the components of serviceability. Structural considerations, including the foundations, superstructure, and mechanical system, are the traditional concern of the engineer.

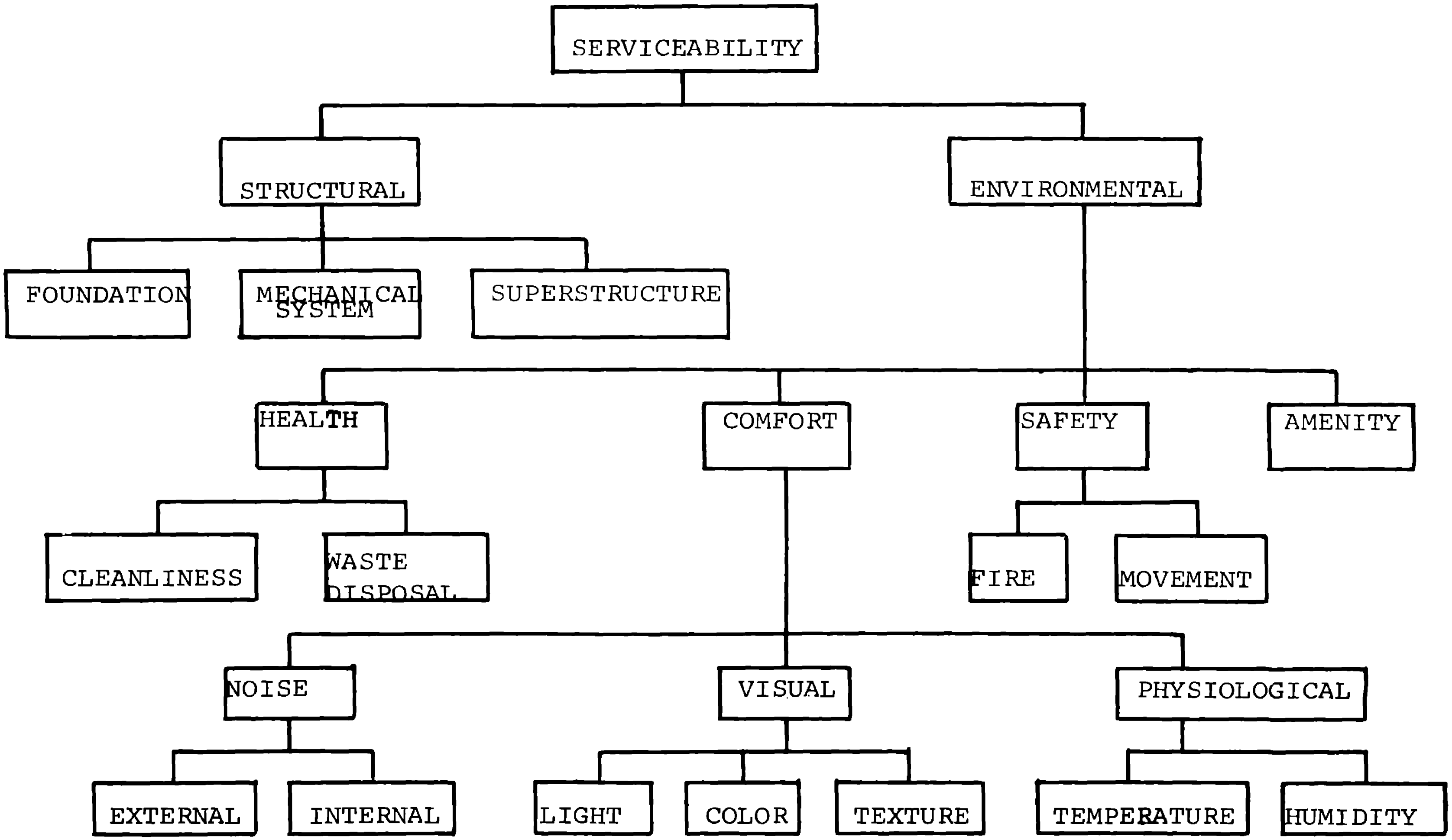


Figure C-1 : Serviceability for urban housing

These factors are often not readily apparent to the user, but are clearly dominant over the other part of a serviceability measure, environmental factors. A structural failure will represent a serious loss of safety and security.

Environmental qualities are those which have the most apparent impact upon the direct user; they are the qualities of his surroundings. There are four components under this classification: health, safety, comfort and amenity.

Health includes the obvious problems of waste disposal. It is estimated that roughly four pounds of garbage per capita are produced in the U.S. (5). Subsystems to collect and eliminate waste materials must be provided.

In addition to waste, there is an overall problem of cleanliness or healthfulness. For example, pipes should not poison the water. Lead paints are an example of a material which would rate relatively low on this component of serviceability.

Safety is defined to cover the hazards of fire losses and danger in circulation. Fire hazards are well known and are considered explicitly in current codes and design methods. Less familiar are the effects on safety, measured perhaps by the likelihood of accidents -- of narrow halls, stairs, poor lighting, etc. Such factors need to be made explicit.

Comfort covers a broad range of factors which lie somewhere between socially approved necessities and the extras which money can buy. The noise level within the house will depend on externally generated sounds -- traffic and aircraft, for example -- and upon internal sources -- children, machinery. Insulation requirements depend directly upon these sources and the availability of noise transmission pathes.

Visual "comfort" is a measure of how well lighting and overall visual aspects meet utilitarian and psychological needs. The interaction of light (intensity), color, and texture will have to be different in the kitchen from what it is in the bedroom, in order to be equally satisfactory. Texture relates to the quality of lighting (glare, for example) and to surface qualities of the environment (wood grains for walls versus concrete blocks).

The term physiological is intended to refer to the effects on comfort resulting from the interaction of temperature and humidity - higher temperature is much more easily tolerated if relative humidity is at a low level. Response to this component will be a function of the type of activity in an area. Hard work or exercise necessitate lower levels of temperature and humidity.

The final component, amenity, is something of a catch-all term. It includes, for example, the relative desirability of real wood versus wood-grained plastic wall panels, or single picture windows versus an equal amount of glass in smaller panes, or the presence of a useable fireplace. These are the things which are not based primarily on need, but which make the difference in level of service perceived by a user. At present, the best approach to this component that can be suggested here is one based on simple orderings of particular alternatives for their relative levels of amenity. The fact that lending institutions can make judgements about a developer's provisions for amenity, and the subsequent effect on marketability of a development, indicates that there is something here which might be evaluated more explicitly.

Figure 1 is a proposed representation of a serviceability measure, in terms of its component subscales. The next step is, of course, to select suitable indicants of response on each of these subscales and to obtain a functional relation between serviceability and these indicants. In some cases, data is already available for these relations. Table 1 suggests some possible indicants and a judgement as to the immediate possibility of developing the desired function from available data. Items marked with a questionmark are felt to be quite likely to require substantial research before a good understanding can be achieved.

B. Reliability, Maintainability, and Service Life

The idea of service life for housing is complicated by the separation of user groups in a market environment. This separation is often observed in terms of income, but will include more subtle differentation such as social status and racial factors. The result of the separation is that as the level of service deteriorates, direct usage of the

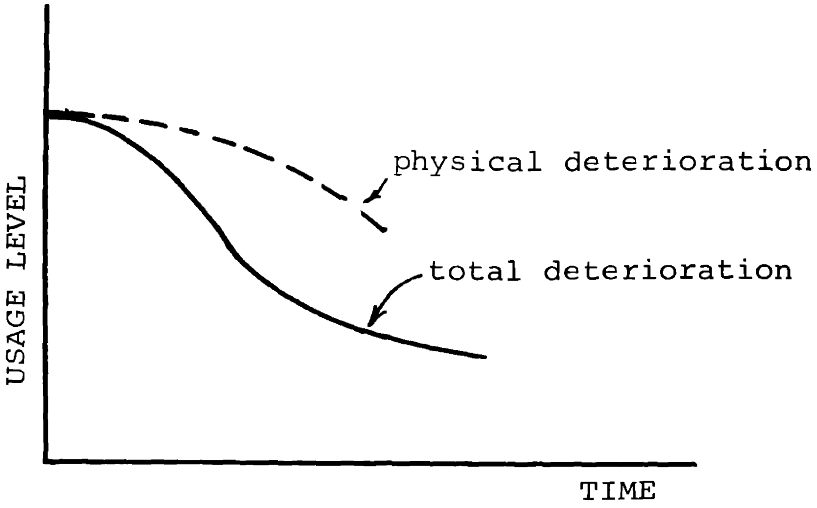


Figure C -2 : Housing depreciation curves

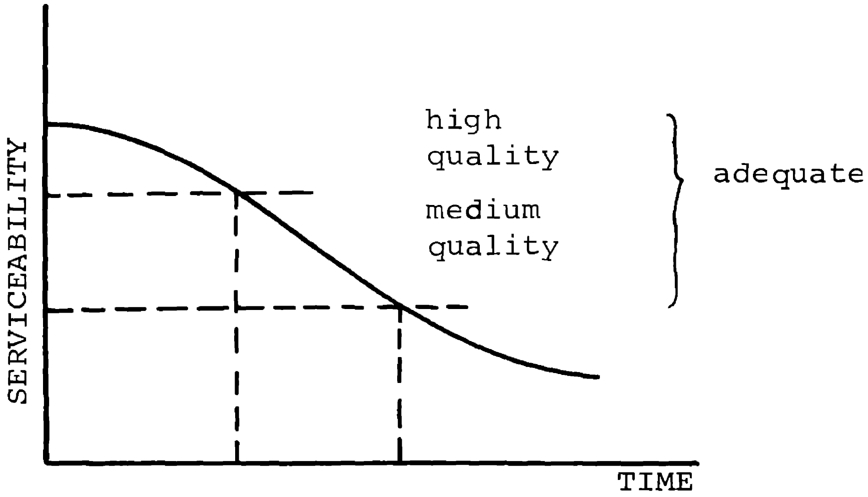


Figure C -3 : The filtering process as an approach to state definition

SERVICEABILITY COMPONENTS	SUGGESTED INDICANTS	AVAIL. OF MEASURE
Structural		
Foundation	differential deflection, average gross deflection	✓
Mechanical system	allowable load	✓
Superstructure	allowable floor load, maximum deflection	✓
Environmental		
Health: Cleanliness	concentration of pollutants	?
Waste disposal	capacity	✓
Comfort: Noise	interior NC rating	✓
Visual	intensity, color temperature	?
Physiological	effective temperature	✓
Safety: Fire	fire rating	✓
Movement	relative probability of accidents	✓
Amenity	status factors, relative desireability	?

(ref.)

Table C-1 : Possibilities for measurement of serviceability in urban housing of the dwelling unit may pass from one direct user group to another, with the new group possibily finding the unit as satisfactory as the previous group did. It may thus be suggested that there are several service life estimates to be made.

The phenomenon in question is referred to as filtering (6). Filtering is the process by which housing ages and changes its level of use. The process is controlled only in part by the physical aging of the constructed facility.

In filtering, a housing unit which starts its service life as high quality, high income housing will, with time, lose some attractiveness. It will depreciate to become middle income housing. With additional time, the housing moves to lower income and perhaps to slum conditions. In typical urban settings each of the three levels spans a time period of 20 to 50 years, giving housing a total life of on the order of 100 years.

Grigsby (6) discusses the process through use of a hypothetical housing depreciation curve (see Figure 2). The total depreciation of the unit is minimal in early years of life. Any decrease in usage level is due primarily to changes with time in heighborhood and users' subjective values. In very old units, the total depreciation rate declines as the potential demand rises (more families can afford the lowered rate), intensified by a relative scarcity of usable low cost housing. This trend often occurs in spite of an acceleration in physical deterioration which is observed.

Grigsby's description may be rephrased on the basis of serviceability, in that losses of serviceability on particular component scales will imply dissatisfaction among a particular, identifiable faction of the potential users. It may further be suggested that this effect is only indirectly related to cost: i.e., a decrease in rent will not persuade the high-level user to be satisfied. Then, lifetime becomes a concept related to rent level, which may be changed. There might be multiple "failure" levels, as suggested in Figure 3, associated with expected rent reductions. The lowest level of failure might be thoroughly unacceptable on a basis of total social welfare in the city.

The definition of usage levels and failure levels presents questions which will require some careful thought. The modeling of lifetime behavior and computation of reliability and maintainability may be carried out in a multi-stage manner, handled separately for each defined level of usage. It might be found desirable to design houses like autos, to be discarded after some particular average service life. These and similar problems must be faced in modeling the lifetime behavior of the system.

A beginning attempt at such modeling might be made by starting with the descriptions of condition used by the U.S. Bureau of Census, wherein housing is described as adequate, substandard, or dilapidated. Substandard means that a dwelling unit is in need of substantial repairs, while dilapidated units exhibit structural failure or lack of inside plumbing. The problem with these three states is that substandard and dilapidated both are quite likely to be failure states (as suggested in Figure 3), although the dilapidated state does indicate a dominance of structural consideration over environmental. Hence, it would probably be desirable to expand the definition of adequate to suggest several states of service.

It would seem likely that the model would best be built as a set of integrated submodels. One group of submodels would include failure modes resulting from factors which are relatively independent of facility usage or wear and tear; for example, safety with respect to fire. The second group would include failure modes which are likely to depend upon operation and maintenance activities, such as visual comfort. (For example, it is expected that light bulbs will require replacement).

It may in general be suggested that given the current technologies of housing, the overall maintainability of a unit must be low, relative to other types of constructed facilities. This suggestion stems from the combination of long service life and heavy usage. Mechanical units will need repair; walls will need repainting; roofs will need patching. A minimum maintenance system becomes expensive.

On the other hand, given dependable maintenance policies, there is no apparent reason why reliability should not be quite high. Preventive maintenance is not a new idea in this field, and could contribute admirably to overall system performance.

C. Search and Selection

In searching for alternatives for solution of a particular problem, it may be desirable to extend the previously alluded to idea of dominance among serviceability subscales, to assist in the first stage of satisficing. It has been suggested that structural subscales are dominant over environmental subscales. Hence, one might begin search by choosing a promising structural system. (See Figure 4).

Safety and health may be viewed as occupying the second level of hierarchy. The argument for this view is based upon the idea of social consensus. That is, it is generally recognized that health and safety are important, and must be provided.

At the lowest level, then, are comfort and amenity. These scales are most apparent to the direct user, and include those factors which are often compared with rent costs by the user in making his selection. Further, the components of the physical system which most directly influence these scales are items such as wall and floor coverings, insulation material, etc., which are in effect attached onto a finished structural-service system.

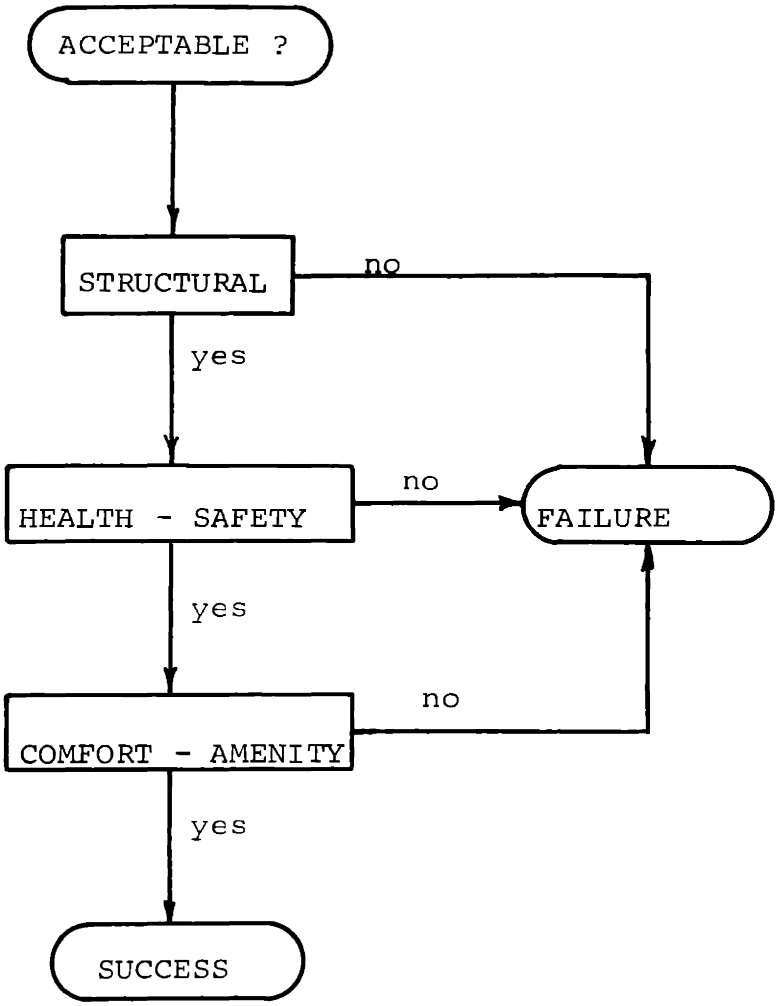


Figure C -4 : Hierarchy of the serviceability measure

Reaching the "success" block in Figure 4 is then the completion of a solution proposal. The output of the search step will then be a satisfactory housing possibility.

The selection decision is complicated by the multiple lifetimes or failure-usage levels of the housing system. The decision maker must consider all stages of use within the context of the total city. Experiments with models such as Forrester's urban dynamics construct (2) suggest the role which varying the relative lengths of usage periods can play. For example, planning housing to have a shorter life can, when linked with a program of regular demolition and reconstruction of inadequate units, increase the city's tax base by encouraging new industry and increasing employment. Of course, this is only suggestive. Forrester's work has been a subject of some controversy (for example, see Hester (7)). Obviously, though, decision will in this broader context be of some importance.

D. Conclusion

As was stated initially, this Appendix is intended primarily to suggest the issues to be faced in the analysis of urban housing, and to present some preliminary work done in this direction. It is felt that the area is quite important, and in need of work. Other aspects of urban analysis are being pursued (see Kilbridge, et al. (8)) and indicate the need for and place of this analysis of the physical system.