A STUDY OF METHODS OF ANALYSING AND REPRESENTING SUBJECTIVE REACTIONS TO TELETEXT SYSTEMS, SERVICES AND CONTENT

Stephen J. Lupker, Brian R. Shelton

and

Philip A. Vernon

Department of Psychology
The University of Western Ontario
London, Ontario
Canada

Report from contract 36001-3-0505
Department of Communications
Scientific Authority: Paul Hearty
March 1, 1985

A STUDY OF METHODS OF ANALYSING AND REPRESENTING SUBJECTIVE REACTIONS TO TELETEXT SYSTEMS, SERVICES AND CONTENT

Stephen J. Lupker, Brian R. Shelton

and

Philip A. Vernon

Department of Psychology
The University of Western Ontario
London, Ontario
Canada

Report from contract 36001-3-0505
Department of Communications
Scientific Authority: Paul Hearty
March 1, 1985

EXECUTIVE SUMMARY

In the present manuscript, we have attempted to do three things: discuss the issues involved in selecting a scaling procedure for teletext systems, recommend and discuss the most viable of the procedures currently available, and present a psychological model of the processes involved in making a scaling judgment. In the first chapter we accomplish the first of those goals. Broad classes of scaling techniques are discussed and it is ultimately concluded that indirect scaling is most appropriate in the present circumstance.

In chapters 2, 3 and 4 we meet our second goal by suggesting a) the optimal way to collect scaling data, b) the best techniques for analyzing those data to produce scale values and c) the most appropriate way of analyzing the resultant scale values. The recommendation is that since the industry has adopted a five-point rating scale for data collection, a scale which also suits our purposes, that we also use a five-point scale ranging from bad to excellent. The best techniques for analyzing those data are Thurstonian (1927) scaling and Allnatt's (1973; 1975; 1979) more recently developed method. Detailed descriptions of how to

carry out each of these techniques, and how to test the assumptions each technique makes, are included. The analysis of variance is felt to be the optimal tool for analyzing the resultant scale values although data transformations or even nonparametric alternatives may be necessary.

Our final goal is achieved in chapter five. Here a model of the entire scaling process is presented and discussed. As is argued, the model has a number of strengths including its generality and testability. Further, it represents an integration of a number of areas of psychological research and, thus, its principles have strong empirical backing. Hopefully, it can serve as a useful framework for understanding the nature of subjective scaling judgments.

TABLE OF CONTENTS

CHAPTER 1 - SCALING ISSUES

The assessment of subjective reactions has had a very long history. The first recorded attempt to describe subjective responses was the scale of stellar magnitudes used by astronomers in 150 BC to measure the perception of the stars (Stevens, 1960). In the mid-1800's psychologists began to study the problem formally, and, since that time, various "schools of thought" have developed. There are large areas of disagreement between experts, and it would be impossible to make recommendations compatable with all theoretical approaches.

In terms of the subjective assessments of teletext systems, there are constraints on the type of procedures which can be used. A review of the procedures indicates that a specific type of assessment, indirect scaling, is the most appropriate approach to apply to teletext evaluations. Procedures will be recommended for developing a theoretical representation of the assessment process and the perceptually important aspects of teletext displays. In subsequent chapters, specific methods will be examined from both statistical and procedural points of view.

I. PSYCHOLOGICAL MEASUREMENT

Measurement is the process of assigning numbers to events or objects. In physics, measurement involves comparing the object to be measured to some calibrated standard, such as a yard stick or balance. The result is a numerical representation which has a clearly defined meaning by virtue of the known characteristics of the measuring device.

The goal of psychophysics is to measure mental events, and the observer is the measuring device. Each judgment the observer makes can be considered a reading from a mental "yard stick". The mental scale is not an objectively calibrated device. Assumptions must be made about the way in which subjective assessments are made. Without these assumptions the subjective assessment has no meaning, just as it has no meaning to describe a physical object as "5" without specifying that we are talking about length and that the units are centimeters. Any substantive conclusions based on subjective responses necessarily imply a theory of the subjective measurement process, whether the assumptions are made explicit or not.

All psychophysical procedures make the assumption that observers are able to follow the instructions they are given to perform the task. The instructions define, at least in part, the nature of the measurement scale. Sometimes there is little doubt that the instructions can be followed. For instance, an observer may be asked to judge which of two objects is larger, more colourful, more interesting, prettier or sexier. The task is quite simple, to make a choice, and the concept (and therefore the measurement) is defined by the observer. In other procedures the observer performs a more difficult task, such as to assign a number to an object which represents its ratio to some known standard. The observer is expected to choose a number twice the value assigned to the standard if the test stimulus is perceived as twice as bright, interesting, attractive, or whatever attribute is specified in the instructions. Here, the observer must define the concept and perform a reasonably difficult task.

The basic psychophysical assumptions are that the observer can:

1) isolate the attribute of interest, and
2) perform the requested judgment.

The extent to which a particular procedure can be expected to meet these requirements determines its viability as a psychophysical technique.

1.1 Categories of Psychophysical Methods

1.1.1 Objective Techniques

There are a number of procedures which can be applied when judgments can be classified as either correct or incorrect. The questions these techniques are used to answer involve the detection of stimuli or the discrimination between two or more alternatives. The techniques completely solve the basic problems of psychophysical measurement, and are included as an exemplary baseline for the adequate measurement of subjective responses.

The basic procedure is to present the observers with stimuli, either in isolation or in sets, and require them to identify the stimulus which has the attribute in question. The observer's choice can be scored as correct or incorrect, and the measure is defined in terms of the accuracy of responses. For example, to determine an observer's sensitivity to acoustic intensity, two sounds differing only in intensity could be presented. The observer would be required to identify the more intense of the two. The observer's ability to perform the task is explicitly determined in the comparison of accuracy to chance performance. If accuracy is greater than chance, the observer can perform the task. The problem of the definition of the attribute used for the judgment is solved by the selection of stimuli such that they only vary in terms of the attribute of interest. If two stimuli differ only on a single attribute and the observers can accurately discriminate the stimuli, then the judgment must be based on that attribute.

Objective procedures require the control of the attribute of interest, while keeping other attributes constant. This is often difficult, even with apparently well-defined stimuli. For instance, the duration of an auditory stimulus can easily be controlled, and the detectability of a change in duration can easily be measured. However, it is difficult to determine whether the discrimination is based on changes in duration, signal energy, the short-term energy spectrum, or some combination of these various cues. Great care and effort must be taken to ensure that stimuli vary only in terms of the attribute of interest.

With respect to the problem of the subjective quality of teletext displays, an observer's response cannot be scored as correct or incorrect. There is no objective definition of subjective quality against which to judge the accuracy of a response. For this reason, objective techniques are inappropriate for the assessment of subjective quality. It is strongly recommended, however, that if situations arise at any point in this project which are amenable to objective techniques, then these techniques should be adopted immediately. These procedures provide the most satisfactory solution to the basic problems of psychophysical measurement, and the natural inertia to alter experimental paradigms should be avoided if objective techniques become applicable.

1.1.2 Subjective Techniques

A subjective psychophysical procedure is one which involves responses which cannot be classified as correct or incorrect. In subjective techniques, the observers are reporting, as best they can, the psychological impact of a stimulus. Since there is no physical device which can measure psychological impact, a response cannot be shown to be incorrect. Objective descriptions of the stimulus can only be used to characterize the stimulus, but not to define accuracy.

In subjective techniques, the observer is presented with stimuli, either individually or in sets, and asked to deliver a descriptive judgment. The instructions define the criterion for the response, but the observer's interpretation ultimately defines the measurement. A judgment might be required to assess the acceptability, attractiveness, suitability, or whatever attribute is requested, of the stimulus in question. The examples were chosen to emphasize a potential problem: the attributes are different because of the nuances of language, but it is not at all clear that different observers will employ the same precise definition. This problem is usually dealt with by employing very broad categories for the attribute being assessed, and not attempting to make fine discriminations of attributes. The observer is sometimes given a scenario to orient the task, so that all the observers will approach the problem from a similar perspective. For instance, an observer might be told to rate a number of teletext frames as Bad, Poor, Fair, Good, Excellent, and told to think of them as television images which might be transmitted into their home.

The issue of adequately defining the attribute being measured can also be addressed by requiring a degree of consistency in responses, both between and within subjects. Unfortunately, this requirement provides only a minimal indication that the attribute is being clearly defined. Consistency only provides assurances that the response is based on some stable stimulus characteristic. It provides no evidence that the judgment was made solely on the attribute of interest. Another approach is to examine the judgments to make sure they make "sense", seem to provide judgments which order stimuli in a reasonable way, or correlate with scores on a related task. Although this seems defensible, it is a circular procedure since the proper subjective ranking of stimuli on the attribute in question was the original purpose of the measurement. Thus, these sorts of procedures can only provide an assessment of the face validity of the responses. Overall, the problem of the definition of the attribute is dealt with by the assumption that the observers can follow the instructions, with a few minimal forms of verification.

The second issue involves the ability of the observer to perform the task, once the attribute of interest is defined. From the perspective of subjective assessments, this issue is assessed by the success achieved in providing numbers which accurately reflect the observer's opinions. This point is judged by the degree to which meaningful scale values can be recovered from the data. Unfortunately, there are large differences of opinion as to what constitutes a meaningful psychological scale. Since the true scale is an unknown quantity, attempts are made to verify the accuracy of the values by examining the statistical properties of the results. Therefore, it is important to deal with this issue in some manner, because it addresses the way in which values are assigned to objects, and thus, the meaning of the measurement.

1.1.2.1 Operational Definitions

One approach is the pragmatic approach of using an operational definition of the scale values. The measurement takes its meaning solely from the instructions and the procedure, and the main goal is to develop a standardized testing situation. The observer is viewed as a "black box" which is confronted with a standard stimulus situation, and produces a standard measurement. No reference is made to the processes by which judgments are made, and the psychophysical scale reduces to a summary of the input-output relations between the stimulus and response. Judgments are taken at their face value, and the adequacy of the measurement is merely the utility of the results. The consistency of observers on the task, and the success of the scale in discriminating between the stimuli, are indicative of an adequate measurement under this criterion.

1.1.2.2 Direct Scaling

Related to the above logic in simplicity, but not quite so atheoretical, are the so-called direct psychophysical procedures. Here, it is assumed that observers can make complex judgments, and are able to report the values of psychological events directly. In particular, observers are asked to make ratio judgments about the attributes of stimuli. The observer simply assigns a number to a stimulus which represents the ratio of the attribute in question to that assigned to a known standard. Procedures of magnitude estimation, magnitude production, and fractionation are of this type.

The assumption being made here is that these techniques define ratio values, so that the obtained judgments represent a direct estimate of the underlying psychological scale. Investigation of a number of rather simple relations, such as that between acoustic power and loudness, luminance and brightness, electrical current and perceived sensation to name a few, have all indicated an encouraging amount of consistency. All such functions can be adequately described as power functions, such that the perceived magnitude is related to the physical magnitude raised to some power. The value of the exponent is unique to the relation being assessed. An example is the sone scale of loudness, where auditory signals of various frequencies and intensities are scaled relative to a 1000 Hz tone presented at 40 dB SL. A scale such as this might be constructed for picture quality, by having observers make magnitude estimates of quality to a standard presentation. All judgments could be refered to a standard, with stimuli judged to be at equal ratios to the standard being regarded as equivalent on the dimension.

Although this procedure may seem attractive, there is a serious problem in the validation of the scale. The technique of magnitude estimation has been argued to apply to continua of quantity, the so-called prothetic continua. Ratio scaling is a valid approach in these instances because there is a true zero, that is, there exist some stimuli that contain no amount of the attribute. The existence of a true zero is required for ratios to be meaningful. If the psychological representation instead has an arbitrary zero, then the value assigned to the standard by the observer is likewise arbitrary, and so too will be the ratio steps on the psychological scale. Attempts to use direct scaling procedures with attributes which have no clear zero have resulted in largely unsatisfactory results. Under some conditions, consistent subjective scales can be obtained, but the resultant scales do not seem descriptive to competent observers (Ward, 1970; Marks, 1974). This problem makes the use of magnitude estimation scaling of a given stimulus to a known standard a risky procedure, if no validation can be provided that the judgment represents a ratio value on a prothetic dimension.

If the scaling of picture image quality is thought to be prothetic, then ratio scaling would certainly be possible, but the issue of validation is complicated because there may be no objective scale of image quality. One approach would be to measure the growth of subjective image quality with increases in the magnitude of various parameters related to overall image quality. Each type of teletext system to be evaluated may, of course, vary on more than one of the parameters of objective quality. Thus, interactions of the parameters must also be assessed to allow a prediction of the overall subjective quality of the system. An examination of the relation between the subjective magnitudes and each of the physical parameters of quality would provide validation of ratio scaling of the judgments, if the plots can be described as power functions. The assessment of interactions between individual parameters in the determination of overall image quality could define the relative salience of each physical parameter as a determinant of overall subjective quality. This result would be a useful one, because it would define the parameters that subjective quality is most sensitive to. However, other procedures which do not make the assumptions of direct judgments can be used to obtain similar sorts of information. The effort required in validating the use of magnitude estimation may turn out to be excessive, especially since the assessment of each single parameter is only one component of the overall subjective quality.

A potentially more acceptable application of direct scaling would be to develop a physical scale of image quality which accurately reflects the sense of objective quality intended for a given application. For instance, if the concern was the transmission of static frames, a measure could be developed which describes the correlation between the transmitted and displayed message. Each teletext option could be quantified with respect to this measure, and the relation between this objective metric and subjective quality could be assessed using magnitude estimation techniques. Other senses of the term "picture quality" would require new physical descriptions of the stimulus to quantify the new meaning of the term in objective terms. The success of the scaling could be assessed, once again, by the success of a power function in describing the relation between subjective and objective quality measures. An approach such as this would depend critically on the success of the physical measure in capturing the essential characteristics of the physical "picture quality".

Although direct scaling techniques have been enormously successful in some areas of subjective assessment, their application to the problem at hand is not a simple matter. This type of approach is possible, but not without either making some risky assumptions, or providing some form of scale validation. A central problem is that direct scaling requires a meaningful quantification of the physical stimulus such that the subjective judgment of interest can be represented on a ratio scale relative to the objective measure. For these reasons, ratio scaling techniques in general are not obvious solutions to the problem at hand.

1.1.2.3 Indirect Scaling

The final category of subjective measurement techniques are those described as indirect scaling. In these methods, there is no need for an objective scale of the attribute in question, because it is not necessary that the procedures relate judgments to any scale of physical magnitude. Rather, the attempt can be made to relate the stimuli to one another, in a psychological representation which is consistent with the observed set of judgments.

All indirect scaling procedures make an explicit set of assumptions about the way in which observers make responses. That is, they make assumptions about the form of the psychological representation of stimuli, and about the transform between the representation and the judgment in question. Since the subjective responses are known and the transform is assumed, the original psychological representation can be defined.

The adequacy of the assumed representation in accounting for the observed set of responses can be evaluated empirically. Given the obtained subjective representation, the measurement model can be used to predict the pattern of subjective responses, based on the solution and the assumed transform. These predicted subjective judgments can be compared to the original data set, and the adequacy of the model in accounting for the observations can be directly tested. Poor models can be discarded, and alternate representations can be assessed.

Given a reasonable fit between the model and the data, there is very little to lose in using indirect scaling techniques. At the very worst, the derived scale values can be taken as transforms of an operationally defined measurement. Since the operational definition of measurement makes no claim of being an optimal representation, it makes little difference whether they are transformed or not. At best, the derived scale values will be an accurate representation of the true subjective values. What is actually obtained is probably a compromise between the worst and best case.

The decision essentially boils down to whether or not it is useful to transform the raw data to estimates of the psychological representation. If the goal of the measurement is merely to test some hypothesis in a single experiment, then the transformation is probably not worth the effort. In terms of a large scale project the exercise is probably useful. It allows the assignment of a numerical value to a stimulus which describes the psychological value of the stimulus on some subjective attribute. The meaning of the measurement is defined by the model used to describe the judgment process, in a manner analogous to the way in which physical measurements derive their meaning. This is a clear advantage to the operational definition approach, where the issue of the assignment of values to events is ignored.

1.1.3 Procedural Conclusions

The major conclusion is that indirect scaling provides the most plausible solution to the measurement of subjective reactions to teletext systems. This follows from 1) the fact that no clear objective measurement is available to define the physical stimulus, which excludes objective psychophysics and complicates direct scaling procedures, and 2) the assertion that the mere operationalization of the measurement protocol does not address the basic issue of the accurate representation of subjective quality. Indirect scaling procedures, on the other hand, focus on the process by which judgments are made, and require no objective description of stimuli. These attributes make such procedures most applicable to the current problem.

The great difficulty in deriving psychophysical scales stems from the fact that when purely subjective attributes are being dealt with, there can be no objective assessment of the accuracy of the scale. As a consequence, all subjective scales derive support from demonstrations of the utility of the derived measure.

II. GENERAL INTRODUCTION TO INDIRECT SCALING

The consideration of the categories of psychophysical assessments has indicated that indirect scaling procedures provide the most promising approach for the subjective evaluation of teletext systems. The argument is based on a very general consideration of the options available in the psychophysical procedures applicable to subjective reactions. The decision to use an indirect scaling approach does not specify a single procedure because there are a number of methods which are included under this general categorization.

The selection of a particular procedure should be based on a number of criteria. One is the ease of measurement, that is, the effort required to collect the raw data. Another consideration is the form of the resultant psychological representation, and the utility of the results obtained to the solution of the problem at hand. Finally, the statistical properties of the procedures and the adequacy of the method from a measurement point of view is of vital concern. The latter consideration will be addressed in great detail in subsequent chapters. At that time, specific indirect scaling procedures will be described and evaluated from a statistical point of view.

The balance of this initial chapter will be concerned with indirect scaling from a general perspective, and will provide comments relevant to the first two considerations. The purpose of this discussion is to provide a background for the more detailed analysis, and to discuss the options available from a global perspective. This analysis will attempt to clarify the goals of indirect scaling methods, and will indicate how each could be applied to the problem of subjective assessment of teletext systems.

2.1 Data Collection Procedures

The raw data required for indirect scaling techniques are obtained by having observers make judgments about the relation between stimuli. The main goal of indirect scaling is to describe the position of specific stimuli or events in their psychological co-ordinates. The scale is derived by a consideration of the relation between the observed psychological responses, without reference to physical attributes. It is not surprising, then, that the main form of the raw data must be an estimation of the psychological relation between the stimuli presented to the observer.

There are many nuances in the specific procedures used, but virtually all can be considered to be a form of one of three major categories of tasks: category judgments, subjective rankings, or direct subjective comparisons. The specific instructions change on the basis of the attribute being considered and the type of scaling being employed, but the presentation of stimuli and the form of the observer's task can be reasonably summarized in this way. In constucting this categorization scheme a thorough review of the literature was undertaken. (See the additional Reference section for a list of those papers reviewed but not cited in the text.) All of the papers reviewed which provide data amenable to indirect scaling used procedures that can be categorized according to this scheme.

2.1.1 Category Judgments

In this type of procedure, observers are asked to place each stimulus presented into a descriptive category. The categories can be defined by the instructions, or in some rare cases, the categories appropriate for the stimuli are chosen by the observer.

The most common procedure is category ranking. The observer is presented with an *n*-point scale, and asked to assign a value to each stimulus which represents its position on the scale. The verbal description of the scale defines the attribute in question, and the meaning of various scale values are often specified. The categories are often described numerically, but verbal descriptions are quite common. Most often there is an attempt to give some absolute values on the scale and anchors are provided for the judgments, but in other cases the observer is given very little information about the intended meaning of the scale values. In these cases, the observer's interpretation is a major determinant of the meaning of the scale.

One variant of this procedure is the sorting task, where the observer is given a number of stimuli and asked to sort them into a number of groupings. Once again, the groupings are sometimes clearly defined, but at other times the group characteristics are left up to the observer.

Another variation is the so-called analogue scale, where observers are given a dial, slide potentiometer, or a continuous scale of some sort and asked to indicate a reading on the scale which represents the amount of the attribute contained by each stimulus. The scale reading is taken as the response, which can be considered a special case of category judgment, where there are a large number of cateqories. The number of categories is determined by the precision of the device used to subdivide the scale.

2.1.2 Subjective Rankings

A second approach is to present the observers with all the stimuli at one time, and have them rank order the entire set on the basis of some attribute. The procedure is usually applied by not allowing tied ranks, but forcing the observer to choose a specific ordinal ranking. The attribute chosen for the ranking is defined by the verbal description provided to the observer. In some instances, the observer is given a subset of the entire set rather than the complete collection, or given a number of subsets and asked to rank each of the smaller groupings. If a subset procedure is employed, then the subsets are usually chosen to contain overlapping elements to allow an estimate of the overall ranking of the complete set of stimuli from the ranking of the subsets.

2.1.3 Direct Subjective Comparisons

In this sort of procedure, an observer is presented with two or more stimuli and asked to choose one stimulus over the others according to some criterion. In the method of paired comparisons, for example, every possible pairing of two stimuli from a set are presented to the observer, and a choice is made on every pair. Likewise, the method of triads asks the observer to choose which of three stimuli is the most dissimilar to the other two on the basis of some attribute. In its most complete form, this type of procedure requires that each stimulus from the set be presented with each other member or combination of members from the entire stimulus set, which may often be a prohibitive requirement. For instance, the choice of the most attractive of two stimuli from a twenty-element set presented in pairs would require 190 choices to be made, and the same judgment with stimuli presented three at a time would necessitate 1140 judgments. In some instances, therefore, the complete set of choices are not sampled.

2.2 Data Analysis Schemes

All three procedures provide data which can be treated in a simple pragmatic manner, that is, as simple operationally defined measurements. The form of category judgment data is a frequency distribution of category choices, which can be used to define the average category rating, be it a mean, median or modal average, or statistical tests can be completed to compare the frequency distributions obtained by different conditions in an experimental arrangement. Likewise, the ranking of stimuli can provide data in the form of average ranks, and direct subjective comparisons can provide frequency data regarding various choices. If viewed as mere dependent measures of a behaviour, these data can be analysed by conventional statistical procedures to make decisions regarding the significance of experimental manipulations.

The preferred type of analysis will be explored more fully in subsequent chapters. The main concern here is to address the issue of indirect scaling, and the type of representations that can be constructed to characterize the obtained data set. There are two distinct sorts of indirect scaling, these being attribute scaling and multi-dimensional representations.

2.2.1 Attribute Scaling

The attempt here is to assign a value to each stimulus in the set which describes the psychological value of the attribute in question. The attribute does not have to be a one-dimensional concept, nor do stimuli have to vary on only one attribute. The technique relies on the ability of observers to isolate the attribute of interest in each member of the stimulus set, and to base judgments on that single attribute. Observers are assumed to reduce various components of an attribute to a one-dimensional judgment, which is an appropriate projection of the various components.

For instance, judgments of the attractiveness of paintings clearly involve elements of colour, form and composition. Presumably, attractiveness is a combination of these components. In attribute scaling, observers are assumed to perform the required combination of factors to define the attribute. In these cases, great care must be taken to define the attribute of interest. Vastly different scales might result, even with the same observers and stimulus set, if paintings were rated for their attractiveness and for their artistic impact. Presumably, the stimuli would have the same psychological representation in both tasks, but the salience of each dimension would be different in the judgments of the two attributes.

Attribute scaling requires some degree of variability of judgments to proceed, since virtually all procedures scale attributes on the basis of some concept of the errors of judgment. The most widely used procedure is the Thurstonian (1927) scaling technique, which assumes that the psychological representation is a normally distributed variable, and that differences in judgments are a reflection of this inherent variability. If all stimuli are described as being very good, for example, no psychological scale can be derived. The Thurstonian scale is then a description of the relative positions of stimuli, in standard deviation units obtained from the model. For this reason, variability of responses is required for this type of scaling. No variability defines an infinite distance between stimuli, since two normal distributions must be separated by an infinite distance in order to not overlap. Attribute scales are best applied, then, to reasonably homogeneous groupings of stimuli, cr at least to stimulus sets which cover the range of the attribute in reasonably small steps.

2.2.2 Multi-dimensional Scaling

The goal in multi-dimensional scaling is totally different from attribute scaling, as are the instructions given to the observer in making judgments. In these procedures, the attempt is to place each stimulus in the set into a space which describes the psychological representation of the stimuli. Observers do not make judgments about the value of a stimulus with respect to an attribute, but rather estimate the similarity, degree of difference, or distance separating stimuli on a given attribute. To scale the attractiveness of paintings, then, observers would be presented with two stimuli from the set, and asked to rate their similarity in terms of attractiveness. The estimates of the distances between stimuli are used to construct a psychological space, which describes the position of all the stimuli from the set, such that the distances between stimuli in the space are consistent with the set of judgments. The result is an *n*-dimensional space, where each dimension is some sort of psychological vector required to describe the stimulus. The goal is to minimize the difference between the observed distance estimates and the distances between stimuli in the space, using the fewest possible dimensions.

The definition of the dimensions of the space requires extreme care and a reasonable approach, since the true dimensionality of the space is not known. An error-free fit can always be obtained by using one dimension less than number of stimuli in the set, and the error of the fit decreases with the addition of new dimensions to the psychological space. In practice, however, judgments can usually be accounted for with a reasonable number of dimensions. The result is a plot of stimuli in a multi-dimensional space which is an estimate of the psychological representation of the stimulus set when judgments are made of a given attribute.

It must be understood that this approach does not give a scaling of an attribute, but describes the dimensionality of an attribute. Presumably, in order to make a judgment about an attribute, the observer must weight each dimension, and project a one-dimensional value to describe the attribute of interest. Multi-dimensional scaling can be made of specific attributes, or of the representation of the entire set in terms of the similarity of the elements of the set. The multi-dimensional solutions for the two cases may not be the same.

2.3 Applications

These scaling techniques provide an exciting set of possibilities with respect to the evaluation of teletext systems. When taken together, these procedures hold promise for a large-scale, broad-based assessment of subjective quality.

To begin, attribute scaling can be used to provide first-order estimates of picture quality. The meaning of the scale values are clear from the particular model used to define the scale units. The adequacy of the scale to account for the data set can be tested by a comparison of the expected and observed data set, given the model used to derive the scale. At the very worst, the representation will allow the discussion of the desired attributes from the framework of the model. The model might not be a true description of the psychological representation, but certainly the results can be interpreted in these terms, "as if" the model were true. The representation would be, at least, a useful fiction. Since many people believe that any sufficiently advanced technology is indistinguishable from magic, this should not be a major concern.

Taking a less pragmatic view, these tools can be used to develop a very complete representation of picture quality. There are a lot of unknowns in this process, and the success of each stage of the development of the theory is uncertain, but a plausible research scenario can be described to explore the possibilities. First, a multi-dimensional scaling of picture quality would be most useful, to describe a psychological space of picture quality and the position of various stimuli from a set of typical examples. The dimensions of the space would be undefined, but an examination of the position of stimuli in the space could provide a clue as to the meaning of each dimension. Formally, correlations could be obtained between the dimension values and physical parameters of picture quality and transmission modes. In the best possible case, the dimensions will simply represent parameters of physical picture quality, or some simple combination of these physical measures. If this was successful, new stimuli could be placed in the space on the basis of these physical measures.

To address the question of subjective quality, correlations could be made between the results of attribute scaling and the dimensional values obtained. In all probability, acceptable picture quality, or quality according to any criterion, would be restricted to a region of the space, rather than randomly distributed through the psychological representation. Assuming a non-random distribution, multiple regression techniques could be used to define the relative salience of each of the psychological dimensions of picture quality. If these dimensions could be given a physical interpretation, then predictions of picture quality could be made on the basis of physical measures of teletext systems.

III. CONCLUSION

This discussion has been based on a review of the literature with our goal being to try to extract the collective wisdom of researchers as to the most effective way to measure subjective reactions. A library search was conducted and articles were reviewed which deal with subjective assessments. Our review indicates that there is no real agreement on how to measure subjective reactions. Individual researchers apply specialized procedures to their particular measurement situation. Most of the research projects were relatively limited in scope, in comparison to the type considered here. The only reasonable summary of the literature is that there is no one approach that can be described as being correct, but that the available techniques should be applied in a reasonable, thoughtful manner. It is a matter of tailoring the measurement to the purposes of the project. This first chapter has attempted to provide a broad theortical base for the selection of an appropriate procedure for the evaluation of teletext systems.

CHAPTER 2 - RECOMMENDED PROCEDURES

In the previous chapter it was recommended that indirect scaling of subjective quality be used in the evaluation of teletext systems. This decision leaves a number of options open, both in terms of the procedures used to collect the data and the methods of analysis employed to derive the subjective scales. The purpose of the present chapter is to resolve these issues and to recommend specific procedures. For all intents and purposes, the issue of the appropriate data collection procedure can be unambiguously solved: There is enough of an evaluation protocol established in the teletext industry to indicate a substantial benefit to the use of categorical judgment procedures. The issue of the most appropriate analytic technique is less clear, and only guidelines can be established at this time. The available options will be discussed, and their relative merits evaluated.

I. DATA COLLECTION PROCEDURES

A number of standard psychophysical procedures have evolved over the history of subjective testing, and none has ever been shown to be superior to the rest on any substantive grounds. Preferences abound, certainly, but these preferences are based on factors other than the ability of the procedures to provide data adequate for the indirect scaling of subjective reactions.

In the case of the evaluation of teletext systems, it is recommended that a categorical judgment procedure be used. The case is clear enough that it would be quite irresponsible to make any other choice.

1.1 Categorical Judgment

In this method, observers are presented with a number of response categories which they are to use to describe the attribute of interest. For example, a scale to estimate the subjective heaviness of objects might employ five categories, such as very light, light, average, heavy, and very heavy. The number of categories is formally irrelevant, but in practice, five to seven categories are typically used. The number of choices simply determines the resolution of the scale, at least in theory.

On each trial, the observer is presented with a stimulus, and asked to classify it according to the provided categories. If stimuli are presented in groups rather than singly, the procedure is usually described as a sorting task, but the logic is essentially the same.

If the main concern is the subjective impressions of a particular individual, the observer must make a large number of repeated judgments on the same stimuli to estimate the variability of responses. If the main concern is the characterization of an "average" observer, then a large number of observers can make single judgments of the stimulus set. In the particular application considered here, the average response would often be the primary concern.

1.2 Decision Model

Categorical judgment was developed by Thurstone (1927), and it is based on a particular model of decision making. In the model, each stimulus, X, is mapped to a subjective dimension, S, by some unknown function, f. The dimension S is the attribute of interest, defined by the instructions given to the observer. There is noise in the mapping function, so that each physical stimulus X_i can be described by a mean psychological value S_i with a specific variance. In Thurstone's conception, this notion of variability in the mapping function is fundamental to the characterization of sensory systems. It is referred to as the discriminal dispersion. In the formal Thurstonian theory, discriminal dispersions are assumed to be well represented by a normal density function, as illustrated in Figure 1.

In order to make categorical judgments, it is assumed that the observers can isolate dimension S from all other descriptive psychological dimensions and establish category boundaries, t_g, in the space. For n categories, n-1 category boundaries must be established. (Noise can also be assumed to be associated with the criterion positions, which would also be described as normally distributed variables.) A more complete characterization, showing a number of stimuli and the n-1 criteria positioned on the S dimension, is depicted in Figure 2.

Assuming that the five categories are referred to as Bad, Poor, Fair, Good, and Excellent, t_1 would be the boundary between Bad and Poor, t_2 would separate Poor and Fair, t_3 would deliniate Fair from Good, and t_4 would separate Good from Excellent. On the presentation of a physical stimulus X_i, the sensory impression Y_i would be produced by the function f. Due to the noise inherent in f, Y is regarded as a normally distributed variable with mean S_i and standard deviation a_i. The observer classifies the stimulus by reporting the category into which the value Y_i falls, given the (momentary) placement of the criteria.

Although this characterization is extremely simple, it is the only seriously proposed model for the decision process involved. The details of the conceptualization may change, that is, some theorists may make the criterion fixed rather than variable, or make the function f take a special form, or choose a probability distribution other than the normal, but the basic concept is the same.

One aspect of the decision process which has developed since Thurstone's description relates to the determination of the criterion placements. Thurstone was relatively mute on the mechanism by which criteria were established. Presumably, he felt that the verbal descriptions of Poor, Good, or Excellent would be sufficient to allow observers to establish criteria. As will be discussed below, the Thurstonian analysis procedure actually solves for the criterion positions, so that in his scheme, the only important consideration was that the variation in criterion positions be minimized in order to reduce the error of measurement in any given scaling task.

The work of Parducci (1965) has brought considerable doubt to the idea that verbal descriptions fix criterion placements across scaling tasks. He has proposed that the observer adjusts the category boundaries such that in the long run, each category will be used equally often over the course of the experiment. Further, Parducci has shown that such a strategy maximizes the information transmitted by the use of the scale. In this sense, the observer is assumed to make optimum use of the categories provided.

Parducci has amassed an impressive amount of evidence consistent with his model. On the other hand, all the support comes from experiments in which the category labels are quite arbitrary. For example, observers might be asked to describe a series of lines as very short, short, medium, long or very long. The meaning of the labels here clearly depends on the context.

If all the lines range from 1" to 20", then the observer will classify 1" lines as very short and 20" lines as very long. The observers quite reasonably do not categorize all lines as very short in anticipation of the presentation of a line 7 miles long. Parducci often maximizes the arbitrariness of categories by simply assigning each category a number rather than a verbal descriptor.

According to this view, category judgment scales are not absolute: The frequency of each response depends on the set of stimuli used in the experiment. The meaning of the categories Good, Bad, Excellent, and so on depends on the stimuli chosen for evaluation, so that the exact same stimulus can be given quite different evaluations when presented to the same observers in different stimulus sets.

In a model explicitly developed to evaluate teletext picture quality, Allnatt and his colleagues (1973; 1975; 1979) have taken quite a different view of the problem of criterion placement (see Section 2.2). In this conceptualization, the psychological continuum, *S*, is mapped onto a second continuum, *t*. On the *t* continuum, criteria are placed such that the range of the scale is divided into equal parts. As far as can be discerned from the relevent papers, there is no real evidence for this suggestion. The assumption seems to have been made to make the mathematical analysis more straightforward. As will be discussed below, the Allnatt analysis scheme provides considerable promise for the characterization of subjective responses to teletext systems, but the assumed process of criterion placement for categorical judgments is an unsubstantiated component of Allnatt's system.

Basically, then, the decision-making model used to represent categorical judgments is essentially a Thurstonian model. Each stimulus is thought of as producing a discriminal dispersion on a subjective attribute. The observer classifies each stimulus by the comparison of the resultant subjective attribute value with the values of the category boundaries. The only real disagreement in the literature concerning this basic model is the process by which observers determine the positions of these category boundaries on the subjective scale.

1.3 Justification of Category Judgment

The justification for the use of categorical judgment as a primary procedure for the evaluation of teletext systems is quite simple. It is as good as any other indirect procedure, it is easy to use, it is widely used in subjective assessment and it has more or less become the standard in the industry.

Category judgment is the method recommended by the CCIR (1974) as the preferred method of Videotext evaluations. The recommendations include specifications of viewing conditions, number of observers, instructions to observers, and the like.

These functions are formally independent of the procedure, but the adherence to a standard procedure can do nothing but reduce the error component of cross-laboratory comparisons. The fact that this procedure is recommended by an international agency which seems to carry some weight in the industry is a real advantage.

Category judgment is the method of preference for the subjective evaluation of related telecommunications products. The grade of service models employed by AT&T (Cavanaugh, Hatch and Sullivan, 1979) and BNR (Lui and Ebert, 1976) are based on subjective evaluations through category judgments. Thus, an added advantage is that consulting and technical expertise can be sought in these highly specialized and related industries. Likewise, advances in technology in that sector could readily be applied to the current problem if the procedural differences were minimized.

From an even more global perspective, category judgments seem to be the most widely used form of subjective evaluations. In compiling the recommendations offered in chapter 1, a series of 95 papers were collected and reviewed. These reports gave procedural details on about 158 subjective measurements made on a wide range of topics. Of these reported measurements, a full 70 (44.3%) involved categorical judgments. The next most popular procedure was accuracy methods (23 measurements - 14.6%), and since accuracy involves an objective procedure, it cannot apply here. To complete the survey, the other procedures used in the literature were choice behaviour methods (23 cases - 14.6%), magnitude estimation (14 cases - 8.9%), ranking (11 cases - 7.0%), verbal descriptions (7 cases - 4.4%), paired comparisons (5 cases - 3.2%), and methods of adjustment (4 cases - 2.5%).

Procedures of ranking and paired comparisons result in subjective measurements very similar to that obtained from the recommended categorical judgment procedures. Both, however are unwieldy for large stimulus sets. In ranking, the entire group should be simultaneously presented for optimum results. If the experiment had a large stimulus set, say 50 pages of teletext, that would mean the simultaneous presentation of 50 monitors. In paired comparisons, observers make preference judgments between all possible pairings of the stimulus set taken two at a time. For a 50 item stimulus set, this means 1225 judgments per observer. In both ranking and paired comparison procedures, some labour-saving presentation regimens are available, but only at the cost of complicating assumptions.

The other related procedures are those involving choice behaviour. In this approach, observers are given the entire set of stimuli, and asked to state which one they prefer. Luce (1959) has shown that by making some reasonable assumptions about the nature of that judgment, a scale of subjective quality or preference can be derived from the frequency of choice from the set. However, it is now apparent that the obtained solution is very similar to the Thurstone solution. In fact, given certain assumptions about the distribution of errors in the choice model, the two are identical (Luce, 1977). Further, a direct comparison of the Luce and Thurstone models has indicated that the latter tends to fit the data with greater precision (Kornbrot, 1978).

The conclusion is inescapable. The vast majority of evaluations, both in the general psychological literature and in the evaluation of teletext graphics employ categorical judgment. It is an easy procedure to use which observers can learn quickly. No other procedure seems to be demonstrably superior, so that there is no reason to deviate from the choice of others. By taking this course of action, compatibility can be maintained with other laboratories, and the project can aid in the refinement of the adopted protocols.

1.4 Specific Procedural Recommendations

The procedure of category judgment has been recommended as a standard procedure for the subjective evaluation of teletext systems. The reasons are to maintain compatibility with the rest of the industry and because there is a well-defined decision model available for the task. At the same time, the procedure is susceptible to specific problems. For example, the criterion placements may change, depending on the stimulus set, and the scale obtained is only unique up to a linear transformation (see below). The following specific recommendations should help to minimize the problems and maximize the advantages of the procedure.

1. Adhere to the measurement recommendations of the CCIR. This will increase the comparability of the measurements with those taken in other laboratories.

2. Test about 50 observers in an experiment, at least as a first guess as to the number of observers required. Simulations of category judgments of telephone grade of service has indicated that this is a reasonable number of observers to produce stable results (Kort, 1983).

3. Give the observers a series of practice trials, probably between 10 and 20 judgments, with stimuli which provide a reasonable sample of the range of teletext qualities to be assessed in the experiment. This will assist in the establishment of stable criteria. If practice trials are not provided and Parducci's suggestions are correct, the initial experimental trials will be dominated by large shifts in criterion placement, as observers try to optimize the information transmitted by the scale. If Parducci's ideas are not correct and the meaning of the categories define the boundaries, the inclusion of these practice trials will not matter much. A little practice never hurt anybody. Unfortunately, contrary to the old adage, it never makes anybody perfect either.

4. Regardless of the stimulus set of interest, always include samples which span the entire range of quality. For instance, even if the prime interest was to assess a number of stimuli with relatively good quality images, poor and excellent images should also be included for consideration by the observers. The purpose here is to prevent drastic shifts in criterion placements, which make comparisons across experiments difficult. The assessment of the same approximate range of quality in pictures across experiments will minimize analysis problems due to criterion placement. Since the relevant stimuli will be presented in a random fashion, these stimuli should also be randomly intermixed.

5. The most important and innovative recommendation has been left to the last. Since this is the initiation of a relatively long-

term and unified research effort, a very simple procedure can be used to unify all the experiments in the set. A few standard stimuli should be created which approximately span the range of picture quality. Choose one poor, one average, and one excellent quality frame with quality varying on as many dimensions as possible. Include these stimuli in every quality assessment experiment performed. This will do two things. First, it will approximately define the range of stimuli across experiments to a standard value. More importantly, this procedure will provide a common standard by which all measurements can be compared.

The last recommendation is extremely important in categorical judgment. In the following section, analysis schemes will be described. The most widely used analysis, Thurstonian scaling, produces a subjective scale linear with true psychological representation. The parameters of that linear transform depend on which stimulus is chosen as a standard, because that stimulus sets the zero point of the scale. The standard deviation of its discriminal dispersion process determines the unit size of the scale.

Without a set standard the procedure will arbitrarily use the lowest ranked stimulus in the set. Therefore, the linear transform between each obtained scale and the true psychological scale can change from one measurement situation to the next. This makes the results quite difficult to compare across experiments, even within the same laboratory. However, if the same stimuli are included in each experiment, the most stable of these can always be used as a referent, standardizing the origin and the unit of the scale. Thus, all the scaled solutions from all experiments should be set in the same linear relation to the "true" psychological scale. This should mean that the evaluations would be directly comparable.

II. DATA ANALYSIS PROCEDURES

Two data analysis methods will be described. Thurstonian scaling is a traditional procedure which is generally used for subjective evaluations. The second procedure is an approach developed by Allnatt, specifically for use in the evaluation of teletext systems. However, some crucial assumptions are made in this analysis, which may or may not be justified. The claims Allnatt makes for this analysis system make it very appealing, because the implications are that the resultant scale of teletext impairment is additive. That is, the claim is that if one noise source impairs subjective quality by a and another independent noise source impairs subjective quality by b, then the effect of both sources presented together is $a+b$.

One bright spot in the analysis problem is the fact that the data collection procedure for both analyses is the same, so that if a clear decision cannot be made between the two analyses, both can be applied to the same data.

2.1 Thurstonian Scaling

The Thurstonian solution involves the estimation of the positions of the discriminal process distributions on the subjective dimension S. In order to obtain the solution, stimulus X_1, the lowest ranking stimulus on the scale, is arbitrarily assigned a scale value of zero. (Although, as noted, any stimulus can serve as the standard, for purposes of this discussion we will detail the procedure as it is typically applied.) The position of X_2 with respect to X_1 is measured in standard deviation units of the discriminal dispersion of stimulus X_1. The position of X_3 with respect to X_2 is similarly determined, and concatenated with the difference between X_1 and X_2, to produce a scale position for X_3. This process is repeated with successive stimuli until the entire stimulus set has been positioned on the subjective dimension S. The overall strategy is like measuring a football field with a six-inch ruler.

Here, the origin of the scale is determined by X_1 and the step size or basic unit is set by the standard deviation of the discriminal dispersion of stimulus X_1. No physical measures are used, so the scale is completely psychological. The "true" psychological scale is not recovered, but the result is linear with that "true" scale, at least according to the model. The parameters of the linear transform are entirely determined by stimulus X_1, or, more generally, by whichever stimulus is chosen to act as an arbitrary reference. If recommendation 5 is followed, the chosen standard common to all experiments will become the referent.

2.1.1 Calculation Details of Thurstonian Scaling

The basic data for categorical judgments is a matrix of size n x K, where n is the number of categories and K is the number of stimuli used in the measurement. A cell in the matrix is the frequency of occurrence of a given judgment for a particular stimulus. For Thurstonian scaling, the first step is to convert these data to probabilities, and to rank the stimuli from the lowest to highest in terms of the obtained judgments. This is only a first-order ranking, and it is done by arbitrarily assigning the values from 1 to n to the n categories. The lowest ranking category (i.e., Bad) is assigned the value 1, and the best category (i.e., Excellent) is assigned the number n. A mean opinion score is calculated for each stimulus, and the stimuli are ranked on the basis of the mean opinion score.

The matrix is then used to obtain a cumulative probability distribution for each stimulus in the set, as a function of the category number. These data represent the probability that a stimulus would be judged at or below the category in question. Since the cumulative distribution sums to one, the last category is lost, and we now have a matrix of size n - 1 x K. Under the normality assumption, the cumulative probabilities are transformed to Z scores to obtain a new matrix. This matrix is a matrix of positions of the n-1 category boundaries described in standard deviation units relative to the mean of the discriminal dispersion process for each stimulus. In the following paragraphs, the subscript i will be used to denote stimuli, and subscript g will describe category boundaries. Variable Z will refer to the values entered in the category boundary matrix, so that Z_{gi} is the Z score of the gth boundary for the ith stimulus.

The matrix of Z_{gi} values provide the raw data for the Thurstonian analysis. Successive stimuli are chosen, first stimulus pairs 1 and 2, then 2 and 3, through to pair K-1 and K, and the standard deviation difference between the pairs is determined. This can be done graphically by plotting the values of Z_{gi} as a function of $Z_{g\ i+1}$. If the assumption of normality is met, even to a first approximation, the plot should be linear. The slope of the least-squares linear fit is the ratio of the standard deviations of the discriminal dispersions. If the standard deviation of stimulus i is denoted as a_i, then the slope, $M_{i\ i+1}$, will be equal to the ratio a_{i+1}/a_i. The Thurstone technique arbitrarily sets a_1 to 1.0, so that all the standard deviations can be solved as $a_{i+1} = a_i \times M_{i\ i+1}$.

The intercept of the same plot, $B_{i\ i+1}$, is the difference between the means of the discriminal dispersions, defined in standard deviation units of a_i. The subjective position S_{i+1} of stimulus X_{i+1} in relation to stimulus X_i is thus $S_{i+1} = S_i +$

$a_iP_{i\;i+1}$. In this way, the mean and variance of the discriminal dispersions of the entire data set are determined by the successive concatenation of results. The reader should note that this process will always involve K-1 plots regardless of whether it is done as outlined here or as suggested in Chapter 3.

In this analysis, it is possible that some stimuli have only one Z value because there was so little variability in responses. These stimuli must be eliminated from the analysis because the successive plots cannot be realized. If their inclusion is critical, their position on the S dimension can be estimated by placing them in their approximately proper position relative to the scaled values, on the basis of their mean opinion score (see above).

Once the stimulus positions S_i and standard deviations a_i are determined, the criterion placements can be determined. The value Z_{gi} is the criterion boundary placement for the ith stimulus, in standard deviation units of stimulus X_i. Therefore, the criterion placement t_g for stimulus X_i can be defined as $Z_{gi}a_i$ + S_i. There can be up to K samples of this value, one for each stimulus in the set, provided that the value Z_{gi} could be calculated for all i. The estimate of the criterion placement is thus the mean of these estimates, so that:

$$\overline{t}_g = \frac{1}{K} \sum_{i=1}^{K} t_{gi}$$

If $\underline{t}_g$ is not provided for a particular stimulus, it is excluded from the sum and $\underline{K}$ is reduced by one.

The procedure obtains, then, estimates of the mean and standard deviations of the discriminal dispersion processes for each stimulus, along with criterion placements for the set. The procedure outlined here is covered in Torgerson (1958), and other analysis procedures are described there as well. The procedure included here was chosen because its steps are intuitive with respect to the model, and it can be easily realized in computer code. The judgments involved are variable enough that the exact calculation procedure is not of primary concern.

2.2 Allnatt's Procedure

The Thurstone solution is a general one and it is widely used. Allnatt's procedure has been developed specifically for teletext evaluations, and the impairment of image quality. The material reviewed here is contained in a series of papers published by Allnatt and his associates over a number of years (e.g., Allnatt, 1973; 1975; 1979).

In the Allnatt approach, category judgments are made of stimuli which vary on a physically quantifiable impairment dimension. In the standard procedure five categories, labelled from Excellent to Bad, are used. The research program Allnatt has undertaken focusses on three main concerns:

1. The nature of the psychophysical function which relates the degree of impairment (or physical quality) with the perceived quality of teletext displays.

2. The relation between the perceived quality derived from psychophysical functions and the recorded categorical responses.

3. The result of combined impairments from multiple, independent noise sources.

Each aspect will be considered in turn.

2.2.1 The Psychophysical Function

The psychophysical function is assumed to be a power function of the form

$$\Psi(D) = a D^{b}$$

where $\Psi(D)$ is the perceived impairment on scale $\underline{S}$, $\underline{D}$ is the physical impairment, and $\underline{a}$ and $\underline{b}$ are constants specific to the units of $\underline{D}$. The justification is the general success of this representation in sensory scaling. Supportive data were discussed in the first chapter.

In some ways, the assumption is fairly weak, in that the only requirement is that the function be power-like, and not necessarily follow the exact form. To a first approximation, this is probably a reasonable guess, because the power function is quite versatile in fitting monotonic functions, regardless of the "true" form.

In other ways, however, the assumption is much stronger in that it requires the mapping to be onto a unidimensional psychological space, what we have been calling the *S* scale. In the case of many physical dimensions (i.e., height, weight), a direct mapping can be reasonably assumed. However, when rating the impairment of teletext systems, the representational space may actually involve two or three dimensions. If so, a further assumption must be made; that these dimensions are combined in a static fashion (perhaps by weighting them 50:50) to produce values on the *S* scale. Variance in either the nature of this multidimensional space or the way in which the dimensions are combined (because of, say, differential experimental instructions) may invalidate the final scaling solution. Thurstone's procedure, while it also requires a unidimensional *S* scale, would be much more robust in the face of this type of variance.

2.2.2 Category Choices

This process is crucial to Allnatt's approach. It corresponds to the decision process in Thurstonian scaling. Unfortunately, it is difficult to extract a clear rationalization of Allnatt's decision process from his descriptions. Quite often, the concept can be expressed, but the mathematical realization is a tad obtuse.

Allnatt suggests that the observer performs the category judgment task by a specific set of operations. First, the observer normalizes the scale of Ψ . This is done by defining a psychological quantity Ψ ($\underline{D}_M$), where $\underline{D}_M$ is the physical impairment required to split a five-point opinion scale in half. Upon presentation of $\underline{D}_M$ the mean opinion rating would be 3.0 on a 1 to 5 scale. Observers express the psychological magnitude of the stimulus $\underline{D}_i$ as a ratio of Ψ ($\underline{D}_M$). This ratio is scaled by a second power function, which can be expressed as a ratio of the original exponent $\underline{b}$, so that we can write

$$\begin{aligned}\Psi_n(D_i) &= (\Psi(D_i)/\Psi(D_M))^{G/b}\\ &= (a\, D_i^{\,b}/a\, D_M^{\,b})^{G/b}\\ &= (D_i\ /D_M)^{G}\end{aligned}$$

The value $\Psi_n(\underline{D}_i)$ represents the psychological value of $\underline{D}_i$ on the normalized scale. The value of $\underline{G}$ is thought of as an observer-dependent parameter, which gives the model additional degrees of freedom.

Since $\underline{D}$ is an impairment parameter, the values of Ψ_n increase with stimulus degradation. This quantity is further transformed to a normalized (0 to 1 representation) acceptability scale $\underline{t}$ by the relation

$$t = \frac{1}{1 + \psi_n}$$

Thus, we have two descriptions of the same thing, a ψ_n scale which is a psychological impairment scale (what we have called the *S* dimension), and the *t* scale which is the corresponding acceptability scale and from which the response is to be determined.

At this point we come back to something very similar to the Thurstone decision model. The *t* scale is divided into *n* equal steps, corresponding to the *n* categories provided. There is variability in the representation of *t*, and the decision process is the process of determining the category into which the value on *t* falls. The category boundaries are placed, again, to equally divide the range of *t*. The proportion of judgments in each category allow us to estimate the distribution function of *t*, *F*(*t*). This is quite similar to the concept employed in the Thurstonian analysis scheme. From this function the median of the distribution (t_m) is derived by interpolation. This value is the scale value which Allnatt's procedure utilizes.

According to the original decision model and the normalization process, the equation for the median of the category judgment distribution for a given impairment $\underline{D}_i$ should be:

$$t_m(D_i) = \frac{1}{1 + (D_i/D_M)^G}$$

This claim can be evaluated by defining a parameter $\underline{J}_i$,

$$J_i = 1/t_m - 1$$

Defined in this way, it must be true that

$$J_i = (D_i/D_M)^G$$

so that we can predict that

$$\log J_i = G \log D_i - G \log D_M$$

Thus, the plot between log $\underline{J}$ and log $\underline{D}$ should obtain a straight line with a slope of $\underline{G}$ and an intercept of $-G\log(D_M)$. These parameters would allow a determination of the psychological scale $\psi(\underline{D})$, although it is not an important aspect of Allnatt's procedure. What is important is that the straight-line function be empirically obtained, since failure to do so invalidates the analysis.

2.2.3 Combining Impairments

This all sounds very strange. There is very little psychological theory to justify the assumptions, and it is even hard to characterize the exact nature of the decision process. However, Allnatt (1975) has provided some evidence for the notion that impairments from independent noise sources can be predicted by the additivity of *J*, such that for noise sources 1 and 2

$$\begin{aligned} J_{1,2} &= J_1 + J_2 \\ &= (D_1/D_{M1})^{G_1} + (D_2/D_{M2})^{G_2} \end{aligned}$$

Thus, the median of the distribution of the category judgments with both impairments can be predicted as

$$t_{m1,2} = \frac{1}{1 + (D_1/D_{M1})^{G_1} + (D_2/D_{M2})^{G_2}}$$

Since all the parameters in the right-hand side of the equation are defined by the separate analyses of judgments with impairments D_1 and D_2, the joint effects should be predictable.

III. SUMMARY

The evaluation of teletext systems should be done using categorical judgment procedures. Specific recommendations detailed in this report should be followed to optimize the use of these procedures.

In terms of the analysis, the same data can be represented by Thurstonian scaling or by the Allnatt procedure. Allnatt's approach assumes specific criterion placements and a general form of the psychophysical function. In addition, it is based on the notion of impairments which can be defined on a physical dimension. The payoff, however, is in the claim that the combined effects of impairment can be predicted. Thurstonian scaling is theoretically established, but is less optimistic in the analytic solution of the effects of compound impairments.

In the short term, both analysis schemes are recommended, until such time as the utility of the two can be empirically compared. After all, the only perfect science is hindsight.

CHAPTER 3 - IMPLEMENTING THE TWO PROCEDURES

In the previous chapter, two issues were addressed: the optimal data collection technique to use in evaluating teletext systems, and the best way to analyze those data. With respect to the first of these issues, the categorical judgment technique was deemed superior to any other techniques for a number of reasons (refer to the previous chapter for a discussion of these reasons). With respect to the analysis question, two methods were suggested: Thurstonian (1927) scaling, and Allnatt's (1973; 1975; 1979) more recently presented technique. At present, neither of these seems to be clearly superior to the other. In the present chapter, a more complete summary and comparison of these two techniques will be presented. For both techniques, discussion will centre on three issues: a) the assumptions underlying the analysis, b) the nuts and bolts of how the analysis is carried out, and c) how to determine whether the technique can be legitimately applied (including means of testing the assumptions).

I. THURSTONIAN SCALING

1.1 Theoretical Underpinnings

As with all scaling techniques, Thurstonian scaling is based on the idea that there is a subjective dimension $\underline{S}$ representing only the attribute of interest. For our purposes, we can consider that attribute to be acceptability. When a stimulus is presented, it undergoes an analysis which ultimately yields a value on that dimension. Subjects must then use this value to produce a response on whatever response scale the experimenter has provided.

Thurstone has actually suggested a number of slightly different approaches to the scaling problem. They vary simply in the assumptions each makes. The approach we are suggesting is referred to as Case IV. It's important assumptions are as follows:

1. The value $\underline{Y}i$ produced by a given stimulus $\underline{X}i$ on the $\underline{S}$ dimension can be characterized as a random selection from a normal distribution having mean $\underline{S}i$ and variance $\underline{a}^{2}i$.

2. On each trial, $\underline{n}$-1 criteria are placed on the $\underline{S}$ dimension dividing it into $\underline{n}$ sections ($\underline{n}$ is the number of categories the observer is asked to use). Each section corresponds to a category. The response given is the category corresponding to the section into which $\underline{Y}i$ falls. Further, although the positions of the criteria may vary from trial to trial, this variation is uncorrelated with the value of $\underline{Y}i$.

There are two important issues associated with the first assumption. The first is the assumed shape of the distribution. Since all subsequent calculations depend on the assumption of normality, the shape of the distribution should be evaluated. The method of evaluation will be discussed in section 1.3.1.

The second issue concerns the idea of a one-to-one mapping from each stimulus X_i onto a mean subjective impression S_i. Since the S_i represent the scale values of the stimuli, their determination is essentially the goal of this analysis. If the analysis is to yield meaningful values it's important that the S_is remain relatively stable both over the course of the experiment and over experiments using identical experimental parameters.

In the previous chapter, a set of experimental procedures was outlined which should maximize the chance of the S_is remaining stable. However, there are no guarantees here nor is there any way to determine whether the assumption holds throughout the experiment. Variations across seemingly identical experiments can, of course, be detected and, if the discrepancies are substantial, the technique would have limited usefulness. That is, ultimately one may want to examine changes in S_i as a function of other variables (e.g., instructions). To do so one must be sure that irrelevant variables like habituation, or perhaps time itself, are not affecting the S_is.

There is one very important issue associated with the second assumption. While it is not crucial that the criteria remain stable throughout the experiment, whatever variation there is must be random rather than systematic. If the variation is random, no problems are created for the analysis. The only change would be that the calculated criterion positions would be estimates of an average position rather than a stable position. However, if the variation were systematic (e.g., if the top two criteria move up the dimension whenever a high quality stimulus is presented), the obtained S_i values would be relatively meaningless. Fortunately, there is a test (to be described in section 1.3.2) which should allow us to determine whether there are stability problems.

1.2 Calculation Technique

The theoretical rationale for calculating the S_is was presented in the previous chapter. Here, we would like to concentrate more on the calculation details through the use of an example. Following the suggestions presented in the previous chapter, observers will be asked to use 5 categories (category 1 reflects the lowest acceptability, category 5 the highest). Two stimuli X_0 and X_6 are included in the experiment to help establish the range of the scale to be used in the observer's ratings. Stimulus X_0 is very poor in quality and serves as a lower anchor. Stimulus X_6 is as close to perfection as can be physically achieved and, thus, serves as an upper anchor. The data from these stimuli will not be considered in the analysis. However, if the ratings given these stimuli are not as expected (mainly 1s and 2s for X_0, mainly 4s and 5s for X_6) it would be a cause for concern. A third stimulus, X_1, is the standard stimulus which is used in all scaling experiments. It is created by combining a number of different types of impairments to produce an intermediate level of acceptability. It will serve as our referent stimulus in the calculation process. Finally, stimuli X_2, X_3, X_4 and X_5 are the stimuli we wish to scale.

Each stimulus will be presented to the observer a numberof times (say 100 to keep our ratios simple). The steps in the analysis would be as follows.

1. Create a data matrix like that presented in Table 1.

2. Calculate a mean opinion score for all stimuli via the formula

$$M_i = \sum_{j=1}^{n} \frac{F_j}{N} \times j$$

where n is the number of categories, j is the category number, Fj is the frequency per category and N is the total number of times the stimulus was presented. Table 1 also contains the mean opinion scores for the seven stimuli.

3. Check the Mi values for the anchor stimuli $X0$ and $X6$. If they are at the appropriate levels, we can assume they have served their purpose and, thus, their data can now be disregarded.

4. Interchange the rows of the remaining stimuli (including the standard) so that the M_is are in descending order as in Table 2.

5. Turn these frequencies into probabilities as in Table 3.

6. Transform each row of this matrix to produce a cumulative

probability matrix as in Table 4. The last column will always be 1.00 and, thus, can be dropped.

7. Using a *Z* table convert the cumulative probabilities in Table 4 into a matrix of *Z* scores as in Table 5.

8. Using *X*1 as the standard, create separate plots of the *Z* scores for *X*1 against the *Z* scores for each of the other stimuli. These are shown in Figure 3. (Note that the *Z* scores for *X*1 go on the Y-axis. Note also that there are 4 plots. This follows from the fact that there are 5 stimuli being used in the analysis.) Correlation coefficients, slopes and intercepts of the best fitting straight line should be calculated for each plot. These are shown on the figure. In each case, the intercept of the line is the scale value, *S*i, for the stimulus being compared to *X*1. The unit is the standard deviation of the distribution for *X*1. Essentially what is being done is that *S*1 has been set to 0 and *a*1 to 1. This is perfectly legitimate since the *S*is are only determined up to a linear transformation in any case. The slope of the line is the ratio of the standard deviations (e.g., *a*i to *a*1). Since *a*1 has been arbitrarily set to 1, the slope can be considered to be our best estimate of *a*i. Thus, the *S*is and *a*is are the intercept and slope values found on the figure.

9. Finally, the positions of the criteria should be calculated. Each stimulus should allow an estimate of the position of each criterion with respect to its own mean. For a given stimulus *X*i the criterion points *t*gi can be estimated by the following formula,

$$t_{gi} = S_i + a_i Z_{gi}$$

e.g., for stimulus X_2

$$t_{12} = S_2 + a_2 Z_{12}$$

$$= -.36 + (.996 \times -1.27)$$

$$= -1.63$$

where the *Z*gi are the values listed in Table 5. These values are listed in Table 6.

10. The overall estimate for the position of each criterion is obtained by averaging over these estimates

$$\bar{t}_g = \frac{1}{K} \sum_{i=1}^{K} t_{gi}$$

where K is the number of stimuli being scaled. These are also contained in Table 6.

This description of the analysis process has been for an ideal data set. One problem that often arises is that the variability for a stimulus might be sufficiently small that some values in the Z matrix might be $\pm \infty$ (e.g., X_5 in the present example). In this circumstance, certain alterations are necessary. First, when determining S_i for this stimulus, points such as these are obviously not plotted (note that the X_1 versus X_5 plot has only 3 points). Second, when determining criterion placements these points are simply omitted, decreasing K by one. (The final criterion placement was calculated in this fashion.)

If X_1 itself has one or more Z values of $\pm\infty$, a situation could be created in which the plot of X_i versus X_1 has only 2 points. This would be extremely unfortunate. In this circumstance, a concatenation technique should be used. After the stimuli have been ordered according to M_is, proceed as before for stimuli both immediately above and immediately below X_1 in the ordering (call these stimuli X_a and X_b). For the stimulus immediately above X_a (call it X_{a+1}), create a plot with Z values of X_{a+1} on the X-axis and those for X_a on the Y-axis. The intercept of this function will be the difference between means for the two stimuli X_a and X_{a+1} in terms of the standard deviation of X_a.

It would then be necessary to change the units of this value so that they are equal to a_1. To do this, multiply the intercept by the ratio a_1/a_a. This value can then be added to S_a to produce the scale value for S_{a+1}. That is,

$$S_{a+1} = S_a + \frac{a_1}{a_a} \times B_{a+1,a}$$

where $B_{a+1,a}$ is the slope of the line relating the Z scores of X_{a+1} and X_a.

A similar procedure would then be carried out using X_b to produce the scale value for the stimulus immediately below it in the ranking. If more stimuli need to be scaled, we can simply continue the concatenation process. For any stimulus whose mean is larger than that for X_1, the stimulus immediately below it in the ordering is used while for any stimulus whose mean is less than that for X_1, the stimulus immediately above it in the ranking is used.

Hopefully, with this concatenation technique, the bulk of the plots will always involve at least 3 points. If any involve only 2 points large estimation errors can arise and the normality assumption cannot be tested (see section 1.3.1). If a given stimulus produces only one *Z* score, the technique itself cannot be used. This would occur, for example, if a given stimulus is only rated good or excellent. The basic problem is that this single point won't allow an estimate of both a mean and a standard deviation for the stimulus. If this occurs, a scale value must be estimated in a somewhat different fashion. After final S_i values for the other stimuli have been determined, the regression equation relating these values to their respective M_i values should be calculated. This equation should then be used to predict the S_i for the problem stimulus based on that stimulus' M_i value.

1.3 Justifying the Technique

One thing to realize about Thurstonian scaling is that it can be applied to any set of stimuli. Unlike Allnatt's technique, which will be discussed shortly, the stimulus set does not have to vary along a quantitative dimension. This fact actually has both positive and negative implications. The positive implication is that all stimuli, even those which are only qualitatively different, can be scaled. The negative implication is that even if we can empirically validate the model, we learn nothing about the scale values of any stimuli not actually used in the experiment.

More specifically, the aim of any technique which requires variation on a physical dimension is to specify a function relating values on this dimension to values on an internal dimension. If such a function can be validated empirically, the effects of additional variations on that dimension can be determined without further empirical work. The other technique to be discussed (Allnatt's) does require variation on a physical dimension, and, if validated, will yield a psychophysical function. In addition, with a method that also needs to be validated empirically, it may allow investigators to specify *a priori* the scale value for stimuli which vary along two physical dimensions. Thus, this technique, if successful, could be a much more powerful and useful tool than Thurstonian scaling in the evaluation of teletext systems.

With respect to Thurstonian scaling, there are two tests that can, and should, be performed before accepting the derived scale values as legitimate. In the first instance, the assumption of normality should be evaluated. The second test is a test of the relative stability (or nonsystematic variation) in the positions of both distributions and criteria along the *S* dimension. If either test is unsuccessful, the results of the procedure would have to be regarded with extreme suspicion.

The more important of these two tests is the second one. Failure here indicates that criterion placements vary systematically with positions of the discriminal distributions. Thus, our ability to locate and talk about the position of these distributions relative to established points of reference would be minimal. The problems created by a lack of normality would be less severe. Fortunately, the more powerful test is the test for criterion stability.

1.3.1 Testing Normality

Testing the normality of the distributions involves an examination of the plots shown in Figure 3. (This test can, and probably should, be carried out before any further analysis is undertaken.) If the normality assumption is correct each plot should be well described by a straight line. The linearity of these relationships can be tested by simply comparing the value of the correlation coefficient to a criterion value. Obtained values less than the criterion would suggest that the assumption should be rejected.

The correlation coefficients can be found on their respective plots in Figure 3. For a four-point plot, a value of .90 should serve as the criterion while, for a three point plot, a value of .988 should be used. (If a graph contains only 2 points, this test simply cannot be made.) In all cases, these criterion values represent those values for an hypothesis test with an alpha of .05 (one-tailed). In every case shown here, the obtained *r* is greater than the criterion, suggesting that the normality assumption may be valid.

This test of normality is obviously not a powerful one. It would be more powerful if more categories were used in the data collection procedure and, consequently, more points appeared in these plots. However, for the sake of consistency across laboratories, we will hold to the recommendation of using 5 categories. Thus, these plots will never involve more than 4 points. If, by and large, they do involve all 4 points, no real problems should arise. However, if many contain only 3 points, the deviation from normality would have to be extreme before it would be detected. Further, as noted, if a plot contained only 2 points, the test simply could not be performed since 2 points always lie on a straight line.

1.3.2 Testing Stability

We have already determined 1) the means of the distributions on *S*, 2) the standard deviations of these distributions, and 3) the relative positions of the criteria in terms of the standard deviation of the referent stimulus. In our "stability" test, we will begin with the assumption that these values are all valid and then attempt to regenerate the original data. If we can do so to a suitable degree of accuracy we can conclude that the means, standard deviations and criterion placements represent stable characteristics. This test will always be the last test in the validation process. If it also is successful we can then regard the distributions' means as legitimate representations of the scale values of the stimuli being examined.

The technique is quite simple. Using the observed mean and standard deviation and the derived criterion placements, the proportions of scores falling into each category can be calculated based on the normality assumption. (The normality assumption must, of course, be validated first.) This is essentially a matter of determining the relative positions of the criteria in each distribution. That is, a Z score is calculated for each criterion in terms of the mean and standard deviation of each distribution (see Table 7). These scores are then used to calculate the expected proportion of responses falling into each category by using a Z table (see Table 8). These proportions can then be turned into expected frequencies (see Table 9) which can be compared against the actual data. (Note that the expected frequencies should be correct to one decimal place because the original data set were integer values.)

At this point, the expected frequencies should be surveyed in order to make sure none are less than 5.0. In the present case, two cells are (the category 1 cells for X_3 and X_4). In cases like this, the expected frequency matrix (and the associated data matrix) must be altered slightly. The problem cells should be combined with their closest neighbor to create expected and obtained frequencies for placing these stimuli into either category 1 or category 2 (see Table 10).

A χ^2 statistic is then computed by applying the following formula to each cell in the altered matrices and then summing over all the cells.

$$\chi^2 = \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

In the present case, the obtained χ^2 value is 12.46. This value is evaluated against a χ^2 distribution with the number of degrees of freedom equal to the number of cells in the altered matrix minus the number of stimuli. Here the degrees of freedom is 18. If the χ^2 is significant, there is reason to suspect that the stability assumption is incorrect. To minimize the likelihood of failing to detect violations of the assumptions, it is best to choose a very liberal alpha value, for example, .10. The .10 cutoff for the χ^2 distribution with 18 degrees of freedom is 26.0. Since our obtained value is well below 26.0, the assumption is not demonstrably violated. Thus, all things considered, the scale values obtained for our stimuli seem to be valid ones.

II. ALLNATT'S SCALING TECHNIQUE

2.1 Theoretical Underpinnings

Like Thurstonian scaling, and all other scaling procedures, Allnatt's scaling technique is based on the notion of a subjective dimension, *S*. In Thurstonian scaling, we talked about *S* as a dimension of acceptability with higher values reflecting higher levels of acceptability. Here, *S* will be an impairment dimension with higher values representing lower levels of acceptability. (These dimensions could conceivably be regarded as the reverse of one another.) According to Allnatt, however, observers do not use the *S* dimension to determine their responses. The obtained values on the *S* dimension are mapped onto a second, response dimension, *t*, which runs from 0 to 1. It is the value on the *t* dimension which is used in the response process.

With respect to these processes, the following assumptions are made.

1. The momentary value, *Y*i, produced by a given stimulus, *X*i, on the *S* dimension can be characterized as a random selection from a distribution having mean *S*i where

$$S_i = a\, D_i^{b}$$

The value *D*i is a measure of impairment on a physical dimension while *a* and *b* are constants specific to the units of the dimension *D*.

2. The *Y*i value is transformed internally to produce a value *t*i on the *t* dimension which can be characterized as a random sample from a distribution with median *t*mi where

$$t_{mi} = \frac{1}{1+\left(\dfrac{S_i}{S_M}\right)^{G/b}} = \frac{1}{1+\left(\dfrac{aD_i^{b}}{aD_M^{b}}\right)^{G/b}} = \frac{1}{1+\left(\dfrac{D_i}{D_M}\right)^{G}} \qquad (1)$$

The value D_M is the physical impairment which will produce a median *t* value exactly in the middle of the *t* dimension (i.e., t = 1/2) while *C* is a parameter dependent on the physical dimension being investigated.

(Note: there is an important difference between the two subscripts *M* and *m*. *M* is used to refer to a particular stimulus, that stimulus which bisects the *t* dimension. In any given experiment it's unlikely that this stimulus would exist. The other subscript, *m*, refers always to the *median value* on the *t* dimension *for a given stimulus*.)

3. The *t* dimension is divided into *n* equal-width sections by *n*-1 firmly fixed criteria. (Again, *n* refers to the number of categories the observers have to use. In the present case, with 5 categories, there would be criteria at .2, .4, .6, and .8 along the *t* dimension.) The response given is the category corresponding to the section into which t_i falls.

There are a number of issues associated with these assumptions. With respect to the first assumption, the idea that a physical dimension and a psychological dimension can be related by a power function is well-documented. However, in the classic circumstances (e.g., height, weight, brightness), two things are true which are not necessarily true here: 1) there is a value on the physical dimension which represents an absolute 0, and 2) the psychological representation has a straightforward one-dimensional form. The first of these attributes is absolutely crucial to the existence of a power function relationship. Without a physical 0 point, there can be no subjective 0 point (0 objectively must produce 0 subjectively - a $0^b = 0$), and the best that could be hoped for would be a linear, rather than a power, relationship between the D_is and S_is. With respect to impairment of teletext systems, the notion of 0 physical impairment is problematic. In fact, it would appear to change with the development of new technologies. Thus, this assumption might have problems right from the start. A discussion of how this aspect of the first assumption can be evaluated will be included in section 2.3.3.

The second of these aspects of assumption 1 was discussed to some extent in the previous chapter. The ultimate representation of a stimulus on the S dimension may be achieved fairly directly. That is, even if the initial representation of a stimulus is in a multidimensional space, as long as the way in which the dimensions are handled *does not vary* (e.g., each of j dimensions may be weighted equally), then the relationship between D and S can be considered to be straightforward. If, however, the way in which the multidimensional representation is handled depends on something like task instructions, the same stimulus could give rise to a number of S_is. Thus, a straightforward power function equation simply could not capture the nature of the relationship between D and S.

One other comment should be made about the first assumption. Nothing is being said here explicitly about the shape or variance of the distribution about S_i. Nonetheless, implicit assumptions are being made. What Allnatt has chosen to do is to state these assumptions in terms of the t dimension. Since there is a one-to-one mapping between S and t, either dimension can be used as the vehicle for stating and testing the assumptions. Interestingly enough, however, the shape and variance of the distribution on the t dimension are so complicated that testing the assumptions at that level is not advised either. Instead, as we shall see, Allnatt recommends a second transformation to a T dimension in which the mathematics are simpler and the ease of testing assumptions is greater.

The second assumption really represents the central contribution of Allnatt's technique. If accurate, it specifies the exact relationship between the physical dimension, D, and scale values on the t dimension. It is this t dimension that Allnatt finds most meaningful psychologically. Thus, the scale values we're after here are those representing central tendency on this dimension rather than the S_is. The test of the proposed relationship between D and t will be described in section 2.3.2.

The validity of the final assumption is absolutely crucial to the success of Allnatt's technique. Even if all the earlier assumptions are correct, this assumption must also be correct if the obtained scale values are to be interpretable. Allnatt and Corbett (1972) suggest that this assumption may fail if the stimulus set includes too narrow a range of impairment levels. Allnatt and Corbett (1972) have worked out a set of instructions for analyzing data under those circumstances. In the present circumstances, no problems of this sort should arise since, as recommended in the previous chapter, the stimulus set will always include one very bad and one very good stimulus. Unfortunately, even under the present circumstances, there is no way to test this assumption independent of Allnatt's second assumption.

2.1.2 The T dimension

As noted above, for analysis purposes, neither the *S* dimension nor the *t* dimension is to be used. Instead, a third dimension, *T*, is recommended. Values on this dimension are related to those on *t* via the equation

$$T=\ln\ (t/1-t) \tag{2}$$

This *T* dimension has *absolutely no* psychological relevance or reality. It does not exist in anyone's head nor does it necessarily represent anything in the real world. It's used solely for analysis purposes. It does have a number of properties that Allnatt regards as important. The first is that the range of this dimension is the whole real line rather than 0 to 1 as in the *t* dimension. In much of Allnatt's earlier work, he struggled with ways of modeling the shape and variability of the distribution on the *t* dimension. However, since the dimension is strictly limited at each end the distribution was always, in some sense, truncated, making modeling difficult. On the *T* dimension, no such problem exists and Allnatt (1973) suggests that the distribution function can be modeled by the equation

$$F(T) = \frac{1}{1 + e^{-g(T-T_m)}} \qquad (3)$$

where *g* is a free parameter and T_m is the transform of the median of the *t* distribution for the stimulus under consideration. This proposed distribution function will, of course, need to be tested.

The other nice property is that, because the transform is monotonic, the median of the *T* distribution is the transform of the median of the *t* distribution. Thus, when T_m is found, t_m is determined through the inverse of the transformation in equation (2).

$$t_m = \frac{1}{1 + e^{-T_m}} \qquad (4)$$

The t_ms are the stimulus scale values which we are ultimately attempting to find.

2.2 Calculation Technique

The theoretical rationale for calculating the T_ms was not presented in a very complete way in the previous chapter. Thus, although the purpose of this section is to outline the calculation technique, more attention will be paid to the theory behind the steps than was in the discussion of Thurstonian scaling.

Since the data collection method here is the same as that for Thurstonian scaling, this example will employ the same data set as scaled earlier. Again, stimuli X_0 and X_6 are the anchor stimuli which helped the observers maintain their criterion placements appropriately. The data from these stimuli can again be disregarded. Stimulus X_1 is our referent stimulus which supposedly represents a middle level of quality. However, as notely previously, it was created by combining impairments from a number of dimensions. Thus, it doesn't fit with the other stimuli which will only vary along one dimension. As such, in practice, the only reason to scale it is to make certain that its t_m value remains relatively constant across experiments. Nonetheless, for the present example, it will be assumed to represent an impairment only along the dimension of interest and, thus, it will be scaled to the same end as stimuli X_2, X_3, X_4 and X_5.

The steps in the analysis will be as follows:

1. Create a data matrix like that presented in Table 1.

2. Eliminate the anchor stimuli and turn the matrix into a probability matrix like that in Table 3. (The ordering of the rows is irrelevant. They can be left as they were in Table 1 or ranked as in Table 3.)

3. Turn this matrix into a cumulative probability matrix as in Table 4. The last column will again contain only 1.00s and, thus, can be dropped.

4a. The values in the rows of this matrix plotted against the placements of the four criteria on the *t* dimension (.2, .4, .6 and .8) would give an estimate of the distribution function on *t*. What we want is an estimate of the distribution function on *T*. Thus, the cumulative probability values should be plotted against the *T* transforms of .2, .4, .6, and .8 (i.e., -1.386, -.405, +.405 and +1.386). In either case a best fitting function could then be drawn through these points and the median (*t*m or *T*m) estimated by interpolation. However, since the precise form of these functions is actually specified, the plotting should not be done by eye. Instead, if this approach is to be taken the *T* dimension should be used and the value *g* should be estimated in a way which allows equation (3) to best fit the data. However, since estimating parameters of logistic functions is overly complicated at best, a simpler way to solve for the *T*ms is found in 4b.

4b. This simpler procedure is actually the standard trick for dealing with logistics, taking logarithms in order to produce linear relationships. Beginning with equation (3) if we invert both sides and subtract 1 we obtain

$$\frac{1}{F\ (T\ )} - 1 = e^{-g(T-T_m)}$$

Taking logarithms produces:

$$\ln\ (\ [1\ /\ F(T)]\ -\ 1\ )\ =\ -g\ T + g\ T_m$$

Thus, a plot of the derived values from the left-hand side of this equation should produce a straight line with slope -*g* and intercept *gT*m . Therefore, what we want to do here is to transform the *F*(*T*) values in the three step process of a)inversion (Table 11), b) subtraction of 1 (Table 12), and c) conversion to logarithms (Table 13).

5. For each stimulus, plot the values in Table 13 against the four values of *T* determined previously (i.e., -1.386, -.405, +.405, +1.386). (These plots are contained in Figure 4.)

6. Correlation coefficients, slopes, intercepts and estimates of *g* (*g* is the negative of the slope) should be determined for each plot (see Figure 4).

7. An average *g* value should be determined to produce the most stable estimate (see Table 14).

8. The intercepts should all be divided by *g* to produce estimates of *Tm* (see Table 14).

9. The *Tm* values should be transformed by equation (4) to produce the scale values, *tm*, on the *t* dimension (see Table 14).

10. While the corresponding values on the *S* dimension may also be desired and, in theory, possible to determine, in practice they are unattainable from the present data. As noted in equation (1), the basic relationship between *t* and *S* is:

$$t_{m_i} = \frac{1}{1 + \left(\frac{S_i}{S_M}\right)^{G/b}}$$

While the equation can be used to solve for *Si*, determining its value requires knowledge of S_M and *b*. (S_M (= $aD_M{}^b$) is the scale value for the stimulus whose *tm* value is .5, *b* is the exponent in the power function relationship.) Shortly we will discuss an evaluation technique which provides a value for D_M, however, *a* and *b* can not be solved for until the analysis outlined in section 2.3.3 has been carried out. This analysis will require additional data collection.

2.3 Justifying the Technique

The thing to keep in mind about Allnatt's technique is that it's based on the notion that the stimuli vary along some measureable physical dimension. Thus, the ultimate product of this analysis is a psychophysical function relating D values to t values. If the existence of such a function can be validated our understanding of the psychological impact of stimuli varying along D would be greatly enhanced. Further, as will be discussed later, it may be possible to predict the effects of varying stimuli along two or more physical dimensions concurrently. However, the requirement that the stimuli to be scaled vary only along a single physical dimension does limit us a bit in terms of the nature of the stimulus set that can be scaled in a given analysis.

There are three aspects of Allnatt's technique that do need to be evaluated in order to have confidence in the obtained scale values. These are 1) that the internal sensation S_i is related to the physical stimulus D_i by a power function, 2) that the shape of the distribution on the T dimension is reasonably logistic and 3) that the psychophysical function relating D and t is as specified in equation (1). The second and third of these can be tested using the same data used in the scaling analysis. Evaluating the first is substantially more complicated and needs an additional experiment. As such, it will be discussed last.

2.3.1 Testing the Form of the T Distribution

This test involves an examination of the plots in Figure 4. (Normally, this test will be carried out before any further analysis is undertaken.) If the assumption about the logistic shape of the T distribution (and the placement of criteria) is correct, each plot should be well described by a straight line. The linearity of these relationships can be tested by comparing the values of the correlation coefficients to a criterion value. Obtained values less than the criterion would suggest that the assumption is incorrect.

The correlation coefficients can be found on their respective plots in Figure 4. As with the test of normality in the Thurstonian analysis, a value of .90 should serve as the criterion for four-point plots while a value of .988 should be used as the criterion for three-point plots. In every case, the obtained value of the correlation coefficient is larger than the criterion, suggesting that the assumption is valid.

As before, these tests of distributions are not strong ones. If the number of categories were larger than five, the test would be more powerful. However, in the present circumstances, these plots will never have more than four points. Hopefully, most will have all four points although the test can be performed with a three-point plot. If a plot only contains one or two points the test cannot be performed.

2.3.2 Testing the Form of the Psychophysical Function

The proposed form of the psychophysical function is given in equation (1). If one inverts both sides of this equation and then subtracts one, the following relationship is obtained:

$$\frac{1}{t_{m_i}} - 1 = \left(\frac{D_i}{D_M}\right)^G$$

The expression on the left-hand side has a couple of uses and has been given a designation of its own, J_{m_i}. If we take logarithms of both sides, we next obtain:

$$\ln\ (J_{m_i}) = G \ln D_i - G \ln D_M$$

Thus, a plot of the values ln (J_{m_i}) against ln D_i should produce a straight line with a slope of G and an intercept of - G ln D_M. (The J_m values are contained in Table 14.)

Until now, we've considered stimuli X_1 to X_5 as arbitrary and not as representing particular values on the D dimension. In order to complete the analysis, we will need to specify values for each stimulus on D. As Allnatt (1979) notes this is more difficult than it sounds, "It is, however, sometimes necessary to spend a little time searching for a suitable objective measure of impairment that can be simply related to its subjective effect" (p. 615). For the present example we plan to use signal to noise ratio in decibel units (dB). The problem here is that dB is already a log scale raising the question of whether it is the log of logs we're interested in or the dB values themselves. For our purposes, we'll just use the scale values themselves, arbitrarily assigned to be:

$$X_1 = 25 \text{ dB},\ X_2 = 20 \text{ dB},\ X_3 = 30 \text{ dB},\ X_4 = 35 \text{ dB, and } X_5 = 15 \text{ dB}.$$

(Note: remember, under normal circumstances the referent, X_1, will vary along a number of impairment dimensions. Thus, it would not be included in this analysis. It's included here just to aid in the presentation of the example.)

What we're about to do is plot the logarithms of the J_ms against our dB values which are already expressed in log units. If the model equation is correct, we should observe a straight line with slope C. This plot is contained in Figure 5, with the slope, intercept and correlation coefficient listed on it.

The test once again involves the obtained value of the correlation coefficient. If it is larger than a criterion value, the fit of the model equation is acceptable. The choice of criterion is not as straightforward as in previous analyses because it will depend on the number of points in the plot which will vary with the number of stimuli being scaled. Here, there are 5 stimuli and, thus, 5 points. We therefore have 3 (K-2) degrees of freedom. The criterion value should be .805. The correlation coefficient is greater than .805 . Thus, the model equation seems to be a reasonable one.

One additional thing which can be determined here is the value of $\underline{D}_M$, the stimulus that produces a scale value on the $\underline{t}$ dimension of .5. The intercept of the line in Figure 5 is [- G ln D_M], and $\underline{G}$ is defined by the slope. Thus, because the $\underline{D}$s are already in log units:

$$D_M = \frac{\text{intercept}}{-\ G} = + \frac{(1.8496)}{.0929} = 19.91$$

2.3.3 Examining the Nature of S

The first assumption of Allnatt's technique, the power law assumption, has two important implications. One is that the psychological representation of a stimulus is ultimately unidimensional in a straightforward way. The issues involved in this assumption and the implications of multidimensional representations will be discussed in chapter 5. The other implication, that the physical and hence psychological dimensions have true 0 points and, thus, that ratios on the $\underline{S}$ dimension are meaningful, is the issue to be discussed here.

The technique to be used does not involve the data already collected for the scaling analysis. Instead, it requires data from an independent experiment which should be carried out before applying Allnatt's technique. The experiment allows an assessment of the physical and corresponding psychological dimensions themselves. If successful, it indicates that ratios on these dimensions are meaningful and, thus, substantiates the notion of psychological and physical 0 points.

The technique was developed by Fagot (1978). A number of stimuli (5-7) varying along the physical dimension of interest are selected. Suppose for the present example that five stimuli are selected (call them *a*, *b*, *c*, *d*, and *e* in increasing magnitude). Pairs of these stimuli are presented to an observer whose job it is to produce a ratio of magnitudes of these stimuli on the *S* dimension. These pairs should be presented randomly a number of times to provide stable estimates of the ratios. These mean ratio estimates can then be placed in a table as shown in Table 15.

The first aspect of the data to examine is refered to as the monotonicity rule. Moving both from left to right across columns and from top to bottom within each column, the ratios should decrease monotonically (i.e, successive ratios should be less than their predecessors). In the present situation, there are no violations of this rule.

The second step in the evaluation procedure is a bit more complicated. All tetrads of the stimuli are listed and (labelling the stimuli in a tetrad 1, 2, 3 and 4 in ascending order) the quantities $R_{14} \times R_{23}$ and $R_{13} \times R_{24}$ are calculated. (R_{XY} is simply the ratio from the data matrix relating stimulus X to stimulus Y.) The resulting ratio products for the example are listed in Table 16.

Demonstrating that the physical dimension has ratio properties is essentially a matter of demonstrating that the ratio products in the two columns of Table 16 are identical. Calculation of a correlation coefficient would be inappropriate here because it would be insensitive to certain types of differences between the two columns (e.g., if the two columns differed by a constant). Instead, a test developed by Bartko (1976) can be used. A one-way repeated-measures analysis of variance (ANOVA) is carried out on the ratio products in Table 16 treating the tetrads as observers and the columns as two levels of an independent variable. The ANOVA table is presented in Table 17.

First an F ratio for columns is calculated. It should be nonsignificant indicating no overall difference between the two columns. Here, the F value is 6.77 while the .05 cutoff for 1 and 4 degrees of freedom is 7.71. Thus, no problems have arisen yet. (In actuality, to maximize the possibility of finding nonratio scale tendencies, it would be better to be more liberal here and use an alpha level of .10. The criterion F value would then be 4.54 meaning that the test would fail. However, for demonstration purposes, we'll assume our test has succeeded so far and continue.)

The final step is to produce an *F* ratio for tetrads. However, the one shown in the ANOVA table is not the one we're looking for. To create the proper *F* ratio we use the fact that we've failed to find an effect for columns and assume that its mean square represents only error. Thus, a new error mean square is created by pooling the sums of squares and degrees of freedom from the column effect with those for error. The resulting pooled mean square (MS_p) is indicated in Table 17. This value is then used to create the *F* ratio for tetrads as shown in the table. The resulting *F* value of 20.05 is compared against the .05 criterion for 4 and 5 degrees of freedom of 5.19 (actually, here it might be better to be a bit more conservative and use an alpha of .01, that value is 11.39). In any case, the obtained value far exceeds either criterion indicating that the variance in the data in Table 17 is almost entirely due to differences between tetrads and not to differences between columns or random error. Thus, we can conclude that, for our purposes, the two columns match and that the creation of a ratio scale is possible.

Whenever this test is undertaken, it will be important to calculate both *F* ratios. In order to validate ratio scaling, we have to show both that there is no overall difference between columns (the first *F* test) and that error variance plays a minor role in the overall variability (the second *F* test). Thus, the point becomes that most of the variance must be due to differences between row means as it was in this example. A measure of the proportion of variance due to the differences between rows (Bartko, 1976) is given by:

$$\frac{MS_{tetrads} - MS_{pooled}}{MS_{tetrads} + (C-1)\ MS_{pooled}}$$

where *C* is the number of columns. Here the value is .905 indicating that row differences account for 90.5% of the variance. The remainder of the variance (i.e., 9.5%) is attributable to column differences and error, two factors which, for the present data set, we have concluded are unimportant.

Successful completion of the test described above indicates that the data satisfy Fagot's (1978) minimum requirement for the creation of a ratio scale (what he calls C3). To complete this evaluation we next must determine whether the scale values are reasonably well described by a power function. There are two ways of accomplishing this. One way would be carry out a magnitude estimation experiment using the dimension of interest. The only caveat here is that the stimulus which is used as the standard should be more intense than any of the comparison stimuli (in the present case, this means it should have a greater amount of impairment). Thus, the task would actually be a fractionation task. (The reason we would have to use the greatest magnitude stimulus as the standard is because satisfying Fagot's C3 requirement only guarantees that we can generate a ratio scale under this specific condition.) The second way to do this would be to recognize that a fractionation experiment has already been done within the context of the study reported in Table 15. That is, the values in the rightmost column of Table 15 are exactly the data needed here. In each case a comparison of less intensity has been compared to the highest magnitude stimulus in the set and a ratio judgment has been given.

The analysis to be done on these data is a standard one. According to the power law equation:

$$S_i = a\,D_i^{\,b}$$

So,

$$\ln S_i = \ln a + b \ln D_i$$

Thus, if the logarithms of the values in the rightmost column of Table 15 are plotted against the logarithms of *D*, we should observe a straight line with slope *b*. These data are plotted in Figure 6. (Remember, the D_is are already in log units.)

The fit of the straight line to the points is again evaluated using a correlation coefficient which in this example is .9909. The criterion for evaluation would be the one-tailed cutoff for K-3 degree of freedom (K is the number of stimuli so here K-3=2). This value is .900. Thus, the fit appears to be a good one. Based on this result and the success of the preceding analysis, the assumption of a power function relationship between D and S seems to be a reasonable one. Also note that this analysis gives us an estimate of .13 for b, the exponent in the power function (and the slope of the best-fitting line).

2.4 Additivity of Effects

Once we have verified the applicability of Allnatt's procedure to more than one physical dimension, the question of coexisting impairments arises. Allnatt (1979) argues that the effects of coexisting impairments are additive in their J_m values (remember $J_m = 1/t_m - 1$). As far as we can tell, he has no theoretical basis for this claim. He does, however, produce one empirical result which supports his position.

In his demonstration, random noise and long-delayed echo were selected as the two dimensions of interest. For both dimensions, single impairment source stimuli were scaled and the relationship given in equation (1) was validated. (No attempt was made, however, to validate the ratio scale assumption.) Predictions were then made for stimuli containing coexisting impairments by adding the J_m scores appropriate to the level of impairment on each dimension to get a total J_m. This score was then reconverted to a t score by the inverse of the J transformation:

$$t_m = 1/[1+J_m]$$

These scores represent the predicted t_ms which were then compared to the t_ms obtained in the scaling procedure. The results, reported graphically, suggest a good match between predicted and observed t_ms.

Allnatt (1979) also reports a similar investigation carried out by the British Broadcasting Corporation using a couple of additional impairment dimensions. Apparently, this analysis was also successful, suggesting that the technique may hold a certain amount of promise. However, as Allnatt (1979) notes, if the effects of impairment on two physical dimensions are visually similar, the additive rule will not hold. Thus, before additivity is ever assumed for any two dimensions, it should be evaluated empirically.

The choice of the optimal procedure for evaluating the match between the observed and predicted t_ms is not clearcut. What we are trying to do is take a large number of observed t_m values and determine whether they equal their corresponding predicted values. Presumably a factorial design will be used in which stimuli are created by crossing p levels of dimension 1 with q levels of dimension 2. Thus, the number of observed t_m

values will be $p \times q$. (There will, of course, be the same number of predicted t_m values.) Our suggestion here is that these values be arranged in two columns such that one column contains the observed t_ms and the other the predicted t_ms. The same analysis as was carried out on the values in Table 16 can then be carried out here. Again, there should be a nonsignificant F when testing the difference between columns and a highly significant F when testing the difference between rows (i.e., stimuli) with the row test being carried out using the pooled mean square. If so, the claim can then be made that the two dimensions do combine additively.

If three or more dimensions are used, the same evaluation procedure should be followed. The only difference is that the number of stimuli and, hence, the number of rows would increase.

III. SUMMARY

For both Thurstone's and Allnatt's techniques we have now completed our discussion of the three issues set out in this chapter's first paragraph: a) the assumptions underlying the technique; b) the nuts and bolts of the analysis procedure; and c) the method of determining whether the technique can be legitimately applied. In reference to the first issue, Thurstonian scaling has fewer and less strict assumptions. The assumptions involve the placement of normal distributions and criteria on the S dimension. Allnatt's set of assumptions involves assumptions about a) the nature of the relationship between the physical dimension, *D*, and *S*, b) the transformation of the *S* dimension to produce the *t* dimension and c) the shape of the distributions and the criterion placements on *t*. In reference to the second issue, the procedures both are fairly cut and dried in that they can be implemented in a straightforward, step-by-step procedure. In reference to the third issue, both techniques allow each of their assumptions to be tested as outlined in the chapter. In chapter four our discussion will turn to ways of analysing our results, that is, the scale values we have discovered. However, because the reader may have missed it, we should first mention an interesting phenomenon.

For both techniques, issues b) and c) were discussed by using sample data and going through the procedures in a step-by-step fashion. The data, of course, were totally fabricated. Nonetheless, both techniques passed all tests with flying colours. As such, one must realize a couple of things. First, at best, these tests of the assumptions are not powerful. It appears unlikely that either technique will be invalidated very often. Second, validation of one technique definitely does not invalidate the other. In most cases both may be reasonable. Therefore, in circumstances in which both can be applied (i.e., when the stimuli vary on a single physical dimension), the choice of technique may be somewhat arbitrary. For the present we just have to accept the fact that you can never tell which way the train went just by looking at the tracks.

CHAPTER 4 - ANALYSING SCALE VALUES

The previous chapter contained a complete discussion of the issues which should be considered when selecting one of the recommended scaling techniques over the other. Assumptions and how to test them were layed out for both techniques and the procedures for generating scale values were detailed. Regardless of which scaling technique was selected, its ultimate purpose was to produce scale values for each stimulus in the experiment. Once this has been done and all relevant tests of assumptions have been completed, additional analyses may be performed. If the Thurstonian technique were used one might wish to determine which stimuli have scale values significantly different from the scale values of the other stimuli. One might also wish to determine whether the scale values vary as a function of different experimental conditions. For example, if the data were collected under 2 or 3 different sets of instructions, the nature of the instructions could have a strong influence on the obtained scale values. Thus, it would be important to analyse the effects of such a variable.

The preferred method of analysis would be the ANOVA. If the question is whether the scale values are significantly different from one another, a simple one-way ANOVA could be carried out with stimuli as the single factor. If one or more additional factors are introduced, a standard multifactor ANOVA would be preferred and interactions could be examined. Planned and/or post hoc comparisons would also be useful.

There are, of course, two major assumptions involved in doing these types of analyses. The first is the normality assumption. Violations of this assumption should not be a problem here, however, since tests for normality have already been performed. If the normality assumption were incorrect, we wouldn't have reached this point. The second assumption is that of homogeneity of variance. The remainder of this chapter will be devoted to a discussion of this issue.

In Allnatt's technique one would not need to test scale values to determine whether they are different from one another. The goal of his technique is to specify mathematically how scale values vary as the physical dimension varies. If successful, the question of which stimuli are different from which has already been answered. However, the question of whether the scale values vary as a function of additional factors would be relevant. To answer this question, we would not recommend testing the scale values directly, however. Instead, it seems more reasonable to test the parameters that give birth to the scale values, $\underline{C}$ and $\underline{D}_M$.

Once again, the ANOVA is the preferred method of analysis. With respect to the normality assumption, the sampling distributions of $\underline{C}$ and $\underline{D}_M$ may not be normal. However, here the central limit theorem can come to our aid. As long as the number of observers per condition is more than twenty, the normality assumption should be satisfied. Since we have previously suggested using up to fifty observers there should be no problems here. Again, however, there could be problems with the homogeneity of variance assumption. It is to that issue that we now turn.

I. HETEROGENEITY OF VARIANCE

In the following examples we will be assuming that the Thurstonian scaling technique has been carried out on K stimuli. The data to be analyzed are the scale values for each stimulus. The expression, n_i, represents the number of scale values produced for stimulus i. This value is equal to the number of observers who have scaled stimulus i. Under normal circumstances all observers will scale all stimuli, meaning n_i will be a constant. However, for completeness sake, the following discussion will include examples in which the n_is are not assumed to be equal.

For present purposes, the most important assumption underlying the ANOVA is that the populations from which the K samples are drawn have equal variances. As Kirk (1982) has pointed out, none of the ANOVA's assumptions are ever fully satisfied by real data; the important questions to ask are what effects do violations of the assumptions have on the significance levels and the power of the test, and what measures can be taken to deal with those assumptions which are found to have been violated.

A number of studies have demonstrated that the ANOVA is relatively robust to violations of the assumption of equal variances, *provided the samples have equal sizes* (Glass, Peckham, and Sanders, 1972). The exact effects of unequal variances are difficult to calculate, but, generally, the actual significance level of the test is somewhat higher than the nominal level. Thus, the analysis is more likely to result in a Type I error (i.e., a false rejection of the null hypothesis). A summary of the effect of unequal variances on alpha -- the probability of a Type I error -- is shown in Table 18. Here, the nominal alpha is .05 and, as can be seen, except for the case of many samples (K = 7) and small sample size (n_i = 3), inequality of variances has only a small effect on the actual alpha. However, most post-hoc multiple-comparison procedures (e.g., Tukey Honestly Significant Difference, Newman-Keuls, Duncan Multiple Range, Fisher Least Significant Difference, and Scheffé) also require homogenous variances, and the robustness of these tests to heterogeneous variances is unknown, even with equal sample sizes (Games, Winkler, and Probert, 1972). With respect to the power of ANOVA, calculations are impossible if the assumption of equal variances is violated since there would be no true population variance (σ^2). Budescu (1982) presents a method for computing the approximate power, but this procedure has not received widespread acceptance.

If sample sizes are unequal -- a situation that should not arise in the present circumstance -- unequal variances can have a serious effect. The direction of this effect depends upon the magnitudes of the variances of the larger samples relative to those of the smaller samples. The mean square error (MSE) -- the denominator of the F ratio in one-way fixed-effects ANOVA -- is a weighted average of the sample variances with greater weight being placed on the variances of the larger samples. Thus, if larger samples have larger variances, MSE will be inflated and the probability of a Type II error (failure to reject a false null hypothesis) will be increased. Conversely, if larger samples have smaller variances, MSE will be smaller than it should be and the Type I error rate (false rejection of the null hypothesis) will be increased (Lindman, 1974).

In this chapter, attention will first be paid to statistical tests designed to detect whether the assumption of equality of variances has been violated. Following this, a number of procedures will be described whose purpose is to equate initially unequal variances or to compensate for unequal variances in other ways. Finally, a section is devoted to analytic procedures which can be used instead of the ANOVA.

1.1 Tests of Equality of Variances

If sample sizes are equal, several authors recommend that the equality-of-variance assumption of ANOVA not be tested (e.g., Keppel, 1982). This is because, as noted, a violation of the assumption has only a minor effect unless the n_i are very small. Somewhat surprisingly, two of the most commonly cited tests of equality of variances -- Hartley's and Cochran's -- can only be performed when sample sizes are equal or nearly equal. These tests' popularity is due to their computational simplicity and, for this reason they are described below. Two additional tests are also described -- Bartlett's and the Box-Scheffé -- which, while computationally more laborious, can be applied when sample sizes are unequal. The Box-Scheffé test has an additional advantage which will be discussed when the test is described. Two simple numerical examples will be employed to illustrate the use of the tests. In each case, the null hypothesis:

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_K^2$$

will be tested against the alternative hypothesis that at least one variance is significantly different from the others.

1.1.1 Hartley's F_{max} Test

The test statistic for this test, F_{max}, is computed as the ratio of the largest sample variance divided by the smallest sample variance. Consider the following sample data. Here four stimuli have been scaled ($K = 4$) by six observers ($n_i = 6$):

$$n_1 = 6 \qquad n_2 = 6 \qquad n_3 = 6 \qquad n_4 = 6$$

$$s_1^2 = 12 \qquad s_2^2 = 8 \qquad s_3^2 = 25 \qquad s_4^2 = 6$$

The largest variance, s_3^2, is 25, the smallest, s_4^2, is 6, so the test statistic:

$$F_{max} = \frac{25}{6} = 4.167$$

The critical value, F, is defined by K (the number of stimuli or samples), $n_i - 1$ (the degrees of freedom associated with each sample variance), and $1 - \alpha$, and can be found in Table B.7 (copied without permission from Winer, 1971). With $\alpha = .05$, the critical F, $F_{4,5,.95}$, is 13.7. Since the obtained test statistic (4.167) does not exceed this critical value, H_0 is not rejected and equality of variances can be assumed. (Note: assuming equality of variances is tantamount to accepting the null hypothesis, a dubious practice unless beta -- the Type II error probability -- is known, but one which most researchers seem willing to adopt in this case.)

1.1.2 Cochran's Test

In this test, the test statistic, C, is computed as the ratio of the largest sample variance to the sum of all K sample variances. Using the same sample data as above, the largest sample variance is S_3^2 (= 25), and the sum of all the variances is 51, so the test statistic:

$$C = \frac{25}{51} = .4902$$

The critical C value is defined by K, $n_i - 1$, and $1 - \alpha$, and can be found in Table B.8 (copied without permission from Winer, 1971). With $\alpha = .05$, the critical C, $C_{4,5,.95}$, is .5895. Again, since the obtained test statistic does not exceed this critical value, H_o is not rejected and equality of variances can be assumed.

With respect to both of the above tests, Winer (1971) notes that if sample sizes are nearly equal, the largest n_i can be used to determine the appropriate degrees of freedom (n_i - 1) with which to enter the critical values Tables. In most cases, the Hartley's and Cochran's tests will lead to the same decision although, given that Cochran's procedure employs more of the information in the sample data, generally it is slightly more sensitive than Hartley's test.

1.1.3 Bartlett's Test

This test is more appropriate than either Hartley's or Cochran's procedure if sample sizes are unequal. It is also more powerful. It should not be used if any n_i is smaller than 3, and most n_i's should be greater than 5. As will be seen, it is more complex computationally than the previous tests but in many cases it is the preferred procedure.

The test statistic for this procedure is a chi-square statistic, computed as:

$$\chi^2 = \frac{2.303}{C}\left[(N - K)\log_{10}MSE - \sum_{i=1}^{K}(n \quad - 1)\log_{10}S_i^{\,2}\right]$$

Where: 2.303 is a constant

K = number of samples (stimuli),

N = total number of observations across all K samples,

MSE = mean square error (defined below),

$n_i - 1$ = degrees of freedom associated with each sample variance (S_i^2),

and $$C = 1 + \frac{1}{3(K - 1)}\left[\sum_{i=1}^{K} \frac{1}{n_i - 1} - \frac{1}{N - K}\right].$$

Using the same data as before, C is computed as:

$$C = 1 + \frac{1}{3(3)}\left[\frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} - \frac{1}{20}\right]$$

$$= 1.0833$$

$$MSE = \sum_{i=1}^{K} \frac{(n_i - 1)s_i^2}{N - K} = \frac{5(12) + 5(8) + 5(25) + 5(6)}{20} = 12.75$$

$$\log_{10} MSE = \log_{10} 12.75 = 1.1055$$

$$\log_{10} s_1^2 = \log_{10} 12 = 1.0792$$

$$\log_{10} s_2^2 = \log_{10} 8 = .9031$$

$$\log_{10} s_3^2 = \log_{10} 25 = 1.3979$$

$$\log_{10} s_4^2 = \log_{10} 6 = .7782$$

and, $$\chi^2 = \frac{2.303}{1.0833}\Big[(20)(1.1055) - \big((5)(1.0792) + (5)(.9031) + (5)(1.3979) + (5)(.7782)\big)\Big]$$

$$= \frac{2.303}{1.0833}\,[22.11 - 20.7920]$$

$$= 2.8020$$

The critical χ^2 is defined by K - 1 and alpha. In this example, the critical χ^2, $\chi^2_{3,.05}$, is 7.81. Since the obtained value of the test statistic, 2.804, does not exceed the critical value, H_o is not rejected and equality of variances can be assumed.

1.1.4 Box-Scheffé Test

Each of the previous three tests -- Hartley's, Cochran's, and Bartlett's -- provides a valid test of equality of variances *if the underlying population distributions are normal.* However, if the normality of the distribution is either unknown or if it is known to be nonnormal, these tests are inappropriate (Box, 1953; Martin and Games, 1977; Games, Keselman, and Clinch, 1979). If they are used when the distribution is not normal, the null hypothesis of equal variances may be falsely rejected and researchers may wrongly believe that they cannot proceed with an ANOVA. This is especially problematic since the ANOVA itself is relatively insensitive to departures from normality. (In the present circumstances, however, a test for normality will already have been successfully carried out. Thus, the following discussion is only for the sake of completeness.)

A test first proposed by Box (1953), later modified by Scheffé (1959), and now referred to as the Box-Scheffé procedure, can be used whenever nonnormality is suspected. It can also be used when sample sizes are unequal and, although it is computationally quite laborious, it is probably worth the effort. To describe this procedure, consider the results from another scaling study using $K = 4$ stimuli, each scaled by $n_i = 8$ observers, as shown in Table 19.

The first step in the procedure is to divide the n_i observations in each sample randomly into a number of subsamples. According to Games, Keselman, and Clinch (1979), the optimum size of the subsamples (n_j) is the nearest integer value to $(n_i)^{1/2}$. In this example, with $n_i = 8$, each subsample should consist of $(8)^{1/2}$, or 3 observations. Thus, the 8 observations in each sample are randomly divided into 3 subsamples of size 3, 3, and 2, as shown in Table 20. (Note: for smaller n_i, the use of $n_j = 2$ for subsamples is not recommended (Gartside, 1972; Games, Keselman, and Clinch, 1979), since this will result in considerably less power than with subsamples of intermediate size.)

The second step is to compute the variances of each of the subsamples (Table 21) and to convert these variances into natural logarithms (Table 22). Finally, as illustrated below, an ANOVA is performed on these logarithms to test the original equality of variance hypothesis.

For the data in Table 22, find the weighted means ($X_{.i}$) for each group, weighting by the subsample variance degrees of freedom (i.e., $n_i - 1 = \gamma_{.i}$). Thus, for group 1:

$$\bar{X}_{.1} = \frac{2(-1.0986) + 2(-1.0986) + 1(1.5041)}{(2 + 2 + 1)}$$

$$= -.5781 \text{ (with } \gamma_{.1} = 5)$$

In like manner, compute:

$$\bar{X}_{.2} = -1.0175 \text{ (with } \gamma_{.2} = 5)$$

$$\bar{X}_{.3} = -1.0175 \text{ (with } \gamma_{.3} = 5)$$

$$\bar{X}_{.4} = .0085 \text{ (with } \gamma_{.4} = 5)$$

And the grand mean of all the cells, $\bar{\bar{X}}_{..}$:

$$\bar{\bar{X}}_{..} = \frac{\sum_{i=1}^{K} \bar{X}_{.i}}{K} = \frac{(-.5781) + (-1.0175) + (-1.0175) + (.0085)}{4}$$

$$= -.6512 \text{ (with } \gamma = 20)$$

Then compute the following 3 statistics:

$$\text{I} = \gamma \, \bar{\bar{X}}_{..}^{2} = 20(-.6512)^2 = 8.4812$$

$$\text{II} = \sum_j \sum_i \gamma_{ji} \bar{X}_{ji}^{2} = 2(-1.0986)^2 + 2(-1.0986)^2 + \ldots + 1(-.6931)^2$$
$$= 24.9005$$

$$\text{III} = \sum_{i=1}^{K} \gamma_{.i} \bar{X}_{.i}^{2} = 5(-.5781)^2 + 5(-1.0175)^2 + 5(-1.0175)^2 + 5(.0085)^2$$
$$= 12.0244$$

Finally, compute the sums of squares for treatment (SST) and for error (SSE):

$$SST = III - I = 12.0244 - 8.4812 = 3.5432$$

$$SSE = II - III = 24.9005 - 12.0244 = 12.8761$$

and corresponding mean squares:

$$MST = SST/(K - 1) = 3.5432/3 = 1.1811$$

$$MSE = SSE/K(n' - 1) = 12.8761/(4 \times 2) = 1.6095$$

where n' is the number of subsamples per group,and the test statistic, F:

$$F = MST/MSE = 1.1811/1.6095 = .7338 .$$

The critical F , $F_{3,8,.05}$ for this example, is 4.07. Since the obtained test statistic does not exceed this critical value, H_o is not rejected and the variances can be assumed to be equal.

Clearly, this procedure is considerably more laborious than either Hartley's or Cochran's test and, if the treatment populations are known to be normal and the sample sizes are equal, these are the recommended tests (Games, Winkler, and Probert, 1972; Church and Wike, 1976; Keppel, 1982). If normality can be assumed but the sample sizes are unequal, Bartlett's test is recommended. If normality cannot be assumed, the Box-Scheffé procedure is preferred.

Note that one disadvantage of the Box-Scheffé procedure results from the initial random assignment of sample observations to subsamples. Conceivably, different researchers could produce different subsamples and could reach different conclusions from their analyses (Games, Keselman, and Clinch, 1979). An alternative procedure, called the jackknife test (Miller, 1968), also divides observations into subgroups but has the advantage that all users will obtain the same results with the same data. In this procedure, the original $\underline{n}_i$ observations are divided into $\underline{n}_i$ subgroups, each with $\underline{n}_i$ - 1 observations (i.e., one observation is dropped in each subgroup). This procedure has been shown to have greater power than the Box-Scheffé test but it also results in an inflated alpha (Martin and Games, 1977). The Box-Scheffé test maintains alpha close to its nominal value and affords reasonable power (Keppel, 1982) so, overall, it is the recommended procedure (see Brown and Forsythe, 1974, however, for a contrary opinion and another test of equality of variances).

1.2 Procedures for Equating Sample Variances

Assuming one of the above tests has been applied and the null hypothesis of equal variances has been rejected, it may still be possible to perform an ANOVA. A number of procedures exist for transforming the original scores ($\underline{X}_{ij}$) to scores ($\underline{Y}_{ij}$) whose scale has more desirable statistical properties. In this section, four of the most commonly-used transformation procedures are described, followed by a brief discussion of their advantages and disadvantages. Finally, some other correction procedures are described which can be adopted in certain ANOVA designs.

1.2.1 Data Transformations

1.2.1.1 Square-Root Transformation

In certain distributions, sample variances are proportional to the sample means (e.g., the Poisson distribution, in which $\sigma^2 = \mu$). (Such a distribution frequently occurs when the data represent frequency counts of events which have small probabilities of occurrence.) If such is the case, transformed scores (Y_{ij}) can be computed from the original scores (X_{ij}) as:

$$Y_{ij} = \sqrt{X_{ij}}$$

If any X_{ij} is less than 10, a more appropriate transformation is given by:

$$Y_{ij} = \sqrt{X_{ij} + 5}$$

or,

$$Y_{ij} = \sqrt{X_{ij}} + \sqrt{X_{ij} + 1}$$

The latter transformation has been recommended by Freeman and Tukey (1950), and tables for the transformation are available in Mosteller and Bush (1954).

1.2.1.2 Logarithmic Transformation

If standard deviations rather than variances are proportional to the sample means -- as often occurs if the data are positively skewed -- an appropriate transformation is:

$$Y_{ij} = \log_{10} X_{ij}$$

or, $$Y_{ij} = \log_{10} (X_{ij} + 1)$$

The latter is particularly effective when some of the original X_{ij} are equal to zero or are very small (Kirk, 1982).

1.2.1.3 Reciprocal Transformation

If the data are skewed such that the sample variances increase as a monotonic nonproportional function of increasing sample means, as an alternative to the logarithmic transformation, an appropriate transformation may be:

$$Y_{ij} = 1/X_{ij}$$

or, $$Y_{ij} = 1/(X_{ij} + 1)$$

The latter should be used if any of the X_{ij} are equal to zero (Kirk, 1982).

1.2.1.4 Arcsin Transformation

Although such would not occur in the present circumstances, if the X_{ij} are proportions, then the sample variances will almost certainly differ from one another as a function of the sample means. In this case, an appropriate transformation is:

$$Y_{ij} = \arcsin \sqrt{X_{ij}}$$

which should make the variances approximately equal, independent of the means.

The major advantage of these procedures is that, under different conditions, they may equalize (or approximately equalize) initially unequal variances, thereby satisfying the assumption for ANOVA. The major disadvantage, however, is that the sample means will also be transformed and, thus, inferences regarding treatment effects must be made with respect to the transformed data. Clearly, this could produce results which are a bit hard to interpret. In the case of proportions, for example, the statement that the means of the proportions differ across samples is easy to understand. In contrast, the statement that the means of the arcsins of the square roots of the proportions differ across samples is considerably harder to interpret (Lindman, 1974). A second disadvantage is that for some data, none of these transformations has a noticeable effect (e.g., Wike and Church, 1982). In many cases, however, the transformations will not only equate variances but also may minimize skew. In such cases the use of transformed scores would increase the power of the ANOVA (Levine and Dunlap, 1982, 1983; but see Games, 1983 for a contrary opinion).

Finally, transformed scores may not be advisable in 2-way (or higher order) ANOVA designs. Transformations can have large and undesirable effects on the nature and size of interaction effects; in fact, interactions which really did exist in the original data may now fail to be significant (Lindman, 1974).

1.2.2 Statistical Corrections

A number of procedures have been proposed to overcome the effect of unequal variances on the actual significance levels in the ANOVA. Attention will be given to some of these in this section.

1.2.2.1 Box Correction

The Box correction (Box, 1953) is applicable in ANOVA designs with random effects (1-way or higher-order), or, in fact, in any design for which the denominator of the *F* ratio is an interaction mean square. The procedure allows the researcher to estimate the actual significance level of an obtained *F* test statistic (or, more properly, a range of values within which the actual significance level falls). In the procedure, the numerator and denominator degrees of freedom (γ_1 and γ_2 , respectively) are divided by a constant, *c*. The exact value of *c* is unknown, since its magnitude depends on the extent to which the equal-variances assumption has been violated, but its upper and lower limits are γ_1 and 1, respectively. Thus, the actual significance level of the *F* test statistic can be found twice: once with *c* set to 1 and degrees of freedom γ_1 /1 and γ_2/1 (i.e., the regular degrees of freedom) and once with *c* set to γ_1 and degrees of freedom, γ_1 / γ_1 (cr 1) and γ_2 / γ_1 . Note that the value of the *F* test statistic remains the same under both conditions but its level of significance will change, as illustrated in the following example.

Suppose γ_1 is 3, γ_2 is 15, and the *F* test statistic is 5.25. In the regular procedure (using $\gamma_1 = 3$, $\gamma_2 = 15$), an *F* of 5.25 has a significance level, *p*, of .02. With the correction degrees of freedom ($\gamma_1 = 1$, $\gamma_2 = 15/3 = 5$), an *F* of 5.25 has a significance level of .08. The researcher can thus conclude that the actual significance level of the test is between .02 and .08, even though the assumption of equal variances has been violated. If the violation were relatively small, the true significance level would be close to .02. If the violation were large, the true significance level would be closer to .08.

1.2.2.2 Geisser-Greenhouse Correction

This procedure (Geisser and Greenhouse, 1958) is applicable in repeated measures ANOVA designs, and also involves a modification of the degrees of freedom. The obtained *F* test statistic is evaluated against a new critical value that assumes maximum heterogeneity of variance. The procedure is quite simple: the original γ_1 and γ_2 are divided by a factor equal to the degrees of freedom associated with the repeated factor (or factors). As an example, suppose we have $n_i = 6$ observers who have each scaled $K = 4$ stimuli. In the regular repeated measures ANOVA, the treatment degrees of freedom (γ_1) is $K - 1$ or 3, the observer degrees of freedom is $n_i - 1$ or 5, and the interaction (or error) degrees of freedom (γ_2) is $(K - 1)(n_i - 1)$ or 15. If the variances are equal, the appropriate critical value for treatments is $F_{3,15}$. However, if the variances are not equal, γ_1 and γ_2 are divided by 3, giving a critical value of $F_{1,5}$. If the test statistic *F* exceeds the corrected critical *F*, the results may be considered significant although the variances are unequal.

1.2.2.3 Approximate F test for Planned Comparisons

Instead of performing an overall ANOVA, a researcher may wish to test a number of specific hypotheses which were planned prior to data collection. The principal advantage of planned comparisons is that they can be one-tailed (if desired) and, consequently, their critical values are usually smaller than those of most post-hoc procedures.

Consider the following data for $K = 4$ random samples, each consisting of $n_i = 8$ observers:

$n_1 = 8$	$n_2 = 8$	$n_3 = 8$	$n_4 = 8$
$\bar{X}_1 = 6.75$	$\bar{X}_2 = 10.375$	$\bar{X}_3 = 8.625$	$\bar{X}_4 = 13.75$
$S_1^2 = 8.214$	$S_2^2 = 8.839$	$S_3^2 = 9.696$	$S_4^2 = 2.796$

Here, the MSE turns out to be 7.386.

First, assume that the variances are statistically equal and consider a comparison of the mean of group 1 with the average of the means of groups 2, 3, and 4. The contrast coefficients (C_i) for this comparison are 3, -1, -1, and -1, for groups 1 to 4 respectively. Thus, the null hypothesis is,

$$H_0 : (3)\mu_1 + (-1)\mu_2 + (-1)\mu_3 + (-1)\mu_4 = 0$$

The F ratio for this contrast is given as:

$$F = \frac{\hat{\psi}^2}{SE_{\hat{\psi}}^2}$$

where, $\hat{\psi} = \sum_{i=1}^{K} C_i\bar{X}_i = (3)(6.75)+(-1)(10.375)+(-1)(8.625)+(-1)(13.75)$

$$= -12.5$$

so, $\hat{\psi}^2 = (-12.5)^2 = 156.25$

and, $SE_{\hat{\psi}}^2 = MSE \sum_{i=1}^{K} (C_i^2/n_i)$

$= 7.386(3^2/8 + -1^2/8 + -1^2/8 + -1^2/8)$

$= 11.079$

Thus, $F = 156.25/11.079 = 14.103$

This F value would be evaluated against a critical $F_{1,N-K} = F_{1,28} = 4.20$ with $\alpha = .05$.

Now, assume that the variances have been found to be unequal, and consider the same comparison, now tested through the approximate F procedure.

As before,

$$\hat{\psi} = \sum_{i=1}^{K} C_i \bar{X}_i = -12.5$$

and,

$= 156.25$

$$SE_{\hat{\psi}}^2 \text{ now } = 1/n \sum_{i=1}^{K} [C_i S_i^2]$$

$$= 1/8\ [(3)^2(8.214) + (-1)^2(8.839) + (-1)^2(9.696) + (-1)^2(2.796)]$$

$$= 11.907$$

and, $F = 156.25/11.907 = 13.122$

(Note: in this example n_i is a constant, the term n is being used to refer to this common value of n_is.)

The approximate F ratio is thus somewhat smaller than the original F (this may not always happen), but, more important, it has to be evaluated against a critical F with 1 and γ_2' degrees of freedom, where,

$$\gamma_2' = (n^3 - n^2)(SE\hat{\psi})^4 / \sum_{i=1}^{K} c_i^4 s_i^4$$

$$= (8^3 - 8^2)(11.906)^2 / \left[(3)^4(8.214)^2 + (-1)^4(8.839)^2 + (-1)^4(9.696)^2 + (-1)^4(2.796)^2\right]$$

$$= 11.349$$

which is always rounded down (in this case to 11).

The critical F, $F_{1,11}$, is 4.84 with $\alpha = .05$.

The main difference, in this example, is the loss in degrees of freedom. Note that this is only an *approximate* test. When the n_i are fairly large, the approximation is quite good. For small n_i, the procedure is not recommended. If sample sizes are unequal, the approximate $SE_{\hat{\psi}}^2$ and γ_2' are computed by:

$$SE_{\hat{\psi}}^2 = \sum_{i=1}^{K} C_i^2 S_i^2/n_i$$

and, $$\gamma_2' = SE_{\hat{\psi}}^4 \sum_{i=1}^{K} [C_i^4 S_i^4/(n_i^3 - n_i^2)]$$

1.2.2.4 Welsh-Aspin Test

When $K = 2$, the usual procedure for testing the significance of the difference between means is the two-sample t test, although an ANOVA can be performed if it is preferred. Like the ANOVA, the t test requires the assumption of equal variances and, if the assumption is violated, the following procedure -- the Welch-Aspin test -- can be used instead. It can also be modified and used as an alternative to the usual post-hoc procedures which would follow an ANOVA.

Consider the following 2-sample data, for which

$n_1 = 16$ $\qquad$ $n_2 = 21$

$\bar{X}_1 = 946.50$ $\qquad$ $\bar{X}_2 = 931.50$

$$S_1^2 = 437.48 \qquad S_2^2 = 150.25$$

The test statistic, *t**, is computed in a manner similar to the usual two-sample *t* statistic, except that individual sample variances are used in the denominator rather than a pooled variance.

Thus,
$$t^* = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$t^* = \frac{946.50 - 931.50}{\sqrt{\frac{437.48}{16} + \frac{150.25}{21}}}$$

$$= \frac{15}{\sqrt{34.497}}$$

$$= 2.554$$

If this test had been a regular *t* test, the critical *t* would have 35 degrees of freedom, and would be 1.645. For the Welch-Aspin, the degrees of freedom, γ *, are computed as:

$$\gamma^* = \frac{(n_1 - 1)(n_2 - 1)}{(n_1 - 1)(1 - C)^2 + (n_2 - 1)C^2}$$

where, $$C = \frac{s_1^{\,2}/n_1}{s_1^{\,2}/n_1 + s_2^{\,2}/n_2}$$

$$= \frac{437.48/16}{34.497}$$

$$= .7926$$

Thus, $$\gamma^* = \frac{(15)(20)}{15(1 - .7926)^2 + 20(.7926)^2}$$

$$= \frac{300}{13.21}$$

$$= 22.712$$

which is always rounded down (in this case to 22).

The appropriate critical value, $t^*_{22,.95}$, is 1.717. Since the obtained test statistic exceeds this value, the difference between the means can be considered significant although the variances are unequal.

When K is > 2 and variances are unequal, a modified Welch-Aspin test can be used to perform multiple comparisons instead of an ANOVA and the usual post-hoc tests. Again, the main differences between this procedure and the usual post-hoc tests are that individual sample variances are used rather than a pooled

MSE, and a modified denominator degrees of freedom (γ_2^*) is used to find the critical value.

Consider the data from the example in section 3.2.2.2, and the same contrast that was used in that demonstration. That is,

$$H_o : (3)\mu_1 + (-1)\mu_2 + (-1)\mu_3 + (-1)\mu_4 = 0$$

For which,

$$\hat{\psi} = -12.5$$

Following the Welch-Aspin model, $SE_{\hat{\psi}}^2$ and t^* are computed as before:

$$SE_{\hat{\psi}}^2 = \sum_{i=1}^{K} c_i^2 \, s_i^2 / n_i \text{ (as in the approximate } F \text{ test)}$$

$$= \frac{(3)^2(8.214)}{8} + \frac{(-1)^2(8.839)}{8} + \frac{(-1)^2(9.696)}{8} + \frac{(-1)^2(2.796)}{8}$$

$$= 11.907$$

and,
$$t^* = \frac{\hat{\psi}}{SE_{\hat{\psi}}} = \frac{-12.5}{\sqrt{11.906}} = -3.622$$

Thus far, the procedure is identical to the approximate F

test illustrated in section 3.2.2.3. Note that $t^* = -3.623$ is the square root of $F = 13.124$.

The critical value against which the test statistic in this model is evaluated is an approximate Scheffé coefficent, S^*, given by:

$$S^* = \sqrt{\gamma_1 (F_{\gamma_1, \gamma_2, 1-\alpha})}$$

where,

$$\gamma_1 = K - 1 = 3$$

and,

$$\frac{1}{\gamma_2^*} = \frac{3}{K^2 - 1}\left[\sum_{i=1}^{K}\frac{1}{n_i - 1} - \frac{2}{\sum_{i=1}^{K} W_i}\sum_{i=1}^{K}\frac{W_i}{n_i - 1} + \frac{1}{(\sum_{i=1}^{K} W_i)^2}\sum_{i=1}^{K}\frac{W_i^2}{n_i - 1}\right]$$

and each $W_i = \dfrac{n_i}{s_i^2}$

The computations for γ_2^* are summarized in Table 23. They lead to:

$$\frac{1}{\gamma_2^*} = \frac{3}{4^2 - 1}\left[.5716 - \frac{(2)(.7950)}{5.5653} + \frac{1.5193}{(5.5653)^2}\right]$$

$$= \frac{3}{15}\left[.5716 - .2857 + .0491\right]$$

$= .067$

So, $\gamma_2^* = \frac{1}{.067} = 14.93$

which is always rounded down (in this case to 14).

The critical value can then be computed as:

$$S^* = \sqrt{(3)(F_{3,14,.95})}$$

$$= \sqrt{(3)(3.34)}$$

$$= 3.165$$

For interest, we can compare this value to the critical value of the Scheffé coefficient which would be used if the variances were statistically equal:

$$S = \sqrt{\gamma_1 (F_{\gamma_1, \gamma_2, 1-\alpha})}$$

$$= \sqrt{(3)(F_{3,28,.95})}$$

$$= \sqrt{(3)(2.95)}$$

$$= 2.975$$

Thus, even though the degrees of freedom have been reduced from 28 to 14, the new critical value is only marginally larger than that from the original ANOVA. The new value can be used to assess the significance of the test statistic from any contrast. Clearly, the procedure is computationally quite laborious but it is an appropriate way to deal with unequal variances while sacrificing very little power.

1.3 Alternatives to the ANOVA

If the researcher wishes to avoid the problem of unequal variances entirely, a number of nonparametric alternatives to the ANOVA can be performed. In most cases, these tests are only slightly less powerful than their ANOVA counterparts when the assumptions of normality and equal variances are satisfied. If the equal-variance assumption is violated, they are an excellent alternative to the ANOVA. However, these tests cannot be applied unless the experimental design involves only a single factor. Thus, interactions between factors cannot be assessed.

1.3.1 Kruskal-Wallis test

Kruskal and Wallis (1952) developed a test statistic for use with a one-way ANOVA design which has a sampling distribution that is approximately χ^2 with K - 1 degrees of freedom. The test is simple to perform by hand and is demonstrated here with the data presented in Table 24.

First, consider all N (= 12) observations and rank them from

1 to N, assigning a rank of 1 to the smallest value (see Table 25). Second, find the sum of the ranks (R_i) within each sample (Table 25). Third, compute the test statistic, H, using:

$$H = \left[\frac{12}{N(N + 1)} \sum_{i=1}^{K} \frac{R_i}{n_i}\right] - 3(N + 1)$$

(where the 12 in the numerator of the first term is a constant, not the value of N)

$$= \frac{12}{12(13)} \left[\frac{14^2}{4} + \frac{41^2}{4} + \frac{23^2}{4}\right] - 3(13)$$

$$= 7.2692$$

The critical value, $\chi^2_{K-1,1-\alpha}$ ($\chi^2_{2,.95}$), is 5.99147 (with α = .05). Since the obtained test statistic exceeds this value, the null hypothesis that the means of the ranked data for the different samples are equal can be rejected.

Post-hoc contrasts can be performed by computing a Z test statistic:

$$Z = \frac{\hat{\psi}}{SE\,\hat{\psi}}$$

where, $$\hat{\psi} = \sum_{i=1}^{K} \frac{C_i R_i}{n_i}$$

and, $$SE^2_{\hat{\psi}} = \frac{N(N + 1)}{12} \sum_{i=1}^{K} \frac{C_i^2}{n_i}$$ (where again the 12 in the denominator of the first term is a constant)

As an example, consider contrasting the mean for group 2 with the average of those for groups 1 and 3. The contrast coefficients (C_i) are thus -1, 2, and -1, so:

$$\hat{\psi} = (-1)\left(\frac{14}{4}\right) + (2)\left(\frac{41}{4}\right) + (-1)\left(\frac{23}{4}\right)$$

$$= 11.25$$

and,

$$SE^2_{\hat{\psi}} = \frac{12(13)}{12}\left[\frac{-1^2}{4} + \frac{2^2}{4} + \frac{-1^2}{4}\right] = 19.5$$

consequently,

$$\underline{Z} = \frac{11.25}{\sqrt{19.5}} = 2.5476$$

The critical value for all post-hoc contrasts is found as $\sqrt{\chi^2_{K-1,1-\alpha}} = \sqrt{5.99147} = \pm\ 2.4477$ in this example. Since the obtained test statistic exceeds this value, the contrast is significant.

If many of the observations are tied, it may be worthwhile applying a correction factor (C) and computing a new test statistic (H*), where: $H^* = \frac{H}{C}$

where, $$C = 1 - \frac{1}{N^3 - N} \sum_{i=1}^{K} (t_i^3 - t_i)$$

and, t_i = the number of observations tied at a given value.

Consider the data in Table 26, which have been assigned rank values in Table 27. Note that when two or more observations are tied, they are each assigned the average of the ranks that they would have been assigned had they been different. For these ranked data:

$$H = \frac{12}{(16)(17)} \left[\frac{29^2}{4} + \frac{49^2}{4} + \frac{29^2}{4} + \frac{29^2}{4} \right] - (3)(17)$$

$$= 3.309$$

To compute the correction factor (C), note that 2 observations are tied with rank 1.5, 3 are tied with rank 4, 2 are tied with rank 6.5, 3 are tied with rank 9, 3 are tied with rank 12, and 2 are tied with rank 14.5.

Thus, $C = 1 - \frac{1}{16^3 - 16}\left[(2^3-2)+(3^3-3)+(2^3-2)+(3^3-3)+(3^3-3)+(2^3-2)\right]$

$= .9779$

and, $H^* = \frac{3.309}{.9779} = 3.384$

Note that even with many ties, as in this example, H^* is only marginally greater than H. Since H^* will always be larger than H, one need not compute H^* unless H does not exceed the critical value. With a large number of ties, and if an H is obtained close to the critical value, H^* may turn out to be significant and should be computed.

1.3.2 Friedman Test

This is the most frequently used nonparametric alternative to the repeated-measures ANOVA. Consider the data in Table 28, representing the scores of $n = 6$ observers scaling $K = 3$ stimuli. First, rank each observer's scores from 1 to K, assigning a rank of 1 to the smallest score (Table 29). Second, find the sum of the ranks (R_i) for each of the stimuli (Table 29). Third, compute the test statistic, χ^2, as:

$$\chi^2 = \frac{12}{nK(K+1)} \sum_{i=1}^{K} R_i^2 - 3n\,(K+1)$$

where: 12 is a constant

n is the number of observers per condition

K is the number of stimuli

3 is a constant

Thus, $$\chi^2 = \frac{12}{(6)(3)(4)}\left[8^2 + 10^2 + 18^2\right] - (3)(6)(4)$$

$$= 9.333$$

The critical value, $\chi^2_{K-1,1-\alpha}$ ($\chi^2_{2,.95}$ with α =.05), is 5.99147. Since the obtained test statistic exceeds this value, it can be concluded that there is at least one significant difference between the mean ranks across trials.

Post-hoc tests are performed in a similar manner to those in the Kruskal-Wallis model, using a Z test statistic, computed as:

$$Z = \frac{\hat{\psi}}{SE\hat{\psi}}$$

As before, $$\hat{\psi} = \sum_{i=1}^{K} \frac{C_i R_i}{n_i}$$

In this model, $$SE^2_{\hat{\psi}} = \frac{K(K + 1)}{12} \sum_{i=1}^{K} \frac{C_i^2}{n_i}$$ (where the 12 is a constant)

Consider the comparison of group 3 with the average of groups 1 and 2:

Then, $\hat{\psi} = (-1)\left(\frac{8}{6}\right) + (-1)\left(\frac{10}{6}\right) + (2)\left(\frac{18}{6}\right)$

$= 3.0$

and, $SE_{\psi}^{2} = \frac{(3)(4)}{12}\left[\frac{(-1)}{6}^{2} + \frac{(-1)}{6}^{2} + \frac{(2)}{6}^{2}\right]$

$= 1.0$

So, $\underline{Z} = \frac{3.0}{\sqrt{1.0}} = 3.0$

The critical value, $\sqrt{\chi^{2}_{K-1,1-\alpha}}$ is ±2.4477. Since the obtained test statistic exceeds this value, the contrast is significant.

When all the assumptions of the ANOVA are satisfied, the Friedman test is somewhat less powerful than the ANOVA, though its power increases as a function of $\underline{K}$. If the assumptions of the ANOVA, including equal variances, are not satisfied, the Friedman test may actually be more powerful, and, thus, would be an excellent alternative. An even more powerful procedure would be to perform multiple matched-pair Wilcoxon tests, but this requires that only pairwise contrasts be performed. If the researcher plans to do only pairwise contrasts, this is the recommended procedure and it is illustrated in the following section.

1.3.3 Matched-Pair Wilcoxon Test

Initially, this test was developed as a nonparametric alternative to the matched-pair or repeated measures t test. As mentioned, it can also be used to perform multiple pairwise contrasts when K is greater than 2 and , by using appropriate Tables, α can be controlled across all the contrasts to some predetermined value (e.g., .05). Since the procedure is identical regardless of the number of groups or the number of contrasts being performed, it is illustrated here with an example with n_i = 8 observers tested on K = 2 stimuli. The data, and the computations involved in this test, are presented in Table 30.

First, find the difference between each observer's scores -- subtracting stimulus 2 from stimulus 1 scores. Second, record the absolute values of the differences. Third, assign ranks to the absolute differences, giving a rank of 1 to the smallest difference. For tied values, assign the average of the ranks that would have been assigned had the values differed. Finally, compute the sum of the ranks associated with initially positive differences (T_+), and the sum of the ranks associated with initially negative differences (T_-). Either T_+ (8.5 in this example) or T_- (27.5 in this example) can be used as the test statistic (T).

To determine the significance level of the test statistic, refer to Table A-21 (copied without permission from Marascuilo and McSweeney, 1977). In this example, with an n_i of 8, the null hypothesis of no difference between the distributions of scores on stimulus 1 and stimulus 2 would have been rejected if T were ≤ 5 or ≥ 31, with probability of a Type I error (p) = .039.

When multiple pairwise contrasts are performed, α can be controlled at a predetermined level by the following procedure. Suppose we have 10 observers tested on $K = 4$ stimui. There are 6 possible pairwise contrasts, and if the overall α were set at .05, each would be performed with $p \leq .05/6$ (or .0083). Referring to Table A-21, for $n_i = 10$, it can be seen that the decision rule for each contrast would be to reject H_o if T were ≤ 4 or ≥ 51. With this decision rule, each contrast has an α of .007, for an overall α of 6(.007) or .042.

If any observers produce the same scale value for 2 stimuli, (i.e., the difference between the scale values is 0) this difference is discarded before the other differences are ranked. Subsequently, the critical value is found with an n' of $n - d_c$ (where d_o is the number of differences equal to 0).

II. CONCLUSIONS AND RECOMMENDATIONS

1. If sample sizes are all equal or close to equal, inequality of variances is seldom a problem for ANOVA and a test of equality of variances need not be performed.

2. If sample sizes are unequal and the underlying distribution is normal, use Bartlett's test to assess the homogeneity of the variances.

3. If sample sizes are unequal and the underlying distribution is nonnormal, the Box-Scheffé test would be the preferred test for assessing the equality of the variances.

4. If the variances are found to be unequal, one of several data transformation procedures may be used to remove the heterogeneity. This may not be advisable if the researcher is interested in studying interactions between independent variables.

5. A number of alternative analytic procedures may be used in place of the regular ANOVA if variances are unequal. If planned comparisons are performed, the approximate *F* test can be used. If post-hoc tests are performed, the Welch-Aspin model is recommended. Either of these procedures can be used to test any hypothesis of interest (e.g., pairwise contrasts, complex contrasts, interaction contrasts, tests for trends, etc.).

6. If the variances are very heterogeneous, use of nonparametric alternatives to ANOVA may result in more power and is, thus, recommended.

CHAPTER 5 - A COMPLETE MODEL OF THE SCALING PROCESS

In the previous chapters, we have outlined rationale and instructions for what we feel to be the optimal data collection and analysis techniques to use in the evaluation of teletext systems. The categorical judgment technique was deemed to be the best of the data collection techniques available. Two analysis procedures for scaling stimuli were recommended, Thurstone's (1927) and Allnatt's (1973;1975;1979). Allnatt's technique is the more powerful of the two but it can only be used if the stimuli vary along a quantitative physical dimension. Thurstone's technique is more flexible and should be used whenever stimulus variation is qualitative or the assumptions of Allnatt's technique are demonstrably incorrect. The assumptions of Thurstone's technique must, of course, be validated before attempting to calculate or further analyze its resultant scale values. Finally, means of analyzing the resultant scale values were suggested. The ANOVA is the preferred analysis method. However, when the homogeneity of variance assumption of the ANOVA is demonstrably incorrect, a number of options are available, including data transformation, statistical corrections and non-parametric techniques.

In both Thurstone's and Allnatt's techniques, there is an assumption of an internal dimension, $\underline{S}$, which the observers use in some way to produce the resultant scale values. In this fifth chapter, we would like to examine the means by which values on this internal dimension are created. In some cases the scale values may arise in a fairly direct fashion. In others, a number of cognitive variables may be important. In this chapter, we will present a general model of this procedure which will encompass both types of situations.

I. THE BASIC MODEL

Central to both Thurstone's (1927) and Allnatt's (1973;1975;1979) scaling procedure is an internal subjective dimension, S, which is presumed to represent the dimension of judgment. For example, if observers are asked to rate the pleasantness of a set of objects, S is a pleasantness dimension. In the present circumstance, in which observers are asked to rate the acceptability of a teletext screen, this dimension is presumed to be an acceptability dimension, although it could be thought of as a dimension of impairment, presumably the reverse of acceptability.

Each time a stimulus X_i is presented, a value Y_i is assumed to be produced on this S dimension. The Y_i values are not constant but are assumed to be random selections from a normal distribution on S with a mean S_i and variance a_i^2. The scale values we determine in Thurstonian scaling are assumed to be these $\underline{S}_i$ values. The scale values produced by Allnatt's technique are assumed to be transformations of the $\underline{S}_i$ values.

Both Thurstone's and Allnatt's scaling techniques go on to make a number of assumptions about how observers treat the $\underline{Y}_i$ values in order to produce a final response. Chapter three provides a detailed analysis of these assumptions for both techniques. What is yet to be discussed are the mechanics by which the $\underline{S}$ dimension is created and $\underline{S}_i$ values determined. These are the issues addressed in this chapter.

In our conceptualization, the $\underline{S}$ dimension is not considered to exist prior to the scaling session. Instead, observers must create this dimension on the basis of the experimental context. In particular, things like the instructions the observer receives, the beliefs he or she brings to the experiment, the types of stimuli being scaled and all the effects of context will play some role in the creation of this dimension. However, there have to be some primitives here. In particular, the assumption is being made that an initial perceptual process always reveals a stable perceptual representation of any stimulus, regardless of task variables. In many cases, this representation will be multidimensional and its dimensions will need to be discovered. The dimensions, of course, must also be presumed to be stable. Experimental and analysis techniques for determining the nature of these dimensions will be addressed in section 2.2.

The observer's task then becomes that of taking this perceptual representation and distilling a value on the *S* dimension. The way this is presumed to be done is to weight the various perceptual dimensions in some fashion. Some of these may have 0 weight, meaning the dimension is irrelevant to the acceptability judgment. The maximum weight a dimension may receive can vary, depending on the relationship between the units of the perceptual dimensions and those of *S*. This weighting process is basically the interesting one in the model because it is through this process that the *S* dimension is defined. That is, only when the observer decides, for example, that each of three perceptual dimensions should be weighted equally does the *S* dimension come into existence. Further, it is here that context effects arise. That is, context is assumed to influence the judgment process by inducing an observer to use a different set of weights for the *same judgment* in different contextual situations.

Different types of judgments will, of course, also produce different weighting schemes. For example, observers would undoubtedly weight, say, three perceptual dimensions, in a different fashion when asked judge pleasantness than when asked to judge acceptability. The general assumption is that observers are free to weight the dimensions in any way they want. A main purpose of our analysis is simply to determine what weighting scheme they've chosen. However, with some types of judgments only certain weighting schemes would be acceptable. In particular, when more mainline physical dimensions are being varied (e.g., radiance, weight) judgments on the corresponding perceptual dimensions (i.e., brightness, heaviness) must be straightforward and must, in fact, follow a power function relationship (Stevens, 1961). This can only occur if the observers isolate a single relevant perceptual dimension and give only it a nonzero weight. If observers include any other perceptual dimensions in this process, for example, if they give a nonzero weight to a dimension corresponding to colour while making heaviness judgments, we would have to conclude that they simply weren't following the task instructions.

II. THE PERCEPTUAL PROCESS

2.1 Can It Be Regarded as Stable?

The first issue to be dealt with is the claim that the perceptual process yields a stable representation of any stimulus irrespective of task demands. This idea is by no means a new one. In fact, if we had stated it at any time before the late 1940's it would have been regarded as a statement of the obvious. However, with the end of that decade and the beginning of the next a somewhat different view of perception emerged. Simply stated, the idea became that perception does not yield a stable representation of a stimulus. Instead, a host of nonperceptual factors (e.g., expectation, familiarity, set) interact with the sensory information, creating a situation where the same sensory stimulus can be perceived differently in two different situations. This view has become known as the New Look in perception (Bruner, 1957).

Although it isn't clear why the New Look view emerged when it did, there appear to have been two lines of research at that time which may have given it its impetus. One was Helson's (1948) work on adaptation level. Briefly, what Helson demonstrated was that the results of even the most basic scaling experiments were influenced by nonperceptual factors, in particular, by memory for other stimuli. The other area that provided an impetus to these notions was the area of perceptual defence (McGinnies,1949). Here, the findings seemed to show that certain words which were regarded by the society of the day as taboo (e.g., raped, whore) could not be perceived as readily as "normal" words (e.g., apple, table). Thus, once again, we seem to be observing an influence of nonperceptual factors on the perceptual process.

In the thirty-five years since the New Look view emerged, an impressive array of data and an impressive set of arguments have emerged to support it. Now classic examples would be two phenomena involving the perception of words, the word-superiority effect (Reicher,1969; Wheeler, 1970) and the semantic priming effect (Meyer and Schvaneveldt,1971; Meyer, Schvaneveldt and

Ruddy, 1975). The word superiority effect refers to the finding that a letter is more readily reported from a briefly displayed word (e.g., the *T* in CAT) than from a briefly displayed nonword (e.g., the *T* in DMT). Semantic priming refers to the finding that a word (e.g., BUTTER) is responded to more rapidly following perception of a related word (e.g., BREAD) than following perception of an unrelated word (e.g., DOCTOR). In both cases, the finding is explained by suggesting that the context in which the stimulus appears alters and, thus, facilitates its perception.

Additional arguments for the New Look view can be based on everyday observations. Try, for example, to listen to someone speaking English and hear the speech sounds as noise (like one can do if they are not familiar with the speaker's language). Even with great effort you will find it almost impossible to do. Thus, here the fact that you have *learned* the English language appears to be influencing how these sounds are perceived.

There is obviously merit in the arguments for the New Look view of perception. The studies (and anecdote) cited above clearly are demonstrations of the influence of nonperceptual factors. Thus, they raise the question of whether our assumption of a stable perceptual representation, uninfluenced by external factors, can be a viable one. The answer is, in fact, still yes. The reason lies in the basic definitional difference between the New Look researchers' conception of perception and the one we have adopted. In particular, we wish to view perception as the process of establishing a perceptual representation while in the New Look the perceptual process also includes the process of interpreting or categorizing that representation.

Probably the easiest way of explaining this distinction is by considering the theory of signal detection. An issue that the researchers of the 19th century had to grapple with was the fact that observers in detection experiments often said a signal was present when no signal, not even a subthreshold one, was presented. The whys and wherefores of this problem went unresolved until the mid 1950's when Tanner and Swets (1954; see also Green and Swets, 1966) brought the theory of signal detection to psychology. In signal detection theory it is realized that the data in these "perceptual" experiments (the observers' responses) reflect not only the observers' perceptions but also their biases, mental set and so on. As such, signal detection theory provides a means for deriving two measures, one, *d*', to index the perceptibility of a stimulus, and a second, β , to index the extent of the observer's bias. Thus, the usefulness of the theory is that it gives a way of separating more cognitive factors from the actual perceptual effects an investigator is trying to study.

In constructing our model, we are attempting to do the same thing, that is, to separate theoretically the decision process involved in deriving the *S* values from the perceptual information on which the *S* values are based. However, to be a useful model, it must have more than theoretical reality. That is, the separation of these two processes must be possible on a practical level as well. (The fact that in some circumstances they can't be makes the New Look approach quite attractive.) While we can't guarantee that such a separation can be accomplished in the present circumstances (i.e., when scaling teletext service parameters), there are a number of reasons for believing we will be successful.

The first reason is that a substantial amount of evidence suggests that perceptions can't simply be changed by learning new facts about the environment. Observationally, illusions, such as those in Figure 7, are perceived by nearly everyone. More importantly, the perception remains even after objective measurement demonstrates to the observer that he or she is viewing an illusion (e.g., measure the lines in Figure 7a to convince yourself they are of equal length, then try to perceive them as such).

A second case involves one's perceptual response to moving one's eyes artificially. If you place your finger on one eyelid and press gently but firmly, you will move your eye slightly. Although you realize that it is your eye which is moving and not the outside world, the typical phenomenal experience observers report is that they perceive the world as moving.

A third reason for believing we will be successful is that many of the data favouring the New Look view do not stand up to close experimental scrutiny. For example, consider the phenomenon referred to as the word superiority effect. In this task, an observer is shown a letter string very briefly and then asked to report the identity of the letter at a particular position. In an attempt to minimize the use of obvious non-perceptual strategies (e.g., guessing on the basis of knowledge of English), observers are only required to select one of two possibilities for the letter in question. If a word had been presented, the two possibilities would complete words (e.g., if the word had been WORD and the fourth position was probed, the two possibilities might be *D* and *K*). If a nonword had been presented neither possibility would complete a word (e.g., if MCRD had been presented and the fourth position was probed, the two possibilities might be *D* and *K*). As noted above, in these circumstances, a letter in a word is reported more accurately than a letter in a nonword.

At first, this result was taken as reasonably strong evidence that letters are perceived differently in word contexts than in nonword contexts. However, more recently, a number of investigators (Thompson and Massaro, 1973; Bjork and Estes, 1973) have demonstrated that the word-superiority effect is an artifact of the experimental situation. In particular, if the two possible responses and the letter position to be probed are known by the observer ahead of time, the effect disappears. Further, using more sensitive measures of what is perceived, Krueger and Shapiro (1979) and Massaro (1979) have demonstrated that the letter's perceptual representation is unaffected by its context. Thus, it appears that the word superiority effect is not a perceptual phenomenon but one which arises at a later stage.

Similar arguments can be directed against the interpretation of other linguistic phenomena as perceptual effects (e.g., the semantic priming effect, our inability to treat speech as noise). That is, while these effects appear to demonstrate the influence of nonperceptual factors on perception, their actual influence occurs at a later level. The main reason these types of stimuli cause the interpretation problems they do is well explained by the automatic versus controlled processing distinction first proposed by Posner and Snyder (1975) and later expanded on by Shiffrin and Schneider (1977). The notion is that, for a beginning reader (or listener), understanding language is a matter of going through each processing step in a conscious controlled manner. However, with a sufficient amount of practice, these processes and the linkages between them require less and less effort and attention. Ultimately, the whole sequence of behaviors becomes automatic in the sense that it requires virtually no processing effort and, in fact, inevitably runs to completion whether the observer wants it to or not. In this way, the perceptual process becomes so intertwined with other processes that teasing them apart becomes an extremely difficult task.

For a New Look theorist, this creates no real problems. Perception is defined to involve all these automatic processes and it is simply studied as such. However, for our purposes, problems are created both for the experimenter trying to study the basic perceptual processes and for an observer trying to use raw perceptual information.

For the stimuli we wish to scale and, in fact, for most nonlinguistic stimuli, the automatic processing issue really shouldn't be important. Whatever automatic processing goes on when people judge teletext screens will, most likely, involve only the linguistic aspects of the text (i.e., its meaning). If the text message is kept constant, while other, more relevant, stimulus parameters are varied, this automatic processing should have little effect on either the perceptual representations or the acceptability judgments. If so, the model assumption that the perceptual process can be separated from what is done next appears to be a viable one and we should, therefore, be able to study the nature of these perceptual representations. (As a caveat, however, the reader should realize that these techniques can't be applied willy-nilly in other scaling situations. If the to-be-scaled stimuli do engage much automatic processing, the results of this processing may obscure the perceptual nature of the stimuli. Thus, it may be the results of this subsequent processing rather than the results of perceptual processing that the observers are using in the scaling process.)

2.2 Determining the Nature of the Perceptual Representation

If we accept the assumption that observers have access to stable perceptual representations of teletext service parameter effects, the next step is to understand the nature of those representations. We start this analysis by assuming that a perceptual representation can be thought of as a point in a multidimensional space. The point is the stable representation we've been talking about. The dimensions themselves represent constant perceptual attributes. If all goes well, the end product of our analysis will be the re-creation of this multidimensional space. From this, we should be able to discover both the nature of the dimensions and how each stimulus is represented.

The analysis technique to be used here is called multidimensional scaling (MDS). The data for MDS are measures of "proximity" between pairs of stimuli. For our purposes, the proximity measures will be direct similarity or dissimilarity judgments. For the task, a large number of stimuli varying on *all* the physical dimensions of interest would be presented in pairs. Observers will be asked to rate either the similarity or dissimilarity of each pair on some scale. After all pairs of stimuli have been presented at least once, a similarity or dissimilarity matrix can be created where the average rating for each stimulus pair is the value contained in the appropriate cell in the matrix. (See Table 31 for a sample dissimilarity matrix.

For the remainder of this discussion, we will assume that the observers are rating the dissimilarity of the stimulus pairs.) Note that the obtained averages can actually be tabled in only one-half of a matrix. The reason is there should be no difference either theoretically or practically between the dissimilarity between $\underline{X}_i$ and $\underline{X}_j$ and that between $\underline{X}_j$ and $\underline{X}_i$. It is this half matrix that provides the data for any MDS analysis.

The scale that observers use to produce their ratings is somewhat arbitrary. Probably the best experimental procedure would be to anchor the scale by presenting two identical stimuli to the observer and suggesting that they be given a dissimilarity rating of 0. Then, two stimuli which are quite different, for example, the clearest picture possible and total noise, could be presented with the suggestion that they be given a rating of 10. Intermediate levels of dissimilarity (i.e., those represented by the stimuli the observer is about to rate) would therefore receive ratings intermediate to the values 0 and 10.

Care should be taken with the instructions given the observer. The subjective dimension of acceptability should not be mentioned. What we are trying to do is discover the dimensions which make up the perceptual space, not to build new dimensions into the ratings. Similarly, variation on irrelevent physical dimensions should be avoided. For example, unless variations in colour are of interest, care should be taken to hold colour constant. However, **all** physical dimensions which are of interest should be represented. The scaling programs can't allow discovery of important perceptual dimensions if stimuli don't vary on those dimensions. Since we don't know how the physical dimensions are represented as perceptual dimensions we can't afford to leave out any physical dimensions that could be important.

Ideally, the stimulus set will be created by factorially crossing all physical dimensions of interest. At least three levels on each dimension should be selected. Thus, if there are four dimensions of interest there would be 3^4 (= 81) stimuli. If each stimulus is compared to each other stimulus once by each observer, there would be 81 * 80/2 (= 3240) judgments per observer. More stability would, of course, be obtained if each judgment were made two or three times by each observer. Obviously, even with only one judgment per pair, the task will take a certain amount of time. However, until something more is known about the perceptual dimensions, it's probably best to be as thorough as possible.

Fortunately, this "discovery scaling" process needs to be done only once. Once a few subjects (5 - 10) have been run in this extensive dissimilarity rating task, we should know what the perceptual dimensions are and how they relate to the physical dimensions. However, prior to each scaling experiment using a different stimulus set, the dissimilarity rating task will have to be carried out with the set of stimuli to be scaled. The purpose here is not to discover perceptual dimensions (presumably, we already know what they are) but to determine the precise coordinates on those dimensions of the stimuli being scaled. These coordinates can then be used in the subsequent analyses. Given that a typical scaling study seldom will involve many stimuli (e.g., more than 10), the number of dissimilarity ratings required should always be less than 100.

The number of MDS programs one has to choose from is quite large. In the older methods (Torgerson, 1952; 1958), dissimilarity ratings are assumed to be proportional to distance in Euclidean space. (These are referred to as metric procedures.) In more recent techniques (Shepard, 1962; Kruskal, 1964), it is simply assumed that there is a monotonic relationship between dissimilarity ratings and distance (nonmetric procedures). In either case, the scaling program attempts to fit the stimuli into an *n*-dimensional space so that the distances between stimuli in the space accurately reflect the dissimilarity ratings. The number *n* is usually allowed to vary between, say, 2 and 6, with a solution returned for each situation.

The solution returned consists of a set of coordinates for each stimulus in an *n*-dimensional space and a measure of how well the dissimilarity ratings match up with the interstimulus distances in this space. The latter measure is called the STRESS or

S-STRESS of the solution. Because there is variability in observers' response processes, the match between distance and dissimilarity is never perfect. Thus, the STRESS value is never zero. In addition, STRESS will decrease as the number of dimensions used increases, simply because the number of free parameters is increasing. Thus, the solution we wish to accept is one which minimizes not only STRESS but also *n*. Typically, STRESS is plotted as a function of *n* (as in Figure 8) and an elbow on the graph is located. (The elbow here is at *n*=3.) The idea is that increasing *n* from the elbow provides only minimal decreases in STRESS, while decreasing *n* increases STRESS quite dramatically. Thus, the best solution is assumed to have three dimensions. (Figure 8 is, of course, artificial. In the real world, the elbow is always much less clear, however, very few solutions need more than three dimensions.)

At this point (in the example), we know we have a three-dimensional solution (i.e., a three-dimensional perceptual space). However, we still do not know what the axes represent or where they are located in the space. We next have to discover not only what but where those axes are. To understand why this problem arises, one needs only realize that distances between points (what the program is interested in) are totally independent of where the axes are. Thus, the program orients the axes somewhat arbitrarily in the solution. The axes, thus, must be rotated to allow the dimensions to be understood. Rotation can be accomplished by formal procedures (e.g., varimax) or by hand. In either case, the point of the rotation (and, in fact, the point of the entire procedure) is to find three dimensions that have psychological (here, perceptual) reality. In many instances (although not always), a physical dimension will have a direct relationship to a perceptual dimension. Thus, the physical dimensions should be kept in mind when doing the rotation.

Selection of a MDS program is more or less up to the individual. Many nonmetric procedures were developed during the 197 (see Schiffman, Reynolds and Young (1981) for a review) and selection of one of these is suggested. The two most highly recommended programs are KYST and ALSCAL. KYST seems to have the best features of almost all the MDS programs written in the 1970s. ALSCAL is nearly as powerful and also includes provisions for doing individual-difference scaling if observer differences are an issue. Ultimately, however, the selected program will probably be the one most accessible in the available statistical packages.

III. PRODUCING A VALUE ON S

3.1 Basic Issues for the Model

3.1.1 Theoretical Issues

Once the perceptual process has been completed and a representation established, observers must next turn this representation into a value on $\underline{S}$. According to the model, this is done by a simple weighting process. The coordinates on the various perceptual dimensions are multiplied by weights, and the products summed, (potentially, a constant could be added to the equation). Thus, if we have, for example, a three dimensional solution in which coordinates for stimulus $\underline{X}_i$ are designated $(\underline{x}_{i1}, \underline{x}_{i2}, \underline{x}_{i3})$, $\underline{S}_i$ can be expressed as:

$$S_i = w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i3} + k$$

where $\underline{w}_j$ is the weight for dimension $\underline{j}$ and $\underline{k}$ is a constant.

As the equation indicates, scale values depend on both the nature of the perceptual representation and the way that observers choose to weight the various dimensions of that representation. Different $\underline{S}$ values could be produced by the same stimulus if circumstances dictated different $\underline{w}_j$s. For example, if observers were specifically told to attend to a particular stimulus aspect that is directly reflected in a perceptual dimension, they would presumably weight that dimension more than they would in other circumstances. Understanding the role of experimental manipulations will, thus, be a matter of understanding how they influence the choice of weights.

The reader should note that we do not require that the sum of the weights equals one. Thus, this equation is more appropriately referred to as a linear equation in $\underline{n}$ variables rather than a weighting equation. The reason the sum of the weights isn't constrained is that units in the perceptual space and units in the $\underline{S}$ dimension have no particular relationship to one another. If the units of the perceptual space were smaller than those of the $\underline{S}$ dimension, coordinate values would be larger than they should be. Thus, weights must shrink these values to produce the $\underline{S}_i$s. As such, the sum of the weights would be less than one. On the other hand, if the units of $\underline{S}$ were smaller, the weights must increase the coordinate values, meaning that their sum would be greater than one.

With respect to the issue of how the weights are affected, there appear to be two general types of variables that may have an effect. The first could be called experimental context variables. These are variables that are introduced to the observer by the experimenter before the experiment starts. Typically, these are introduced through instructions but they can also arise from any other experimental manipulations that create expectations. Examples would be things like the instruction to be critical or lax or, perhaps, the instructions to assume various roles when judging. How these instructions could influence the weights is, of course, an empirical question. However, one could envision that instructions to pay attention to particular physical dimensions would increase the weight put on relevant perceptual dimensions and decrease the weight put on irrelevant perceptual dimensions. Instructions to be lax or critical may have the effect of increasing (or decreasing) all the weights and so on.

The other type of variables that could affect the assigned weights would be referred to as stimulus variables. For example, if the stimuli were all high in quality, observers might adopt a more critical weighting scheme (Allnatt and Corbett, 1972). Similarily, if the stimuli were all quite unfamiliar, observers might adopt a more neutral weighting scheme while they try to determine which dimensions are more important.

One thing to be aware of is that while we're hypothesizing that variables such as these affect the weighting process there are other possibilities. In particular, as Parducci (1965) has suggested, these variables may affect criterion placements. For example, if all the stimuli are of high quality all the criteria could be placed on the high end of the acceptability dimension. Again, the exact effect the variables are having can be determined empirically. If it turns out that Parducci's suggestion is correct, even part of the time, the model can be amended with no harm being done to its basic structure.

This model, as conceptualized, has different implications for the two scaling models discussed earlier (Thurstone's and Allnatt's). In particular, the present model and Thurstone's model are completely compatible with one another. Thus, as we will suggest, Thurstone's model can be incorporated into our model as a sort of a "back-end", a description of how the scaling process is completed. Allnatt's model, on the other hand, restricts the present model. The restrictions necessary for the two models to fit together will be discussed in subsequent sections. However, at this point, with the present model being as general as it is, there is no reason to believe that these restrictions are totally unreasonable.

3.1.2 Analysis Issues

The basic conceptualization that each S_i is a linear function of n variables is actually borrowed from Anderson (1968). In Anderson's scaling procedure, stimuli are created by selecting a number of physical dimensions (either quantitative or qualitative), selecting a number of levels on each dimension and then factorially crossing these dimensions. For example, if we have two physical dimensions with four levels for each, there would be 4^2 (or 16) stimuli to scale.

Each level, i, on a given dimension, j, is presumed to give rise to an impression, I_{ij}. These impressions are then weighted by a dimension specific weight, w_j, and summed. Thus, the resultant scale value for a stimulus created by mixing, say level 3 on dimension 1 and level 4 on dimension 2 (i.e., S_{34}) would be

$$S_{34} = w_1 I_{31} + w_2 I_{42} + k$$

where $\underline{k}$ is a constant. There would, of course, be 16 of these equations, one for each stimulus being scaled.

In Anderson's analysis, the next step would be to use all 16 scale value equations to solve for the $\underline{I}_{ij}$. The calculated $\underline{I}_{ij}$ values are then regarded as measurements on an interval, rather than a ratio, scale (i.e., the resultant values are linear functions of the "true" $\underline{I}_{ij}$). In our situation, however, we already have the $\underline{I}_{ij}$; they are the perceptual coordinates derived from our dissimilarity judgments. We also have the values on $\underline{S}$ from the category scaling analysis. What we need are the weights, $\underline{w}_j$, and the constant, $\underline{k}$. Thus, our system of equations will have one unknown for each perceptual dimension plus one more for the constant. In the above example, we'd have 16 equations in 3 unknowns.

If there ever were fewer equations than unknowns, we would have what is referred to as an underdetermined system and we would be unable to solve for the weights. This situation will probably never arise here as it's unlikely that we'd ever have more than four perceptual dimensions (thus, five unknowns) or fewer than five stimuli being scaled. If the number of equations and unknowns are equal (also an unlikely possibility) the weights can be solved for unambiguously. In the more common situation, there will be more equations than unknowns, as above. This situation is referred to as an overdetermined system. One set of values for the unknowns may work for some stimuli but not for others. Thus, we have to produce best estimates of the weights which fit these equations as well as possible. This is best accomplished by applying a multiple regression analysis.

In multiple regression analysis, an equation of the form

$$Y = \beta_0 + \beta X_1 + \beta X_2 + \beta X_3 + \ \ldots \ \beta X_k$$

is used to input known values, the Xs, and to predict scores on another variable, the Ys. This is done through estimation of the β s. The β values are selected by comparing predicted values on the Y variable ($\hat{Y}$s) to observed values on the Y variables and minimizing the sum of the squared differences, $\Sigma(Y - \hat{Y})^2$. Obviously, if the equations and the procedure work $\Sigma(Y - \hat{Y})^2$ should be as small as possible. If so, the procedure has selected β s which fit well in the equations. A measure of how good the fit is is provided by the multiple regression procedure and is referred to as r^2. To the extent that r^2 is close to 1 (it's always between 0 and 1, inclusive) we've been successful in finding β s. A test of how well we've done (i.e., a test of how big r^2 is) is also provided by the procedure. An F value is produced which is evaluated against a critical F with numerator degrees of freedom equal to the number of β s solved for minus one, and denominator degrees of freedom equal to the number of scores on the Y variable minus the number of β s solved for.

In our circumstance, the Y scores are the S_is found in the scaling analysis. The $\underline{X}$ scores are the coordinates on the various perceptual dimensions and the β s are the weights and constant in the set of equations. If the multiple regression procedure produces an $\underline{r}^2$ large enough that $\underline{F}$ is significant, we can feel fairly confident that the β s the procedure has produced are good estimates of the weights and constant. The fit would never be perfect (i.e., $\underline{r}^2$ will never be 1.0), of course, because the $\underline{X}$ scores are only estimates of the "true" coordinate values. However, if the $\underline{F}$ was not significant, it would indicate that the model wasn't doing an adequate job of describing the data. As such, its validity would be called into question.

The procedure can be applied to any stimuli for which we have $\underline{S}_i$ values (from Thurstone's or Allnatt's procedure) and for which we have coordinates in perceptual space. The next question is whether experimental context or stimulus variables change the weights (or $\underline{k}$). Thus, suppose the perceptual space were three dimensional in a certain circumstance and, we have estimated three weights and a $\underline{k}$ for a given observer under each of two instructional conditions. The way to determine whether there has been an effect of instructions is to submit these data to a multivariate analysis of variance (MANOVA).

The MANOVA is a technique for doing an analysis of variance with more than one dependent variable. In this example, we would have four dependent variables (3 weights and $\underline{k}$). Because there are only two conditions (two different types of instructions), our independent variable has only two levels. (There could, of course, be as many levels and as many independent variables as an experimenter might want.) It should be noted that, as with the ANOVA, normality and equality of variance are being assumed. With a large number of observers, the normality problem disappears. However, we may still have a violation of the equal variance assumption. Thus, each dependent variable should be analyzed as outlined in chapter 4 to determine whether it obeys the equal variance assumption. If not, an appropriate transformation technique should be applied.

The technique for doing a MANOVA is quite complicated and won't be described here. It's, perhaps, best to submit your data to a MANOVA program in some available statistical package. What will be produced by the program is, as in the ANOVA, an *F* ratio. If it is significant, we can assume our independent variable has had an effect. If we wish to determine where that effect lies (i.e., with which dependent variable) we can then do separate ANOVAs on each of our dependent variables.

3.2 Implications for Thurstonian Scaling

As mentioned earlier, the model and analyses discussed above can be applied with essentially no caveats to the S_is derived from Thurstonian scaling. Perceptual dimensions can be discovered by means of either metric or nonmetric MDS techniques.

The coordinates can be used straightforwardly in the multiple regression analysis. The resulting weights and constant can then be taken at face value and changes in weights due to experimental manipulations can be evaluated through the MANOVA.

One additional strength of the Thurstonian scaling procedure is that it permits determination of whether changes are being induced in the weights or in the criterion placements. That is, as discussed in chapter 3, the results of Thurstonian scaling are both a set of scale values (the $\underline{S}_i$) and a set of criterion placements. If an experimental manipulation affects only the criterion placements, only these values will differ across different conditions. The $\underline{S}_i$s will remain constant. Similarily, if only the weights are being affected, only the $\underline{S}_i$s and not the criterion positions will change value. Thus, the effects of any manipulation will be transparent.

Such is not the case with Allnatt's technique. With this technique, either experimentally induced criterion changes or experimentally induced weight changes will produce changes in the calculated scale values. There is, of course, a test, described in chapter 3, for the stability of criteria in Allnatt's procedure. However, it's so weak that, in most instances, a lack of stability would never be detected. Thus, most experimentally induced changes will have to be modelled as affecting the $\underline{S}$ dimension, whether or not such is the case.

3.3 Implication for Allnatt's Technique

As mentioned previously, Allnatt's scaling technique does not yield scale values on the S dimension. Instead, a value is produced on a t dimension which runs from 0 to 1. Thus, in order to use the multiple regression procedure, it is necessary to transform t values to S values via the inverse of equation (1) in chapter 3

$$S_i = \frac{(1 - \frac{1}{t_{m_i}})^{b/G}}{S_M} \qquad (1)$$

The values for b, G and S_M will, of course, be needed. The value for G is determined in the basic scaling procedure. The value for b can be found by carrying out a fractionation task as described in section 2.3.3 in chapter 3. The value for S_M is determined by realizing that

$$S_M = a\, D_M^{\,b}$$

The value for D_M is derived at the same time as that for G and the value for a is derived at the same time as the value for b. Thus, if the analysis described in chapter 3 is carried out fully, all the necessary information will be available to produce the S_is. From these, it's then possible to carry out all the analyses presented in this chapter.

One important thing to realize about Allnatt's technique, which isn't true about Thurstone's technique, concerns the nature of the S_is. In Thurstonian scaling, the S_is are only assumed to be values on an interval scale. Thus, they are no more than linear representations of the "true" scale values. Allnatt's technique assumes the S_i values are on a ratio scale. In particular, the assumption is made that Stevens' power law

$$S_i = a\, D_i^b$$

holds. Thus, the S_i values are, at most, multiples of the "true" scale values. This assumption, of course, represents a restriction on the present model but one that can be incorporated under certain circumstances.

For purposes of understanding under what circumstances Allnatt's ideas can fit into the present model, let's consider that there are three possible scenarios for how a given physical dimension, D, relates to the perceptual space. The first is that the MDS solution returns one corresponding perceptual dimension for each physical dimension. The second is where there is more than one perceptual dimension for a given physical dimension. (We don't have to worry if a physical dimension has no representation in perceptual space. This just means the observer regards it as irrelevant.) The third is where two or more physical dimensions amalgamate to form a single perceptual dimension.

In the first instance (a one-to-one relationship between physical dimensions and perceptual dimensions), let's consider each dimension separately for a minute. For each separate dimension, it's not overly difficult to maintain a power function relationship between the physical dimension and S. The requirement is simply that both the relationship between the physical and the perceptual and the relationship between the perceptual and S be power functions. That is, letting P be the value on the perceptual dimension, if $P = c\ D^{n}$ and $S = d\ P^{m}$ then

$$S = d\ (c\ D^{n})^{m} = d\ c^{m}\ D^{mn} = a\ D^{b}$$

where $b = mn$, and c^{m} is just a constant.

If m is 1, then the equation

$$S = d\ P^{m}$$

is nothing more than a weighting equation in one variable with a constant of zero. On an intuitive level, the equation would represent the observer simply using the perceptual dimension as the S dimension, although the P value may be multiplied by a constant. In many ways, this kind of use of the P dimension makes good sense. Assuming again that the observers have access to P, it's likely that they would keep the transformation between it and $\underline{S}$ as simple as possible. The only way to make it simpler would be to set d to 1 also.

The next step would be to determine what happens when physical dimensions having these single perceptual representations are combined. The model says the resultant scale value is a weighted sum of the perceptual coordinates. Thus, the scale value for a stimulus having coordinate $\underline{i}$ on dimension 1 and coordinate $\underline{j}$ on dimension 2 would be

$$S_{ij} = w_1 S_{i_1} + w_2 S_{j_2} + k$$

$$= w_1 c_1 D_{i_1}^{n_1} + w_2 c_2 D_{j_2}^{n_2} + k$$

where $\underline{k}$ must be assumed to be 0 if Allnatt is correct. However, Allnatt also has proposed a model equation for handling these kind of stimuli. He claims that it isn't the $\underline{S}$ values which are additive but a transform of these, the $\underline{J}_m$ s, where

$$J_{m_i} = \frac{1}{t_{m_i}} - 1$$

and, as before

$$t_{m_i} = \frac{1}{1+\left(\frac{S_i}{S_M}\right)^G} \qquad (2)$$

In other words, Allnatt's processing model suggests

$$J_{m_{ij}} = J_{m_i} + J_{m_j}$$

These two ideas, ours and Allnatt's, appear to be quite different on the surface. However, an examination of equation (1) suggests the relationship

$$S_i = \frac{J_{m_i}}{S_M}^{b/G}$$

If $\underline{b} \approx \underline{G}$ (which Allnatt suggests it may generally be - Allnatt, 1975) then the $\underline{J}_m$ s are nothing more than multiplicative functions of the scale values. Thus, if $\underline{b} \approx \underline{G}$ (and if the further restriction that $\underline{k}$ = 0 is invoked) the present processing model and Allnatt's processing model actually predict essentially the same thing. Certainly distinguishing between them empirically, even when $\underline{b}$ and $\underline{G}$ aren't identical, would be virtually impossible. Thus, in the circumstance in which every physical dimension is represented by a single perceptual dimension, Allnatt's model can be incorporated into the present conceptual framework.

In the present circumstance (scaling teletext systems), a one-to-one relationship between the physical and the perceptual may or may not arise. However, such isomorphic relationships probably are quite common in general. For example, most simple sensory manipulations (brightness, loudness, heaviness, etc.) will probably yield a one-dimensional perceptual representation. Thus, in each of these situations, it makes sense that there would be a power function relationship between *D* and *S*. There presumably would be a power function relationship between *D* and the one dimensional perceptual space and a very simple relationship between the perceptual space and *S*. (Again, note that the second transformation cannot even involve the addition of a constant or the ratio scale properties of *S* and, hence, the power function relationship, will be lost.)

One final issue with respect to physical dimensions that have one-dimensional perceptual representations should be mentioned. The nonmetric MDS programs return solutions in which the distances between values are only monotonically related to the dissimilarity ratings. If we assume that the dissimilarity ratings reflect distances on a ratio scale, then metric programs should produce the best representation of the perceptual dimension. However, the number of dissimilarity judgments per stimulus pair will be too small (maybe one per pair) to produce stable distance estimates. Thus, metric procedures are probably best avoided. Yet, nonmetric procedures may distort the perceptual space a bit since they treat the dissimilarity judgments as only monotonically related to true distance. Failure, then, to observer a power function between *D* and *P* may be partly attributable to the MDS procedure. The only way to guard against these problems would be to take a large number of measurements and to assure that the perceptual space is well mapped out before dealing with it.

We next have to consider the situation where a given physical dimension has no perceptual correlate but instead maps onto two or more perceptual dimensions. If this occurs there is no way to assure that there will be a power function between the physical dimension and *S*. In particular, suppose we have a physical dimension *D* which maps into a perceptual space having dimensions P_1 and P_2. Even if both P_1 and P_2 are power functions of *D* ($P_1=a_1D^{b_1}$ and $P_2=a_2D^{b_2}$) no additive mixture of the values on P_1 and P_2 can be a power function of *D*. In order to produce *S* values via a power function, it would be necessary to attach a weight of zero to one of the perceptual dimensions. (Even attaching a very small weight would disturb the power function relationship, although it's not clear that this could be detected empirically. As noted in chapter 3, testing for a power function relationship involves fitting a straight line to a set of points and a straight line fits just about anything.)

If observers do tend to handle these situations by attaching a weight of zero to one perceptual dimension, it should be possible to alter that strategy experimentally. In particular, the dimension with the zero weight would first have to be identified through a MDS analysis. Then, perhaps, a set of instructions could be created to force attention to that dimension. If this manipulation is successful a nonzero weight will now be attached to this dimension. Unless the observer then chooses to give zero weight to the other dimension the effect will be to alter the value of S (and, hence, of t) and destroy the power function relationship. In contrast, for physical dimensions which have only 1 related perceptual dimension, the scale values of S can only be altered by increasing or decreasing the single weight parameter. In other words, only a multiplicative change is possible. Since S is a ratio scale, changes of this sort would be irrelevant to the power function relationship. Further, although values on S would change, the t values should not. Reexamination of equation (2) demonstrates why this is so. In converting from an S_i value to a t value the S_i value is divided by a normalizing scale value, S_M (S_M is the scale value for the stimulus which has a t value of 1/2). If an experimental manipulation alters S_i multiplicatively, it will probably also alter S_M in the same fashion. Thus, neither the ratio nor the t value should change.

Finally, we must consider the situation where two (or more) dimensions (D_1 and D_2) amalgamate to form a single perceptual dimension. The question is how can these two dimensions combine and still maintain a power function relationship between each D_i and P when examined separately. The answer is, that only certain ways of combining the dimensions allow this to occur. For example, if they combine in an additive power function:

$$P = a (D_1 + D_2)^n$$

the relationship between *P* and either D_i will not reflect a power function. To see this, take logarithms of both sides:

$$\log P = \log a + n \log (D_1 + D_2)$$

If we have a power function between D_i and *P*, this must be the equation of a straight line when the other D_i equals a constant. In fact, it isn't a straight line relationship of D_1 or D_2 unless the other equals 0. On the other hand, if D_1 and D_2 combine in a multiplicative power function

$$P = a D_1^{n_1} D_2^{n_2} \qquad (3)$$

Its true that

$$\log P = \log a + n_1 \log D_1 + n_2 \log D_2$$

Here, there is a straight line relationship between log *P* and either log D_1 or log D_2. Thus, both physical dimensions would be related to *P* appropriately and, once again Allnatt's conceptualizaton could be incorporated into the present conceptual framework. The only issue the reader must take note of is that the $\underline{P}$ dimension which emerges from the MDS analysis should be reasonably well described by equation (3).

On a closing note, it should be mentioned that $\underline{S}_M$ is not regarded as a constant in Allnatt's conceptualization. In fact, it's viewed as the variable which creates all the context effects. If it could be shown that an experimental manipulation altered $\underline{t}$ values in Allnatt's procedure, he would initially try to explain this through changes in $\underline{S}_M$. For example, under normal circumstances, the stimulus which produces $\underline{S}_M$ ($\underline{D}_M$) would, presumably, be near the middle of the $\underline{D}$ dimension. However, if a narrow range of stimuli is used, $\underline{D}_M$ may tend to drift toward the middle of that range, altering all the $\underline{t}$ values. The ability of this one parameter ($\underline{S}_M$) to capture all the effects of experimental manipulations is, of course, an empirical question. However, until and unless the relevant experimentation and model-fitting are done, we have no reason to claim that our model with its many parameters explains context effects better than Allnatt's single parameter.

IV. GENERAL SUMMARY AND CONCLUSIONS

In the present manuscript we have attempted to do three things: acquaint the reader with the issues involved in selecting a scaling procedure, recommend and discuss the optimal scaling procedures for the present purposes, and provide a general processing model for how scaling operations are undertaken. In chapter 1 the first of these goals was accomplished with the suggestion being that, for present purposes, indirect scaling techniques are optimal. In chapter 2 the reader was introduced to the two indirect scaling techniques that appear to be most useful for present purposes, Thurstone's (1927) technique and Allnatt's (1973; 1975; 1979) technique. In chapters 3 and 4 these techniques were examined in considerably greater detail. Methods for examining the techniques themselves and for analyzing their results were discussed. Finally in chapter 5 we've put together a general processing model. Thus, the three goals of the project appear to have been accomplished. In closing, we'd like to provide a brief overview of the model as well as a brief discussion of what we feel are its important strengths.

A schematic diagram of the model is provided in Figures 9 and 10 (Figure 9 contains the general schematic. Figure 10 fills in some of the details). Any stimulus, $\underline{X}_i$, is presumed to give rise to a perceptual representation. This representation is assumed to be stable in that it is uninfluenced by context. Further, it's assumed to be characterized by a set of coordinates in an $\underline{n}$-dimension space ($\underline{x}_{i1}$, $\underline{x}_{i2}$, $\underline{x}_{i3}$ ... $\underline{x}_{in}$). The scale value is created from this representation by taking a weighted sum of these coordinates as follows

$$S_i = w_1 x_{i1} + w_2 x_{i2} + \ldots + w_n x_{in} + k$$

where the $\underline{w}_j$ represent the weights and the $\underline{k}$ is a scaling constant.

The process of assigning weights is the heart of the model. It is here that essentially all effects of experimental manipulations are manifest. Different types of judgments (e.g., acceptability vs. pleasantness, heaviness vs. denseness) will undoubtedly produce different weighting schemes. Further, different contexts for making the same judgments (e.g., instructions to rate the stimuli as an engineer would vs. instructions to rate the stimuli as an everyday viewer would) would also produce differences in the basic weighting scheme.

Once $\underline{S}_i$ has been established, it is used to produce a response on the response scale provided by the experimenter. In order to maintain continuity with previous research, it is recommended that a 1 to 5 (i.e., $\underline{A}$ to $\underline{E}$) rating scale be used. The model, as stated, makes no assumptions about the conversion from the $\underline{S}$ dimension to a response. However, Thurstonian analysis is perfectly compatible with all the assumptions of the mdoel and is, thus, regarded as the best "back-end" of the process. The $\underline{S}_i$s are assumed to reflect average rather than deterministic values on a newly created $\underline{S}$ dimension. That is, there is random variability in the actual value, $\underline{Y}_i$, that a given stimulus produces on $\underline{S}$. The $\underline{S}$ dimension is presumed to be divided into $\underline{n}$ segments by $\underline{n}$-1 criteria. The response given corresponds to the segment into which $\underline{Y}_i$ falls at the moment the stimulus is being evaluated. Because of the variability about $\underline{S}_i$ the same stimulus will not always provide a value in the same segment and, thus, will lead to different responses under identical circumstances.

In general, Thurstonian analysis is preferred to Allnatt's analysis as a back-end to the model. In order for Allnatt's analysis to be applicable, the model would need a number of additional assumptions, detailed earlier in this chapter, which seem unnecessarily restrictive. In essence, if Allnatt's analysis is an accurate description it necessitates a somewhat different model as a front-end. In fact, Allnatt's theorizing has been extended to include a discussion of some of these front-end processes. Thus, one could say his model already has an implicit front-end. Obviously, we feel the model, as proposed, is the best way of viewing the process at present. What follows is a discussion of the model's strengths.

One obvious strength of the model is it represents an attempt to integrate a number of different lines of successful psychological research. Earlier in the present chapter the support for the idea of a stable perceptual representation was discussed. The notion of a multidimensional space underlying the perceptual representation of stimuli has received considerable support in recent years (Carroll and Wish, 1974; Shepard, 1963). Anderson's (1968) notion that perceptual dimensions are combined to form a weighted sum has also proven to be a quite viable one, as has the idea of differential weighting of the dimensions. Finally, the model even has the flexibility to encompass Stevens' (1961) findings that the nature of the function relating physical dimensions and psychological dimensions is a power function for many types of stimuli.

A second strength of the model is its generality. That is, it's applicable for virtually every type of scaling judgment and it provides a framework for understanding how different scaling judgments are related to one another. In much of the work on scaling, the *S* dimension, regardless of what it represents - prettiness, acceptability, artistic value - is taken as a primitive. The models, in some sense, evolve around the judgment itself (e.g., Allnatt's theorizing is a case in point). In the present model the primitives are not the *S* dimensions but the perceptual representations of the stimuli. Regardless of the nature of the judgment the observers are about to make, observers are assumed to perceive the stimuli in a stable fashion. The *S* dimension that the experiment calls for is then created from a weighting of the perceptual dimensions. Thus, prettiness ratings would involve one set of weights while acceptability ratings would involve another. In essence, any kind of scaling judgment about any stimulus that can be perceived, can fit into this framework. Further, the framework provides a means of understanding how and why the different judgments are different. By examining the different weighting schemes used when making, for example, prettiness judgments on the one hand versus acceptibility judgments on the other, we can gain a better understanding of what perceptual dimensions are most relevant to each of the judgments.

The flexibility of changing weights as a result of experimental manipulations also has implications for how context can influence judgments. As noted in discussing Thurstonian analysis, one effect of context could be to create different placements of criteria on the S dimension. However, it's also possible that different contexts can actually produce different values for the S_is. In the present framework the explanation of this result would be quite straightforward. The observers simply selected a different set of weights in the two different contexts, perhaps emphasizing dimension A in context 1 and dimension B in context 2. Again, by looking at how the weights change, information can be gained about how various perceptual dimensions relate to various contextual manipulations. In Allnatt's framework there is also a way of accounting for the effects of context, the physical value of the normalizing stimulus, D_M (see equation 1 in chapter 3). However, as mentioned previously, it seems quite unlikely that variations in D_M alone could account for contextual effects beyond those Allnatt and Corbett (1972) have already investigated. Whether there will be effects beyond those is an empirical question. If so, more work on the front-end of their model will be necessary.

A third strength of the model, and one which is not immediately obvious, is it testability. As Figure 9 makes clear, the model is a stage model. The first stage, the perceptual stage, and the third stage, the response stage, can be empirically evaluated independent of the other stages. The means of determining whether the response process can be described in terms of Thurstonian analysis has been discussed extensively in chapter 3. The perceptual stage can be evaluated by considering the fit of the multidimensional solution of the scaling analysis described in the present chapter. If the STRESS is low and the dimensions are reasonable the notion of a workable perceptual representation is supported. The second stage, in which the $\underline{S}_i$ are determined, is not independent of stage 1. An adequate evaluation of stage 2 requires that the perceptual representations of the stimuli be known. If the evaluation of stage 1 was successful, this information will be available. If not, it won't. Thus, the model really does hinge on the notion that each stimulus provides a stable perceptual representation. If this assumption is incorrect, the model as it stands could be rejected.

On the basis of the argument just presented we feel that the model offers a substantial advance over anything that has gone before it in the area of evaluating teletext systems. The model is general, yet testable, and it integrates a number of lines of psychological research. Putting Thurstone's response model at the back end of ours completes the picture. Trying to attach Allnatt's response model is much less desirable. Generality would be lost as the assumptions of our model were made more restrictive and, as the reader may remember, Allnatt's technique itself requires that the stimuli vary on a quantitative dimension. Since the psychological literature does not seem to contain any empirical demonstrations that any of these restrictions are necessary, Allnatt's model must remain a second choice. The only cost of our model is in terms of the time and effort involved in determining the nature of the psychological space. That is, a number of subjects will have to make a large number (>3000) of judgments of stimulus pairs. However, once this extensive data collection is complete it won't need to be done again. Thus, overall, our feeling is that the positive aspects of the model far outweigh any of the difficulties involved in testing it.

REFERENCES

Allnatt, J. Opinion-distribution model for subjective rating studies. *International Journal of Man-Machine Studies*, 1973, *5*, 1-15.

Allnatt, J. Subjective rating and apparent magnitude. *International Journal of Man-Machine Studies*, 1975, *7*, 801-816.

Allnatt, J. Television measurements through psychophysics to subjective picture quality. *Radio and Electronic Engineer*, 1979, *49*, 611-619.

Allnatt, J. W. and Corbett, J.M. Adaptation in observers during television quality-grading tests. I. Adaptation as a function of the conditioning situation. *Ergonomics*, 1972, *15*, 353-365.

Anderson, N.H. A simple model for information integration. In R.P. Ableson, E. Aronson, W.J. McGuire, T.M. Newcomb, M.J. Rosenburg and P.H. Tannenbaam (Eds.), *Theories of cognitive consistency: A sourcebook.* Chicago: Rand McNally, 1968.

Bartko, J.J. On various intraclass correlation reliability coefficients. *Psychological Bulletin*, 1976, *83*, 762-765.

Bjork, E.L. and Estes, W.K. Letter identification in relation to linguistic context and masking conditions. *Memory and Cognition*, 1973, *1*, 217-223.

Box, G.E.P. Non-normality and tests on variances. *Biometrika*, 1953, *40*, 318-335.

Bruner, J.S. On perceptual readiness. *Psychological Review*, 1957, *64*, 123-152.

Budescu, D. The power of the F test in normal populations with heterogeneous variances. *Educational and Psychological Measurement*, 1982, *42*, 409-416.

Brown, M.B. and Forsythe, A.B. Robust tests for homogeneity of variance. *Journal of the American Statistical Association*, 1974, *69*, 364-368.

CCIR. Method for subjective assessment of the quality of television pictures. 1974.

Carroll, J.D., and Wish, M. Multidimensional perceptual models and measurement methods. In E.C. Carterette and M.P. Friedman (Eds.), *Handbook of Perception*, Vol. 2., New York: Academic Press, 1974.

Cavanaugh, J.R., Hatch, R.W. and Sullivan, J.L. Models for the subjective effects of loss, noise, and talker echo on telephone connections. *Bell System Technical Journal*, 1976, *55*, 1319-1371.

Church, J.D. and Wike, E.L. The robustness of homogeneity of variance tests for asymmetric distributions: A Monte Carlo study. *Bulletin of the Psychonomic Society*, 1976, *7*, 417-420.

Fagot, R.F. A theory of relative judgment. *Perception and Psychophysics*, 1978, *24*, 243-252.

Freeman, M.F. and Tukey, J.W. Transformations related to the angular and the square root. *Annals of Mathematical Statistics*, 1950, *21*, 607-611.

Games, P.A. Curvilinear transformations of the dependent variable. *Psychological Bulletin*, 1983, *93*, 382-387.

Games, P.A., Keselman, H.J. and Clinch, J.J. Tests for homogeneity of variance in factorial designs. *Psychological Bulletin*, 1979, *86*, 978-984.

Games, P.A., Winkler, H.B., and Probert, D.A. Robust tests for homogeneity of variance. *Educational and Psychological Measurement*, 1972, *32*, 887-909.

Gartside, P.S. A study of methods for comparing several variances. *Journal of the American Statistical Association*, 1972, *67*, 342-346.

Geisser, S. and Greenhouse, S.W. An extension of Box's results on the use of the F distribution in multivariate analysis. *Annals of Mathematical Statistics*, 1958, *29*, 885-891.

Glass, G.V., Peckham, P.D., and Sanders, J.R. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 1972, *42*, 237-288.

Green, D.M. and Swets, J.A. *Signal detection theory and psychophysics*. New York: Wiley, 1966.

Helson, H. Adaptation-level as a basis for a quantitative theory of frames of reference. *Psychological Review*, 1948, *55*, 297-313.

Keppel, G. *Design and Analysis: A Researcher's Handbook*, 2nd edition. New Jersey: Prentice-Hall, Inc., 1982.

Kirk, R.E. *Experimental Design: Procedures for the Behavioral Sciences*, 2nd edition. Monterey, Calif.: Brooks/Cole, 1982.

Kornbort, D.W. Theoretical and empirical comparisons of Luce's choice model and logistic Thurstone model of categorical judgment. *Perception and Psychophysics*, 1978, *24*, 193-208.

Kort, B. Models and methods for evaluating customer acceptance of telephone connections. *IEEE Journal*, 1983, 20.6.1-9.

Krueger, L.E. and Shapiro, R.G. Letter detection with rapid serial visual presentation: Evidence against word superiority at feature extraction. *Journal of Experimental Psychology: Human Perception and Performance*, 1979, *5*, 657-673.

Krusal, W.H. and Wallis, W.A. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 1952, *47*, 583-621.

Kruskal, J.B. Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis. *Psychometrika*, 1964, *29*, 1-28, 115-129.

Levine, D.W. and Dunlap, W.P. Power of the F test with skewed data: Should we transform or not? *Psychological Bulletin*, 1982, *92*, 272-280.

Levine, D.W. and Dunlap, W.P. Data transformation, power, and skew: A rejoiner to Games. *Psychological Bulletin*, 1983, *93*, 596-599.

Lindman, H.R. *Analysis of Variance in Complex Experimental Designs*. San Francisco: W.H. Freeman and Co., 1974.

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

Luce, R.D. The choice axiom after twenty years. *Journal of Mathematical Psychology*, 1977, *15*, 215-223.

Lui, P.C. and Ebert, J.G. GOSPAK - A computer package for transmisssion performance studies. *IEEE National Telecommunications Conference*, 1976, *12*, 23.3-1-5.

Marascuilo, L.A. and McSweeney, M. *Nonparametric and Distribution-Free Methods for the Social Sciences*. Monterey, Calif.: Brook/Cole, 1977.

Martin, C.G. and Games, P.A. ANOVA tests for homogeneity of variance: Nonnormality and unequal samples. *Journal of Educational Statistics*, 1977, *2*, 187-206.

Massaro, D.W. Letter information and orthographic context in word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 1979, *5*, 595-609.

McGinnies, E. Emotionality and perceptual defence. *Psychological Review*, 1949, *56*, 244-251.

Meyer, D.E. and Schvaneveldt, R.W. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 1971, *90*, 227-234.

Meyer, D.E., Schvaneveldt, R.W. and Ruddy, M.G. Loci of contextual effects on visual word recognition. In P.M.A. Rabbitt and S. Dornic (Eds.) *Attention and Performance V*. New York: Academic Press, 1975.

Miller, R.G., Jr. Jackknifing variances. *Annals of Mathematical Statistics*, 1968, *39*, 567-582.

Mosteller, F. and Bush, R.R. Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of Social Psychology*. Reading, Mass.: Addison-Wesley, 1954.

Parducci, A. Category judgment: A range-frequency model. *Psychological Review*, 1965, *72*, 407-418.

Posner, M.I. and Snyder, C.R.R. Attention and cognitive control. In R.L. Solso (Ed.) *Information Processing and Cognition: The Loyola Symposium*. Hillsdale, N.J.: Erlbaum, 1975.

Reicher, G.M. Perceptual recognition as a function of meaningfulnes of stimulus material. *Journal of Experimental Psychology*, 1969, *81*, 275-281.

Schiffman, S.S., Reynolds, M.L. and Young, F.W. *Introduction to multidimensional scaling*. New York: Academic Press, 1981.

Scheffé, H. *The Analysis of Variance*. New York: Wiley, 1959.

Shepard, R.N. The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika*, 1962, *27*, 125-140, 219-246.

Shepard, R.N. Analysis of proximities as a technique for the study of information processing in man. *Human Factors*, 1963, *5*, 33-48.

Shiffrin, R.M. and Schneider, W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, 1977, *84*, 127-190.

Stevens, S.S. On the new psychophysics. *Scandanavian Journal of Psychology*, 1960, *1*, 27-35.

Stevens, S.S. The psychology of sensory function. In W.A. Rosenlith (Ed.) *Sensory communication*. Cambridge, Mass.: MIT Press, 1961.

Tanner, W.P. and Swets, J.A. A decision-making theory of visual detection. *Psychological Review*, 1954, *61*, 401-409.

Thompson, M.C. and Massaro, D.W. Visual information and redundancy in reading. *Journal of Experimental Psychology*, 1973, *98*, 49-54.

Thurstone, L.L. A law of comparative judgment. *Psychological Review*, 1927, *34*, 273-286.

Torgerson, W.S. Multidimensional scaling: 1. Theory and method. *Psychometrika*, 1952, *17*, 401-419.

Togerson, W.S. *Theory and methods of scaling*. New York: Wiley, 1958.

Wheeler, D.D. Processes in word recognition. *Cognitive Psychology*, 1970, *1*, 59-85.

Wike, E.L. and Church, J.D. Nonrobustness in F tests: 1. A replication and extension of Bradley's study. *Bulletin of the Psychonomic Society*, 1982, *20*, 165-167.

Winer, B.J. *Statistical Principles in Experimental Design*, 2nd edition. New York: McGraw-Hill, 1971.

ADDITIONAL REFERENCES

Aaron, M. Effect of the menstral cycle on subjective ratings of sweetness. *Perceptual and Motor Skills*, 1975, *40*, 974.

Baird, J.C., Green, D.M. and Luce, R.D. Variability and sequential effects in cross-modality matching of area and loudness. *Journal of Experimental Psychology: Human Perception and Performance*, 1980, *6*, 277-289.

Bartoshuk, L.M. Water taste in man. *Perception and Psychophysics*, 1968, *3*, 69-72.

Bechtel, G.G. Folded and unfolded scaling from preferential paired comparisions. *Journal of Mathematical Psychology*, 1968, *5*, 333-357.

Berstein, I.H., Lin, T.D. and McClellan, P. Cross versus within racial judgments of attractiveness. *Perception and Psychophysics*, 1982, *32*, 495-503.

Blackwood, L.G. and Carpenter, E.H. The importance of anti-urbanism in determinining residential preferences and migration patterns. *Rural Sociology*, 1978, *43*, 31-47.

Boller, F. and Marcie, P. Possible role of abnormal auditory feedback in conduction aphasia. *Neuropsychologia*, 1978, *16*, 521-524.

Boller, F. Vrtunski, P.B., Younjai, K. and Mark, J.L. Delayed auditory feedback and aphasia. *Cortex*, 1978, *14*, 212-226.

Borich, G.D. Preferences for color, form, borders, lines and dots by preschool children and adults. *Perceptual and Motor Skills*, 1970, *31*, 811-817.

Cardello. A.V. Comparison of taste qualities elicited by tactile, electrical, and chemical stimulation of single human taste papillae. *Perception and Psychophysics*, 1981, *29*, 163-169.

Coombs, C.H. On the use of inconsistency of preferences in psychological measurement. *Journal of Experimental Psychology*, 1958, *55*, 1-7.

Coombs, C.H., Donnell, M.L. and Kirk, D.B. An experimental study of risk preference in lotteries. *Journal of Experimental Psychology: Human Perception and Performance*, 1978, *4*, 497-512.

Coombs, C.H. and Huang, L.C. Tests of the betweeness property of expected utility. *Journal of Mathematical Psychology*, 1976, *13*, 323-337.

Coombs, C.H. and Huang, L.C. Polynomial psychophysics of risk. *Journal of Mathematical Psychology*, 1970, *7*, 317-338.

Coombs, C.H. and Huang, L. Tests of a portfolio theory of risk preference. *Journal of Experimental Psychology*, 1970, *85*, 23-29.

Coombs, C.H. and Lehner, P.E. Evaluation of two alternative models of a theory of risk: I. Are moments of distributions useful in assessing risk? *Journal of Experimental Psychology: Human Perception and Performance*, 1981, *7*, 1110-1123.

Corbett, J.M., Taylor, J.R. and Allnatt, J.W. Subjective quality of colour-television pictures impaired by video crosstalk. *Proceedings. The Institution of Electrical Engineers*, 1969, *116*, 181-184.

Coren, S. and Miller, J. Size contrast as a function of figural similarity. *Perception and Psychophysics*, 1974, *16*, 355-357.

Coren, S. and Porac, C. The validity and reliability of self-report items for the measurement of lateral preference. *British Journal of Psychology*, 1978, *69*, 207-211.

Coren, S., Porac, C. and Duncan, P. Lateral preference behaviors in preschool children and young adults. *Child Development*, 1981, *52*, 443-450.

Craig, K.D., Best, H. and Ward, L.M. Social modeling imfluences on psychophysical judgments of electrical stimulation. *Journal of Abnormal Psychology*, 1975, *84*, 366-373.

Crowley, P.M. Effect of training upon objectivity of moral judgment in grade-school children. *Journal of Personality and School Psychology*, 1968, *8*, 228-232.

Curtis, D.W. and Rule, S.J. Judgment of duration relations: simultaneous and sequential presentation. *Perception and Psychophysics*, 1977, *22*, 578-584.

Dapolito, F., Guttenplan, H. and Steinitz, H. Accuracy of subjective judgments of information in long-term memory. *Psychonomic Science*, 1968, *13*, 227-228.

Day, H. Evaluations of subjective complexity, pleasingness and interestingness for a series of random polygons varying in complexity. *Perception and Psychophysics*, 1967, *2*, 281-286.

Dean, C.E. Measurements of the effects of interference in television reception. *Proceedings of the Institution of Radio Engineers*, 1960, *48*, 1035-1049.

Feeley, J.T. Content interests and media prefeences of middle-graders: Differences in a decade. *Reading World*, 1982, *22*, 11-16.

Fredendall, G.L. and Behrend, W.L. Picture quality- procedures for evaluating subjective effects of interference. *Proceedings of the Institution of Radio Engineers*, 1960, *48*, 1030-1034.

Frijters, J.E.R. Three-stimulus procedures in olfactory psychophysics: An experimental comparison of Thurstone-Ura and three alternative forced choice models of signal detection theory. *Perception and Psychophysics*, 1980, *28*, 390-397.

Golding, S.L and Rorer, L.G. Illusory correlation and subjective judgment. *Journal of Abnormal Psychology*, 1972, *80*, 249-260.

Gordon, H.L and Hooker, C.A. Opinions of alcholholics concerning the effectiveness of various treatment methods. *Newsletter for Research in Psychology*, 1969, *11*, 24-26.

Gottfredson, S.D. Evaluating psychological research reports. Dimensions, reliability, and correlates of quality judgments. *American Psychologist*, 1978, *33*, 920-934.

Grazin, K.L. and Williams, R.H. Patterns of behavioral characteristics as indicants of recreation preferences: A canonical analysis. *Research Quarterly*, 1978, *49*, 135-145.

Green, D.M., Luce, R.D. and Duncan, J.E. Variability and sequential effects in magnitude production and estimation of auditory intensity. *Perception and Psychophysics*, 1977, *22*, 450-456.

Griswold, B.J. and Luce, R.D. Choices amoung uncertain outcomes: A test of a decomposition and two assumptions of transitivity. *American Journal of Psychology*, 1962, *75*, 35-44.

Hallworth, H.J. and Waite, G. A factorial study of value judgments amoung adolescent girls. *British Journal of Psychology*, 1963, *16*, 37-46.

Hardiman, G.W. and Zernich, T. Preferences for the visual arts: A review of recent studies. *Perceptual and Motor Skills*, 1977, *44*, 455-463.

Henry, D.L. and Jacobs, K.W. Color eroticism and color preference. *Perceptual and Motor Skills*, 1978, *47*, 106.

Hocevar, D. A comparison of statistical infrequency and subjective judgment as criteria in the measurement of originality. _Journal of Personality Assessment_, 1979, _43_, 297-299.

Houlden, P., La Tour, S., Walker, L. and Thibaut, J. Preference for modes of dispute resolution as a function of process and decision control. _Journal of Experimental and Social Psychology_, 1978, _14_, 15-30.

Jesteadt, W., Luce, R.D. and Green, D.M. Sequential effects in judgments of loudness. _Journal of Experimental Psychology: Human Perception and Performance_, 1977, _3_, 92-104.

Juhasz, J.B. and Paxson, L. Personality and preference for painting style. _Perceptual and Motor Skills_, 1978, _46_, 347-349.

Kahneman, D. and Tversky, A. On the psychology of prediction. _Psychological Review_, 1973, _80_, 237-251.

Kaplan, S., Kaplan, R. and Wendt, J.S. Rated preference and complexity for natural and urban visual material. _Perception and Psychophysics_, 1972, _12_, 354-356.

Kornbrot, D.E. Theoretical and empirical comparisons of Luce's choice model and logistic Thurstone model of categorical judgment. _Perception and Psychophysics_, 1978, _24_, 193-208.

Krantz, D.H. and Tversky, A. Similarity of rectangles: An analysis of subjective dimensions. _Journal of Mathematical Psychology_, 1975, _12_, 4-34.

Laming, D. Luce's choice axiom compared with choice-reaction data. _British Journal of Mathematical and Statistical Psychology_, 1977, _30_, 141-153.

Laponce, J.A. Measuring party preference: The problem of ambivalence. _Canadian Journal of Political Science_, 1978, _11_, 139-152.

Laughlin, P.R. and Laughlin, R.M. Source effects in the judgment of social argot. _Journal of Social Psychology_, 1969, _78_, 249-254.

Lessman, A.M. The subjective effects of echoes in 525-line monochrome and NTSC color television and the resulting echo time weighting. _Journal of the Society of Motion Picture and Television Engineering_, 1972, _81_, 907-916.

Link, S.W. The relative judgment theory of two choice response time. _Journal of Mathematical Psychology_, 1975, _12_, 114-135.

Link, S.W. Applying RT deadlines to discrimination reaction time. _Psychonomic Science_, 1971, _25_, 355-358.

Link, S.W. and Tindall, A.D. Speed and accuracy in comparative judgments of line length. *Perception and Psychophysics*, 1971, *9*, 284-288.

Lu, K.H. A measure of agreement amoung subjective judgments. *Educational and Psychological Measurement*, 1971, *31*, 75-84.

Luce, R.D. and Mo, S.S. Magnitude estimation of heaviness and loudness by individual subjects: A test of a probabilistic response theory. *British Journal of Mathematical and Statistical Psychology*, 1965, *18*, 159-174.

Maloney, K.B. and Hopkins, B.L. The modification of sentence structure and its relationship to subjective judgments of creativity in writing. *Journal of Applied Behavior Analysis*, 1978, *6*, 425-433.

Marks, L.E. *Sensory Processes: The New Psychophysics*. New York: Acadimic Press, 1974.

O'Mahony, M. A note on the reliability of subjecive judgment in discerning when the mouth is clear of salt taste stimuli. *Perception and Psychophysics*, 1973, *14*, 437-439.

Parducci, A. Category judgment: A range-frequency model. *Psychological Review*, 1965, *72*, 407-418.

Pincus, S. and Waters, L.K. Informational social influence and product quality judgments. *Journal of Applied Psychology*, 1977, *62*, 615-619.

Prosser, R.D. and Allnatt, J.W. Subjective quality of television pictures impaired by random noise. *Proceedings. The Institution of Electrical Engineers*, 1965, *112*, 1099-1102.

Rapoport, A. Choice behavior in a Markovian decision task. *Journal of Mathematical Psychology*, 1962, *5*, 163-181.

Rapoport, A. and Tversky, A. Choice behavior in an optional stopping task. *Organizational Behavior and Human Performance*, 1970, *5*, 105-120.

Rapoport, A. and Tversky, A. Cost and accessibility of offers as determinants of optional stopping. *Psychonomic Science*, 1966, *4*, 145-146.

Reid, G.M. Subjective tests on visual telecomminications standards. *Proceedings. Institution of Electrical Engineers*, 1980, *127*, 3-8.

Richards, I.G. An analysis of individual differences in similarity judgments about complex random forms. *Perception and Psychophysics*, 1972, *11*, 143-149.

Rosenberg, S. and Gordon, A. Identification of facial expressions from affective descriptions: A probabilistic choice analysis of referential ambiguity. *Journal of Personality and Social Psychology*, 1968, *10*, 157-166.

Rumelhart, D.L and Greeno, J.G. Similarity between stimuli. *Journal of Mathematical Psychology*, 1971, *8*, 370-381.

Rump, E.E. Is there a general factor of preference for complexity? *Perception and Psychophysics*, 1968, *3*, 346-348.

Seaton, B. and Vogel, R.H. Conflicts, trade-offs and preference measurements. *Health Sciences Research*, 1978, *13*, 146-156.

Shanteau, J.C. and Anderson, N.H. Test of a conflict model for preference judgment. *Journal of Mathematical Psychology*, 1969, *6*, 312-325.

Shapiro, Z. and Venezia, I. On the aggregation across subjects in analyzing individual choice behaviour. *Journal of Mathematical Psychology*, 1977, *16*, 60-67.

Shevell, S.K. and Atkinson, R.C. A theoretical comparison of list scanning models. *Journal of Mathematical Psychology*, 1974, *11*, 79-106.

Slater, P. The analysis of personal preference. *British Journal of Statistical Psychology*, 1960, *13*, 119-135.

Stang, D,J, Castellaneta, J.A., Constantinidis, G. and Fortuno, C.R. Actual versus perceived talkativeness as determinants of judged leadership, popularity and likeableness. *Bulletin of the Psychonomic Society*, 1976, *8*, 44-46.

Stevens, S.S. On the new psychophysics. *Scandanavian Journal of Psychology*, 1960, *1*, 27-35.

Taylor, J.R., Bragg, E.J.W. and Corbett, J.M. Subjective quality of visual telephone pictures impaired by video crosstalk. *Proceedings. Institution of Electrical Engineers*, 1977, *124*, 987-992.

Tursky, B. and O'Connell, D. Reliability and interjudgment predictability of subjective judgments of electrocutaneous stimulation. *Psychphysiology*, 1972, *9*, 290-295.

Tursky, B. and Watson, P.D. Controlled physical and subjective intensities of electric shock. *Psychophysiology*, 1964, *1*, 151-162.

Tversky, A. Features of similarity. _Psychological Review_, 1977, _84_, 327-352.

Tversky, A. Elimination by aspects: A theory of choice. _Psychological Review_, 1972, _79_, 281-299.

Tversky, A. Intransitivity of preferences. _Psychological Review_, 1969, _76_, 31-48.

Tversky, A. and Kahneman, D. Availability: A heuristic for judging frequency and probability. _Cognitive Psychology_, 1973, _5_, 207-232.

Tversky, A. and Krantz, D.H. Similarity of schematic faces: A test of interdimensional additivity. _Perception and Psychophysics_, 1969, _5_, 124-128.

Tversky, A. and Sattath, S. Preference trees. _Psychological Review_, 1979, _86_, 542-573.

Valenzi, E.R. and Andrews, I.R. Effects of price information on product quality ratings. _Journal of Applied Psychology_, 1971, _55_, 87-91.

Vrtunski, P.B., Martinez, M. and Boller, F. Evaluation of delayed auditory feedback (DAF) effect: Comparison between subjective judgments and objective measures. _Cortex_, 1979, _15_, 337-341.

Ward, W.D. Musical perception. In J.V. Tobias (ed.), _Foundations of Modern Auditory Theory_. New York: Academic Press, 1970.

Ward, L.M. and Lockhead, G.R. Response system processes in absolute judgments. _Perception and Psychophysics_, 1971, _9_, 73-78.

White, T.A. and Reid, G.M. Quality of PAL colour television pictures impaired by random noise: Stability of subjective assessment. _Proceedings. Institution of Electrical Engineers_, 1981, _127_, 231-236.

LIST OF TABLES

Table 1. Sample Data Matrix. (Cell entries are frequencies.)

		Categories					
		1	2	3	4	5	M_i
Stimuli	X_0	87	12	1	0	0	1.14
	X_1	5	16	30	30	19	3.42
	X_2	10	20	40	18	12	3.02
	X_3	2	8	20	30	40	3.98
	X_4	1	5	18	22	54	4.23
	X_5	20	30	30	20	0	2.50
	X_6	0	0	0	9	91	4.91

Table 2. Rearranged Data Matrix

		Categories					
		1	2	3	4	5	M_i
Stimuli	X_4	1	5	18	22	54	4.23
	X_3	2	8	20	30	40	3.98
	X_1	5	16	30	30	19	3.42
	X_2	10	20	40	18	12	3.02
	X_5	20	30	30	20	0	2.50

Table 3. Probability Matrix

		Categories				
		1	2	3	4	5
Stimuli	X_4	.01	.05	.18	.22	.54
	X_3	.02	.08	.20	.30	.40
	X_1	.05	.16	.30	.30	.19
	X_2	.10	.20	.40	.18	.12
	X_5	.20	.30	.30	.20	.00

Table 4. Cumulative Probability Matrix

		Categories			
		1	2	3	4
	X_4	.01	.06	.24	.46
	X_3	.02	.10	.30	.60
Stimuli	X_1	.05	.21	.51	.81
	X_2	.10	.30	.70	.88
	X_5	.20	.50	.80	1.00

Table 5. Z-Score Matrix

		Categories			
		1	2	3	4
Stimuli	X_4	-2.327	-1.555	-.706	-.100
	X_3	-2.054	-1.282	-.524	+.253
	X_1	-1.645	-.806	+.025	+.878
	X_2	-1.282	-.524	+.524	+1.175
	X_5	-.841	.000	+.841	∞

Table 6. Criterion Position Matrix

		Categories			
		1	2	3	4
Stimuli	X_4	-1.670	-.813	+.130	+.803
	X_3	-1.648	-.803	+.026	+.876
	X_1	-1.645	-.806	+.025	+.878
	X_2	-1.630	-.880	+.158	+.803
	X_5	-1.644	-.809	+.026	∞
	$\bar{t}_g$	-1.647	-.822	+.073	+.840

Table 7. Relative Positions of the Criteria

		Criterion			
		1	2	3	4
Stimuli	X_4	-2.23	-1.51	-.74	-.07
	X_3	-1.98	-1.25	-.46	+.22
	X_1	-1.64	-.81	+.02	+.88
	X_2	-1.20	-.40	+.46	+1.10
	X_5	-.76	+.05	+.92	+1.65

Table 8. Expected Percentages in Each Category

		Categories				
		1	2	3	4	5
Stimuli	X_4	.0129	.0526	.1641	.2425	.5279
	X_3	.0239	.0817	.2172	.2643	.4129
	X_1	.0505	.1585	.2990	.3026	.1894
	X_2	.1151	.2295	.3326	.1871	.1357
	X_5	.2236	.2963	.3013	.1293	.0495

Table 9. Expected Frequencies in Each Category

		Categories				
		1	2	3	4	5
Stimuli	X_4	1.3	5.3	16.4	24.2	52.8
	X_3	2.4	8.2	21.7	26.4	41.3
	X_1	5.0	15.8	29.9	30.3	18.9
	X_2	11.5	23.0	33.3	18.7	13.6
	X_5	22.4	29.6	30.1	12.9	5.0

Table 10. Altered Matrices

Expected Frequency Matrix

		Categories				
		1	2	3	4	5
Stimuli	X_4	---	6.6	16.4	24.2	52.8
	X_3	---	10.6	21.7	26.4	41.3
	X_1	5.0	15.8	29.9	30.3	18.9
	X_2	11.5	23.0	33.3	18.7	13.6
	X_5	22.4	29.6	30.1	12.9	5.0

Obtained Frequency Matrix

		Categories				
		1	2	3	4	5
Stimuli	X_4	---	6	18	22	54
	X_3	---	10	20	30	40
	X_1	5	16	30	30	19
	X_2	10	20	40	18	12
	X_5	20	30	30	20	0

Table 11. Inversions (reciprocals) of the Probabilities in Table 4

		Categories			
		1	2	3	4
	X_4	100.000	16.667	4.167	2.174
	X_3	50.000	10.000	3.333	1.667
Stimuli	X_1	20.000	4.762	1.961	1.235
	X_2	10.000	3.333	1.429	1.136
	X_5	5.000	2.000	1.250	1.000

Table 12. Values from Table 11 Minus 1.00

		Categories			
		1	2	3	4
	X_4	99.000	15.667	3.167	1.174
	X_3	49.000	9.000	2.333	.667
Stimuli	X_1	19.000	3.762	.961	.235
	X_2	9.000	2.333	.429	.136
	X_5	4.000	1.000	.250	.000

Table 13. Natural Logarithms of Values in Table 12

		Categories			
		1	2	3	4
Stimuli	X_4	4.595	2.752	1.153	.160
	X_3	3.892	2.197	.847	-.405
	X_1	2.944	1.325	-.040	-1.448
	X_2	2.197	.847	-.847	-1.992
	X_5	1.386	.000	-1.386	$-\infty$

Table 14. Parameter Values for Allnatt's Basic Technique

Estimates of g

X_5	X_2	X_1	X_3	X_4
1.543	1.557	1.595	1.559	1.629

$\bar{g}$ = 1.577

Estimates of Tm (intercept estimates divided by $\bar{g}$)

X_5	X_2	X_1	X_3	X_4
-.452	.032	.393	1.036	1.373

Estimates of tm

X_5	X_2	X_1	X_3	X_4
.388	.508	.597	.738	.796

Estimates of Jm (|1/tm|-1)

X_5	X_2	X_1	X_3	X_4
1.578	.969	.675	.355	.253

Table 15. Sample Ratio Matrix

		Larger Stimulus			
		b	c	d	e
Smaller Stimulus	a	.80	.40	.20	.10
	b		.75	.50	.25
	c			.70	.40
	d				.75

Table 16. Tetrads and Their Associated Ratio Products

Tetrads	$R_{14}xR_{23}$	$R_{13}xR_{24}$
(a,b,c,d)	$R_{ad}xR_{bc}$=(.20)(.75) =.15	$R_{ac}xR_{bd}$=(.40)(.50) =.20
(a,b,c,e)	$R_{ae}xR_{bc}$=(.10)(.75) =.075	$R_{ac}xR_{be}$=(.40)(.25) =.10
(a,b,d,e)	$R_{ae}xR_{bd}$=(.10)(.50) =.05	$R_{ad}xR_{be}$=(.20)(.25) =.05
(a,c,d,e)	$R_{ae}xR_{cd}$=(.10)(.70) =.07	$R_{ad}xR_{ce}$=(.20)(.40) =.08
(b,c,d,e)	$R_{be}xR_{cd}$=(.25)(.70) =.175	$R_{bd}xR_{ce}$=(.50)(.40) =.20

Table 17. ANOVA Table for the Data from Table 16

Source	df	SS	MS	F
Tetrads	4	.030875	.00771875	43.18
Columns	1	.00121	.00121	6.77
Error	4	.000715	.00017875	
Total	9	.0328		

$$MS_p = \frac{.00121+.000715}{1+4} = \frac{.001925}{5} = .000385$$

$$F_T = \frac{.00771875}{.000385} = 20.05$$

Table 18. Effect of Unequal Variances on the Significance Level of ANOVA[a]

K	n	Ratio of Sample Variances	Actual Significance Level
2	7	1:2	.051
		1:5	.058
		1:10	.063
3	5	1:2:3	.058
		1:1:3	.059
5	5	1:1:1:1:3	.074
7	3	1:1:1:1:1:1:7	.120

[a] Nominal α = .05

This table is copied without permission from Lindman, 1974.

Table B.7 Distribution of $F_{\max}$ Statistic*

df for s_x^2	$1-\alpha$	k number of variances								
		2	3	4	5	6	7	8	9	10
4	.95 .99	9.60 23.2	15.5 37.	20.6 49.	25.2 59.	29.5 69.	33.6 79.	37.5 89.	41.4 97.	44.6 106.
5	.95 .99	7.15 14.9	10.8 22.	13.7 28.	16.3 33.	18.7 38.	20.8 42.	22.9 46.	24.7 50.	26.5 54.
6	.95 .99	5.82 11.1	8.38 15.5	10.4 19.1	12.1 22.	13.7 25.	15.0 27.	16.3 30.	17.5 32.	18.6 34.
7	.95 .99	4.99 8.89	6.94 12.1	8.44 14.5	9.70 16.5	10.8 18.4	11.8 20	12.7 22.	13.5 23.	14.3 24.
8	.95 .99	4.43 7.50	6.00 9.9	7.18 11.7	8.12 13.2	9.03 14.5	9.78 15.8	10.5 16.9	11.1 17.9	11.7 18.9
9	.95 .99	4.03 6.54	5.34 8.5	6.31 9.9	7.11 11.1	7.80 12.1	8.41 13.1	8.95 13.9	9.45 14.7	9.91 15.3
10	.95 .99	3.72 5.85	4.85 7.4	5.67 8.6	6.34 9.6	6.92 10.4	7.42 11.1	7.87 11.8	8.28 12.4	8.66 12.9
12	.95 .99	3.28 4.91	4.16 6.1	4.79 6.9	5.30 7.6	5.72 8.2	6.09 8.7	6.42 9.1	6.72 9.5	7.00 9.9
15	.95 .99	2.86 4.07	3.54 4.9	4.01 5.5	4.37 6.0	4.68 6.4	4.95 6.7	5.19 7.1	5.40 7.3	5.59 7.5
20	.95 .99	2.46 3.32	2.95 3.8	3.29 4.3	3.54 4.6	3.76 4.9	3.94 5.1	4.10 5.3	4.24 5.5	4.37 5.6
30	.95 .99	2.07 2.63	2.40 3.0	2.61 3.3	2.78 3.4	2.91 3.6	3.02 3.7	3.12 3.8	3.21 3.9	3.29 4.0
60	.95 .99	1.67 1.96	1.85 2.2	1.96 2.3	2.04 2.4	2.11 2.4	2.17 2.5	2.22 2.5	2.26 2.6	2.30 2.6
∞	.95 .99	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00

* This table is abridged from Table 31 in *Biometrika Tables for Statisticians*, vol. 1. (2d ed.) New York: Cambridge, 1958. Edited by E. S. Pearson and H. O. Hartley. Reproduced with the kind permission of E. S. Pearson and the trustees of *Biometrika*.

Table B.8 Critical Values for Cochran's Test for Homogeneity of Variance*

$C = (\text{largest } s_j^2)/\Sigma s_j^2$

df for s_j^2	$1-\alpha$	number of variances										
		2	3	4	5	6	7	8	9	10	15	20
1	.95	.9985	.9669	.9065	.8412	.7808	.7271	.6798	.6385	.6020	.4709	.3894
	.99	.9999	.9933	.9676	.9279	.8828	.8376	.7945	.7544	.7175	.5747	.4799
2	.95	.9750	.8709	.7679	.6838	.6161	.5612	.5157	.4775	.4450	.3346	.2705
	.99	.9950	.9423	.8643	.7885	.7218	.6644	.6152	.5727	.5358	.4069	.3297
3	.95	.9392	.7977	.6841	.5981	.5321	.4800	.4377	.4027	.3733	.2758	.2205
	.99	.9794	.8831	.7814	.6957	.6258	.5685	.5209	.4810	.4469	.3317	.2654
4	.95	.9057	.7457	.6287	.5441	.4803	.4307	.3910	.3584	.3311	.2419	.1921
	.99	.9586	.8335	.7212	.6329	.5635	.5080	.4627	.4251	.3934	.2882	.2288
5	.95	.8772	.7071	.5895	.5065	.4447	.3974	.3595	.3286	.3029	.2195	.1735
	.99	.9373	.7933	.6761	.5875	.5195	.4659	.4226	.3870	.3572	.2593	.2048
6	.95	.8534	.6771	.5598	.4783	.4184	.3726	.3362	.3067	.2823	.2034	.1602
	.99	.9172	.7606	.6410	.5531	.4866	.4347	.3932	.3592	.3308	.2386	.1877
7	.95	.8332	.6530	.5365	.4564	.3980	.3535	.3185	.2901	.2666	.1911	.1501
	.99	.8988	.7335	.6129	.5259	.4608	.4105	.3704	.3378	.3106	.2228	.1748
8	.95	.8159	.6333	.5175	.4387	.3817	.3384	.3043	.2768	.2541	.1815	.1422
	.99	.8823	.7107	.5897	.5037	.4401	.3911	.3522	.3207	.2945	.2104	.1646
9	.95	.8010	.6167	.5017	.4241	.3682	.3259	.2926	.2659	.2439	.1736	.1357
	.99	.8674	.6912	.5702	.4854	.4229	.3751	.3373	.3067	.2813	.2002	.1567
16	.95	.7341	.5466	.4366	.3645	.3135	.2756	.2462	.2226	.2032	.1429	.1108
	.99	.7949	.6059	.4884	.4094	.3529	.3105	.2779	.2514	.2297	.1612	.1248
36	.95	.6602	.4748	.3720	.3066	.2612	.2278	.2022	.1820	.1655	.1144	.0879
	.99	.7067	.5153	.4057	.3351	.2858	.2494	.2214	.1992	.1811	.1251	.0960
144	.95	.5813	.4031	.3093	.2513	.2119	.1833	.1616	.1446	.1308	.0889	.0675
	.99	.6062	.4230	.3251	.2644	.2229	.1929	.1700	.1521	.1376	.0934	.0709

* Reproduced with permission from C. Eisenhart, M. W. Hastay, and W. A. Wallis, *Techniques of Statistical Analysis*, chap. 15. New York: McGraw-Hill, 1947

Table 19. Original Data for Box-Scheffé Test

	SAMPLES			
	1	2	3	4
	1	2	5	7
	2	3	5	8
	2	3	6	8
	2	3	6	9
	3	4	6	9
	3	4	7	10
	3	4	7	10
	6	5	8	11
$\bar{X}_i$	2.75	3.50	6.25	9.00
s_i^2	1.488	.926	1.035	1.309

Table 20. Data of Table 19 Divided into Subsamples

		SAMPLES			
		1	2	3	4
SUBSAMPLES	1	3	4	6	8
		3	5	5	9
		2	4	6	8
	2	2	3	8	11
		2	3	7	7
		1	2	7	10
	3	6	4	5	10
		3	3	6	9

Table 21. Variances of Subsample Data in Table 20

		SAMPLES			
		1	2	3	4
	1	.3333	.3333	.3333	.3333
SUBSAMPLES	2	.3333	.3333	.3333	4.3333
	3	4.5000	.5000	.5000	.5000

Table 22. Natural Logarithms of Variances in Table 21

		SAMPLES			
		1	2	3	4
	1	-1.0986	-1.0986	-1.0986	-1.0986
SUBSAMPLES	2	-1.0986	-1.0986	-1.0986	1.4663
	3	1.5041	-.6931	-.6931	-.6931

Table 23. Computations Needed to Obtain γ_2^* for the Welch-Aspin Test

SAMPLE	n_i	n_i-1	S_i^2	W_i	$W_i/(n_i-1)$	$W_i^2/(n_i-1)$	$1/(n_i-1)$
1	8	7	8.214	.9739	.1391	.1355	.1429
2	8	7	8.839	.9051	.1293	.1170	.1429
3	8	7	9.696	.8251	.1179	.0973	.1429
4	8	7	2.796	2.8612	.4087	1.1695	.1429
TOTALS				5.5653	.7950	1.5193	.5716

Table 24. Original Data for the Kruskal-Wallis Test (Example 1)

SAMPLE		
1	2	3
12	28	14
16	32	11
10	23	17
13	35	24

Table 25. Ranked Data for the Kruskal-Wallis Test (Example 1)

	SAMPLE		
	1	2	3
	3	10	5
	6	11	2
	1	8	7
	4	12	9
TOTALS(R_i)	14	41	23

Table 26. Original Data for the Kruskal-Wallis Test (Example 2)

SAMPLE			
1	2	3	4
6	7	4	7
8	10	6	4
6	12	12	10
10	13	8	8

Table 27. Ranked Data for the Kruskal-Wallis Test (Example 2)

	SAMPLE			
	1	2	3	4
	4	6.5	1.5	6.5
	9	12	4	1.5
	4	14.5	14.5	12
	12	16	9	9
TOTALS(R_i)	29	49	29	29

Table 28. Original Data for the Friedman Test

		STIMULI		
		1	2	3
	1	4	6	7
	2	5	4	8
OBSERVERS	3	3	7	8
	4	5	6	8
	5	3	7	9
	6	6	5	8

Table 29. Ranked Data for the Friedman Test

		STIMULI		
		1	2	3
	1	1	2	3
	2	2	1	3
OBSERVERS	3	1	2	3
	4	1	2	3
	5	1	2	3
	6	2	1	3
TOTALS(R_i)		8	10	18

Table 30. Data and Computations for the Matched-Pair Wilcoxon Test

Observer	Stimulus 1 Scale Values	Stimulus 2 Scale Values	Difference	Absolute Difference	Ranks of Absolute Difference	Sign of Initial Difference
1	4	8	-4	4	5.5	-
2	6	9	-3	3	3.5	-
3	3	12	-9	9	8	-
4	8	7	1	1	1	+
5	10	6	4	4	5.5	+
6	4	9	-5	5	7	-
7	5	3	2	2	2	+
8	7	10	-3	3	3.5	-

$T_+ = 1 + 5.5 + 2 = 8.5$

$T_- = 5.5 + 3.5 + 8 + 7 + 3.5 = 27.5$

TABLE A–21. Distribution of the signed-rank statistic T

The percentiles listed cover the range $\alpha = .005$ to $.125$ for every sample size up to $n = 20$. Values $T_{(+)}$ are such that the probability is α that the signed rank statistic is less than or equal to $T_{(+)}$. The values $T_{(-)}$ are such that the probability is α that T is greater than or equal to $T_{(-)}$.

$T_{(+)}$	$T_{(-)}$	α
	$n = 1$	
0	1	.500
	$n = 2$	
0	3	250
	$n = 3$	
0	6	125
	$n = 4$	
0	10	.062
1	9	.125
	$n = 5$	
0	15	.031
1	14	.062
2	13	094
3	12	156
	$n = 6$	
0	21	016
1	20	031
2	19	.047
3	18	.078
4	17	.109
5	16	.156
	$n = 7$	
0	28	008
1	27	.016
2	26	.023
3	25	039
4	24	055
5	23	078
6	22	.109
7	21	148
	$n = 8$	
0	36	.004
1	35	008
2	34	.012
3	33	.020
4	32	.027
5	31	039
6	30	.055
7	29	074
8	28	098
9	27	125
	$n = 9$	
1	44	.004
2	43	006
3	42	.010
	$n = 9$ (*Cont.*)	
4	41	.014
5	40	020
6	39	027
7	38	037
8	37	049
9	36	064
10	35	082
11	34	102
12	33	125
	$n = 10$	
3	52	.005
4	51	.007
5	50	010
6	49	.014
7	48	019
8	47	.024
9	46	.032
10	45	.042
11	44	053
12	43	065
13	42	080
14	41	097
15	40	116
16	39	138
	$n = 11$	
5	61	005
6	60	007
7	59	009
8	58	012
9	57	016
10	56	021
11	55	027
12	54	.034
13	53	042
14	52	.051
15	51	062
16	50	074
17	49	087
18	48	103
19	47	120
20	46	139
	$n = 12$	
7	71	.005
8	70	.006
	$n = 12$ (*Cont.*)	
9	69	.008
10	68	010
11	67	.013
12	66	017
13	65	021
14	64	026
15	63	.032
16	62	039
17	61	.046
18	60	.055
19	59	065
20	58	076
21	57	088
22	56	.102
23	55	.117
24	54	133
	$n = 13$	
9	82	.004
10	81	005
11	80	.007
12	79	.009
13	78	.011
14	77	013
15	76	.016
16	75	.020
17	74	024
18	73	029
19	72	.034
20	71	040
21	70	.047
22	69	.055
23	68	.064
24	67	.073
25	66	084
26	65	095
27	64	108
28	63	.122
29	62	.137
	$n = 14$	
12	93	004
13	92	005
14	91	007
15	90	.008
16	89	.010
	$n = 14$ (*Cont.*)	
17	88	.012
18	87	015
19	86	.018
20	85	.021
21	81	025
22	83	.029
23	82	.034
24	81	039
25	80	.045
26	79	.052
27	78	.059
28	77	.068
29	76.	.077
30	75	.086
31	74	.097
32	73	.108
33	72	.121
34	71	.134
	$n = 15$	
15	105	.004
16	104	.005
17	103	.006
18	102	008
19	101	.009
20	100	.011
21	99	.013
22	98	.015
23	97	.018
24	96	.021
25	95	.024
26	94	.028
27	93	.032
28	92	.036
29	91	042
30	90	.047
31	89	053
32	88	.060
33	87	.068
34	86	.076
35	85	.084
36	84	094
37	83	.104
38	82	.115
39	81	.126

TABLE A–21 (continued)

$T_{(+)}$	$T_{(-)}$	α	$T_{(+)}$	$T_{(-)}$	α	$T_{(+)}$	$T_{(-)}$	α	$T_{(+)}$	$T_{(-)}$	α
	n = 16			*n* = 17 (*Cont.*)			*n* = 18 (*Cont.*)			*n* = 19 (*Cont.*)	
19	117	005	36	117	028	51	120	071	64	126	113
20	116	005	37	116	032	52	119	077	65	125	.121
21	115	007	38	115	036	53	118	084	66	124	129
22	114	008	39	114	040	54	117	091		*n* = 20	
23	113	009	40	113	044	55	116	098	37	173	.005
24	112	011	41	112	049	56	115	106	38	172	005
25	111	012	42	111	054	57	114	114	39	171	.006
26	110	014	43	110	060	58	113	123	40	170	.007
27	109	017	44	109	066	59	112	132	41	169	.008
28	108	019	45	108	073		*n* = 19		42	168	.009
29	107	022	46	107	080	32	158	005	43	167	.010
30	106	025	47	106	087	33	157	005	44	166	011
31	105	029	48	105	095	34	156	006	45	165	012
32	104	033	49	104	103	35	155	007	46	164	013
33	103	037	50	103	112	36	154	008	47	163	015
34	102	042	51	102	122	37	153	009	48	162	016
35	101	047	52	101	132	38	152	010	49	161	.018
36	100	052		*n* = 18		39	151	011	50	160	020
37	99	058	27	144	004	40	150	013	51	159	022
38	98	065	28	143	005	41	149	014	52	158	024
39	97	072	29	142	006	42	148	016	53	157	.027
40	96	080	30	141	007	43	147	018	54	156	029
41	95	088	31	140	008	44	146	020	55	155	032
42	94	096	32	139	009	45	145	022	56	154	.035
43	93	106	33	138	010	46	144	025	57	153	038
44	92	116	34	137	012	47	143	027	58	152	.041
45	91	126	35	136	013	48	142	030	59	151	.045
46	90	137	36	135	015	49	141	.033	60	150	049
	n = 17		37	134	.017	50	140	.036	61	149	.053
23	130	005	38	133	019	51	139	040	62	148	057
24	129	005	39	132	022	52	138	044	63	147	.062
25	128	006	40	131	024	53	137	048	64	146	.066
26	127	007	41	130	027	54	136	052	65	145	.071
27	126	.009	42	129	030	55	135	057	66	144	.077
28	125	010	43	128	033	56	134	062	67	143	.082
29	124	012	44	127	037	57	133	067	68	142	.088
30	123	013	45	126	041	58	132	072	69	141	095
31	122	015	46	125	.045	59	131	.078	70	140	.101
32	121	017	47	124	049	60	130	.084	71	139	.108
33	120	020	48	123	054	61	129	.091	72	138	.115
34	119	022	49	122	059	62	128	098	73	137	.123
35	118	025	50	121	065	63	127	105	74	136	.131

From *Introduction to Statistical Analysis* (3rd ed.), by W. J. Dixon and F. J. Massey, Jr. Copyright © 1969 by McGraw-Hill, Inc. Used with permission of McGraw-Hill Book Company.

Table 31. Sample Dissimilarity Matrix.

		Stimuli						
		X_1	X_2	X_3	X_4	X_5	X_6	X_7
Stimuli	X_1	-						
	X_2	8.7	-					
	X_3	3.4	6.3	-				
	X_4	0.6	1.1	2.1	-			
	X_5	9.5	3.5	8.8	5.3	-		
	X_6	5.0	4.3	7.0	9.1	1.6	-	
	X_7	4.9	2.2	9.5	0.4	0.4	3.2	-

LIST OF FIGURES

Figure 1. Discriminal Dispersion of Stimulus $\underline{X}_i$ on Attribute $\underline{S}$.

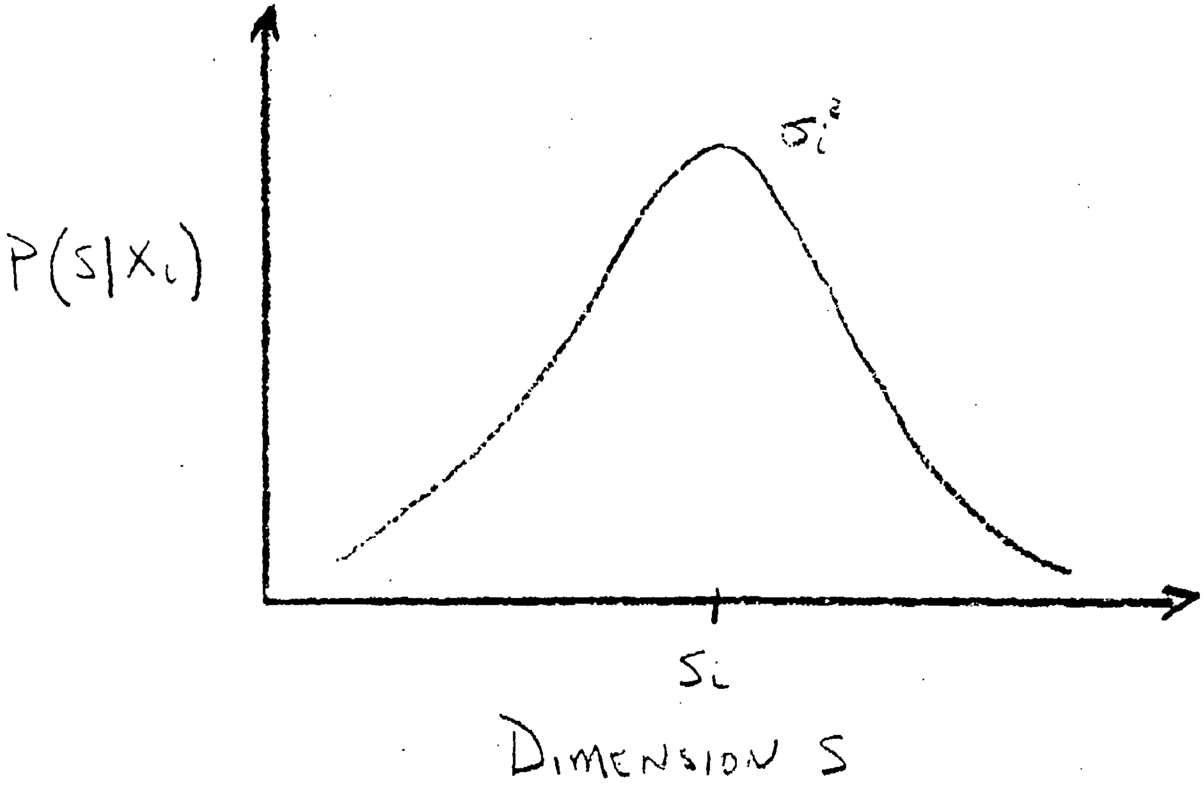


Figure 2. Thurstonian Characterization of the Categorical Judgment Situation.

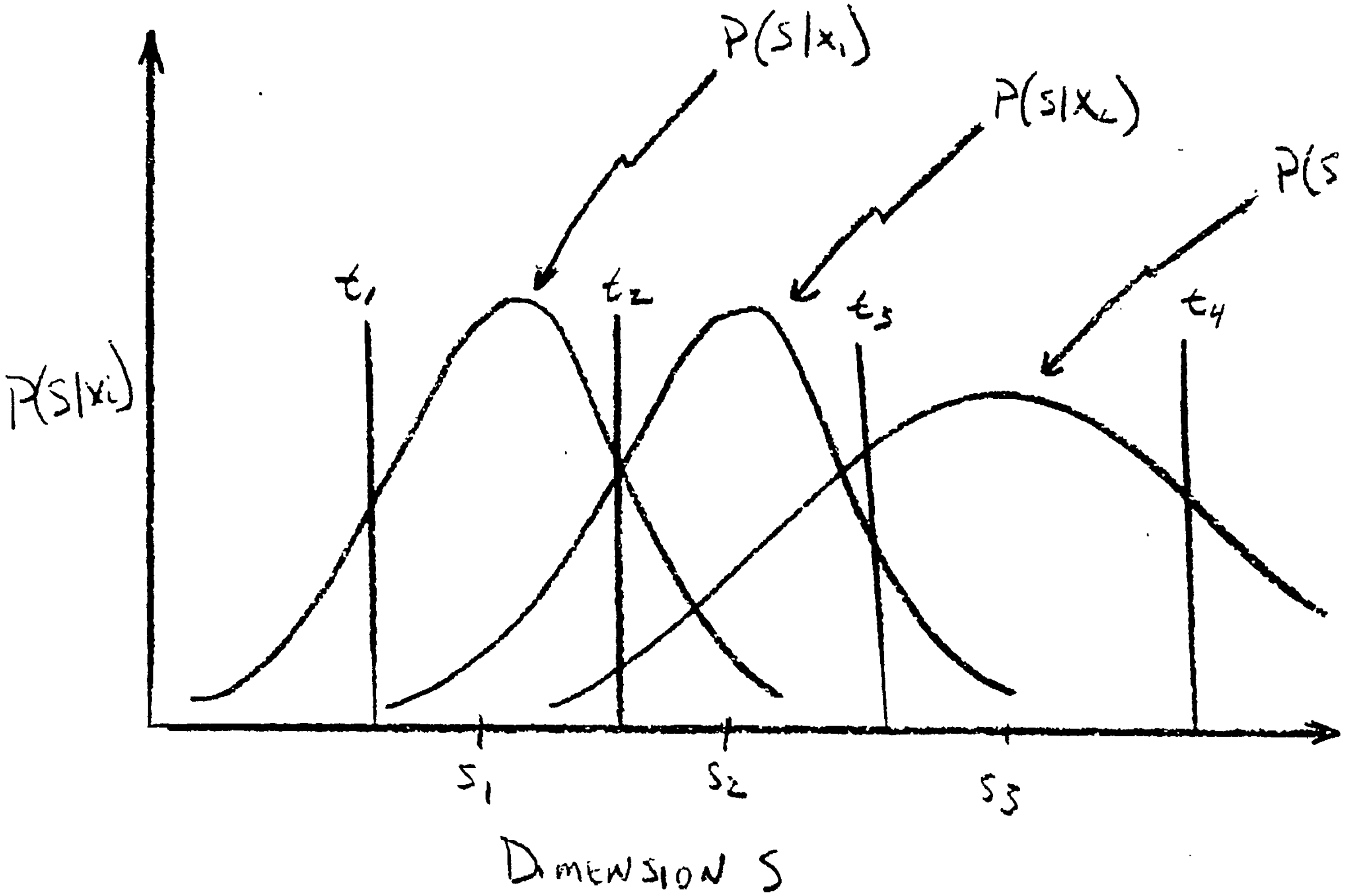


Figure 3. Plots of Z-scores for $\underline{X}_i$ against those for $\underline{X}_1$ in Thurstonian Analysis.

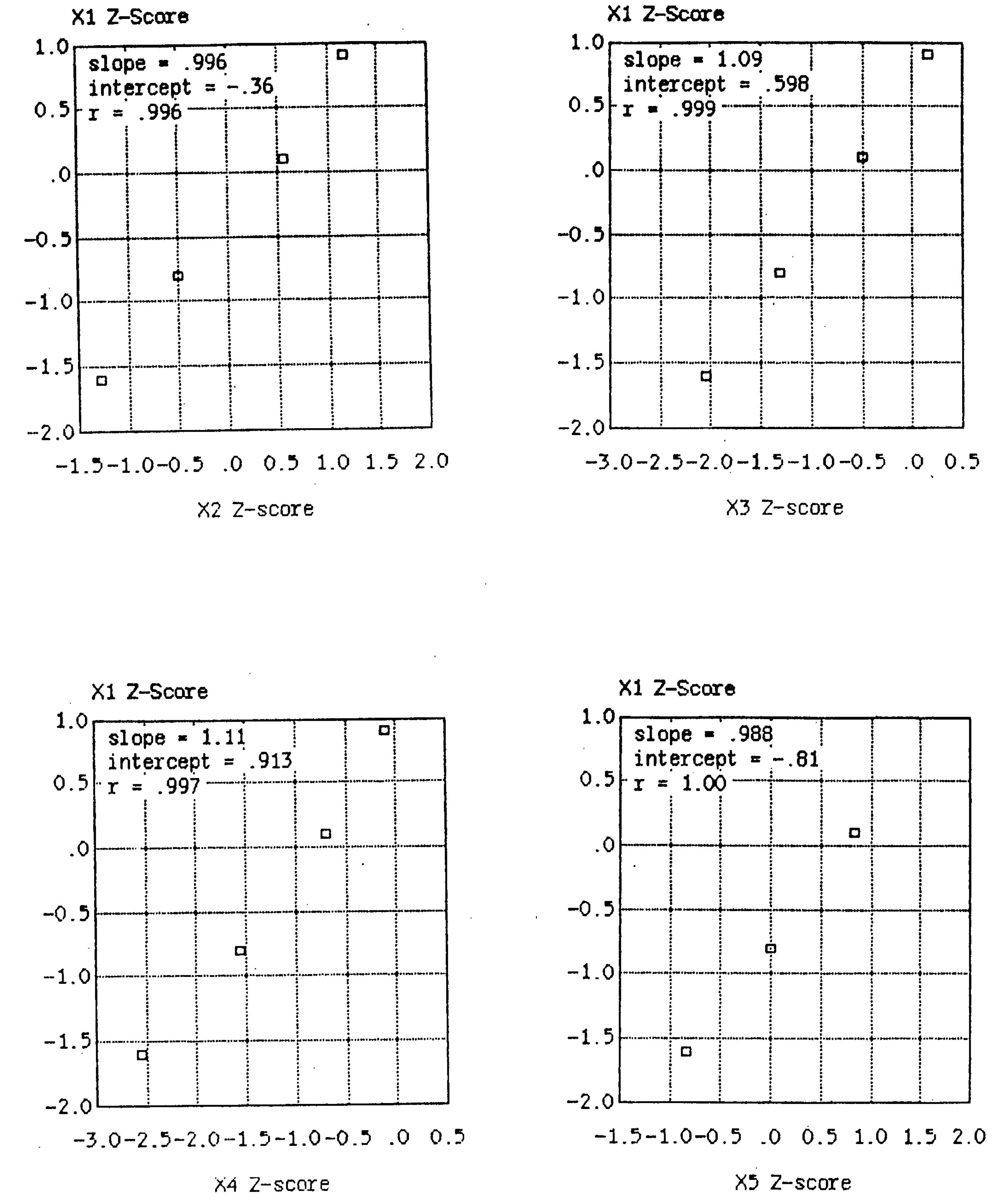
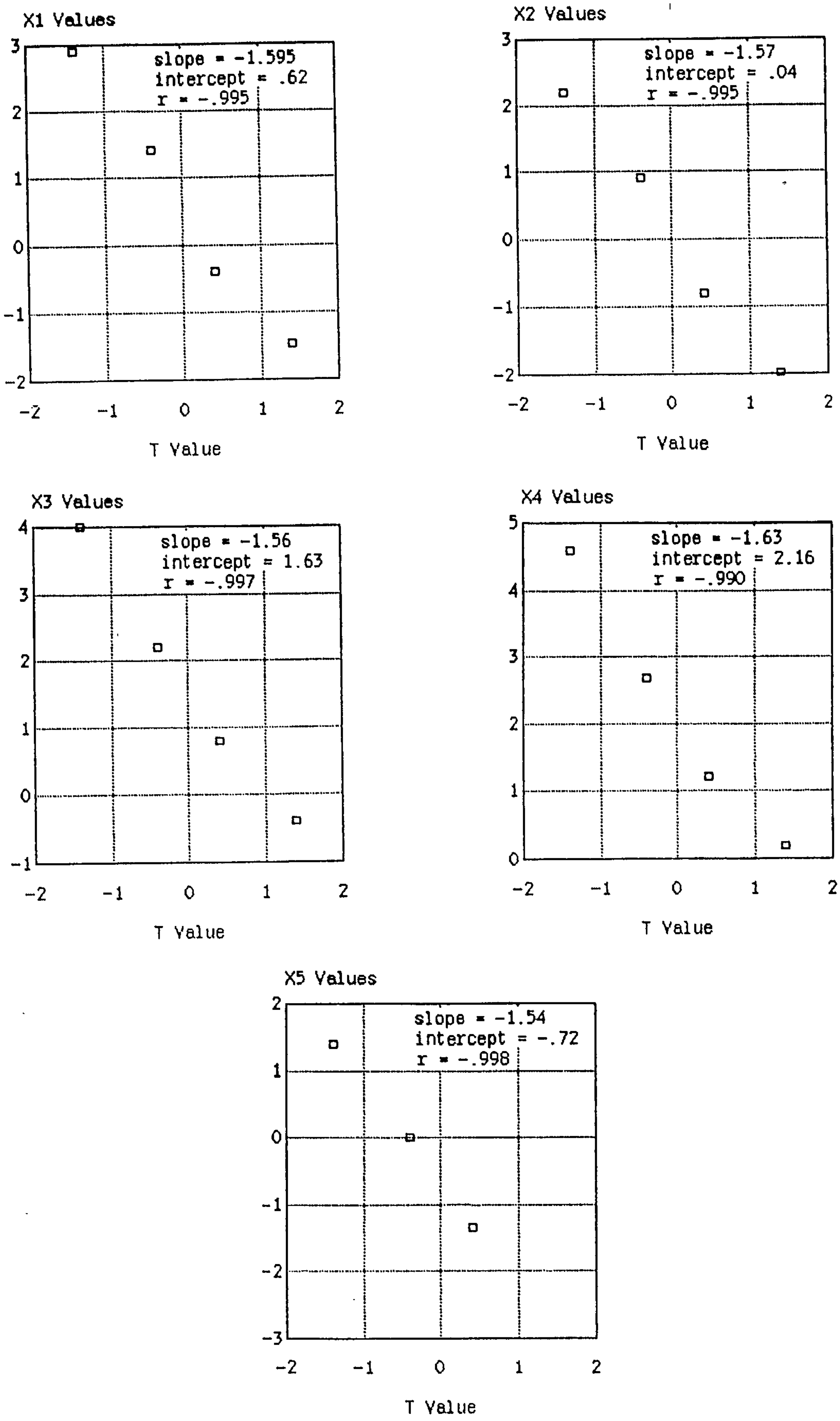


Figure 4. Plots of ln ([1/F(T)]-1) against transformed Criterion Positions for all $\underline{X}_i$ in Allnatt's Analysis.



X1 Values

X2 Values

X3 Values

X4 Values

X5 Values

Figure 5. Plot of ln ($\underline{J}_m$) against Impairment Measure (in dB).

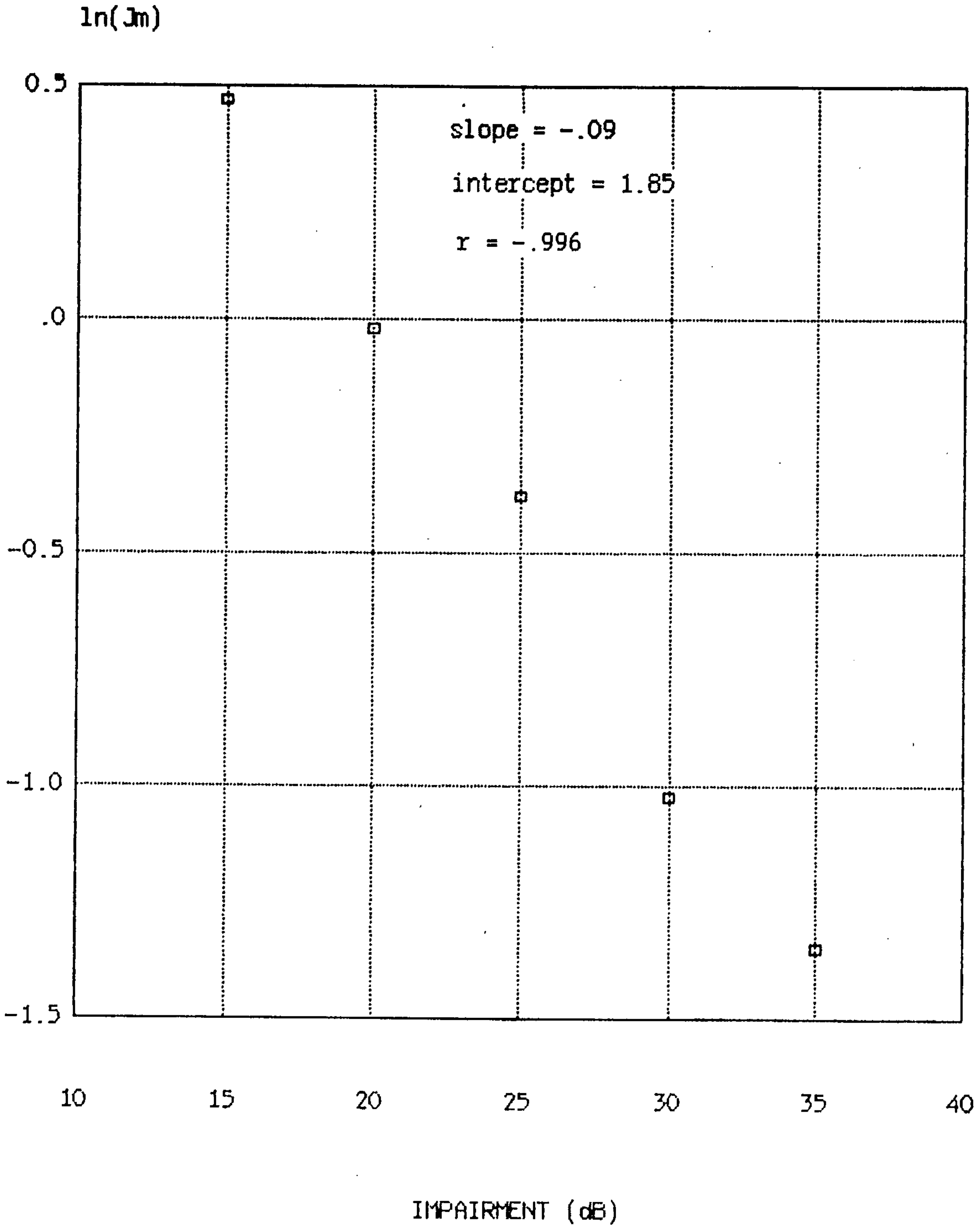


Figure 6. Plot of ln ($\underline{S}_i$) against Impairment Measure (in dB).

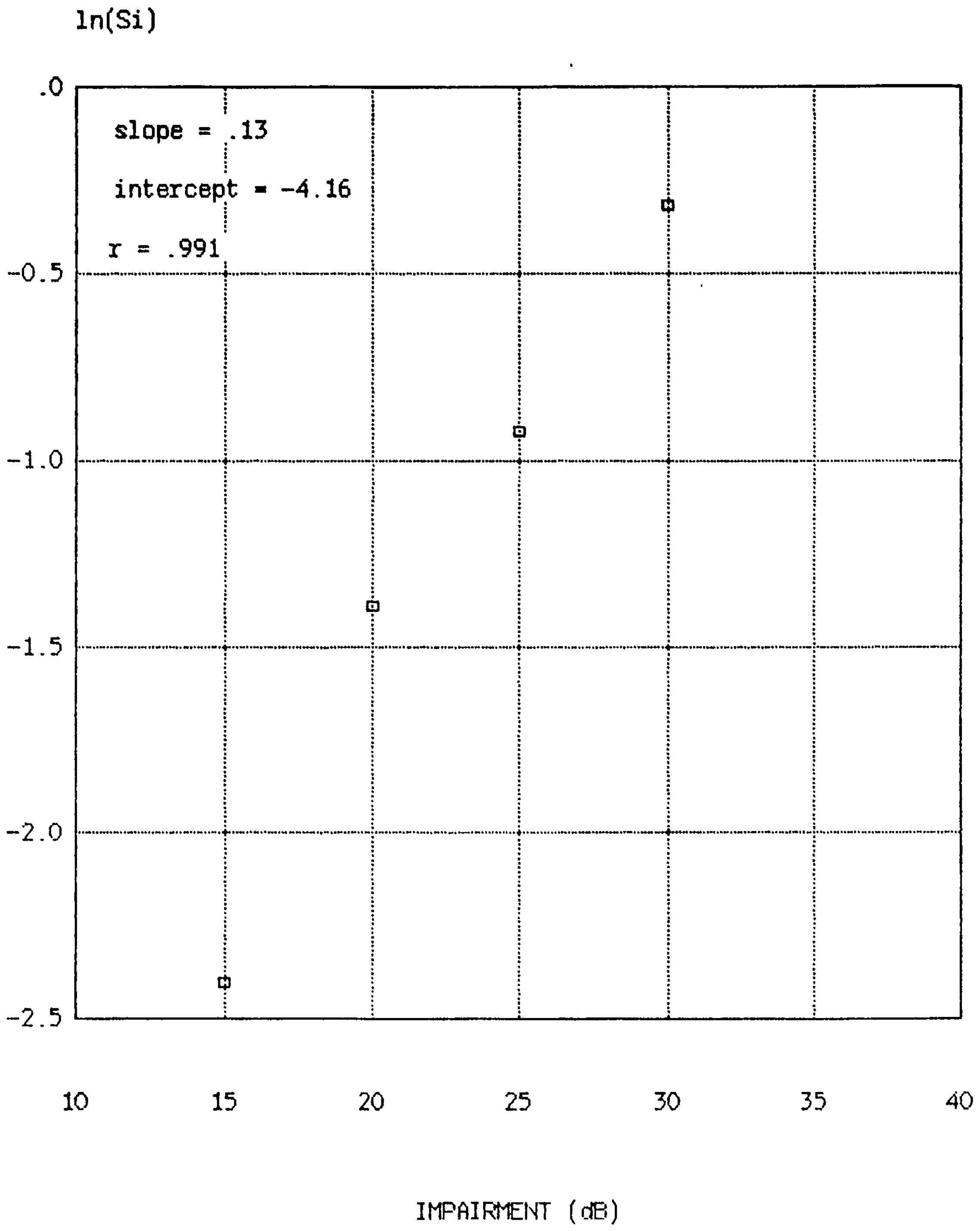
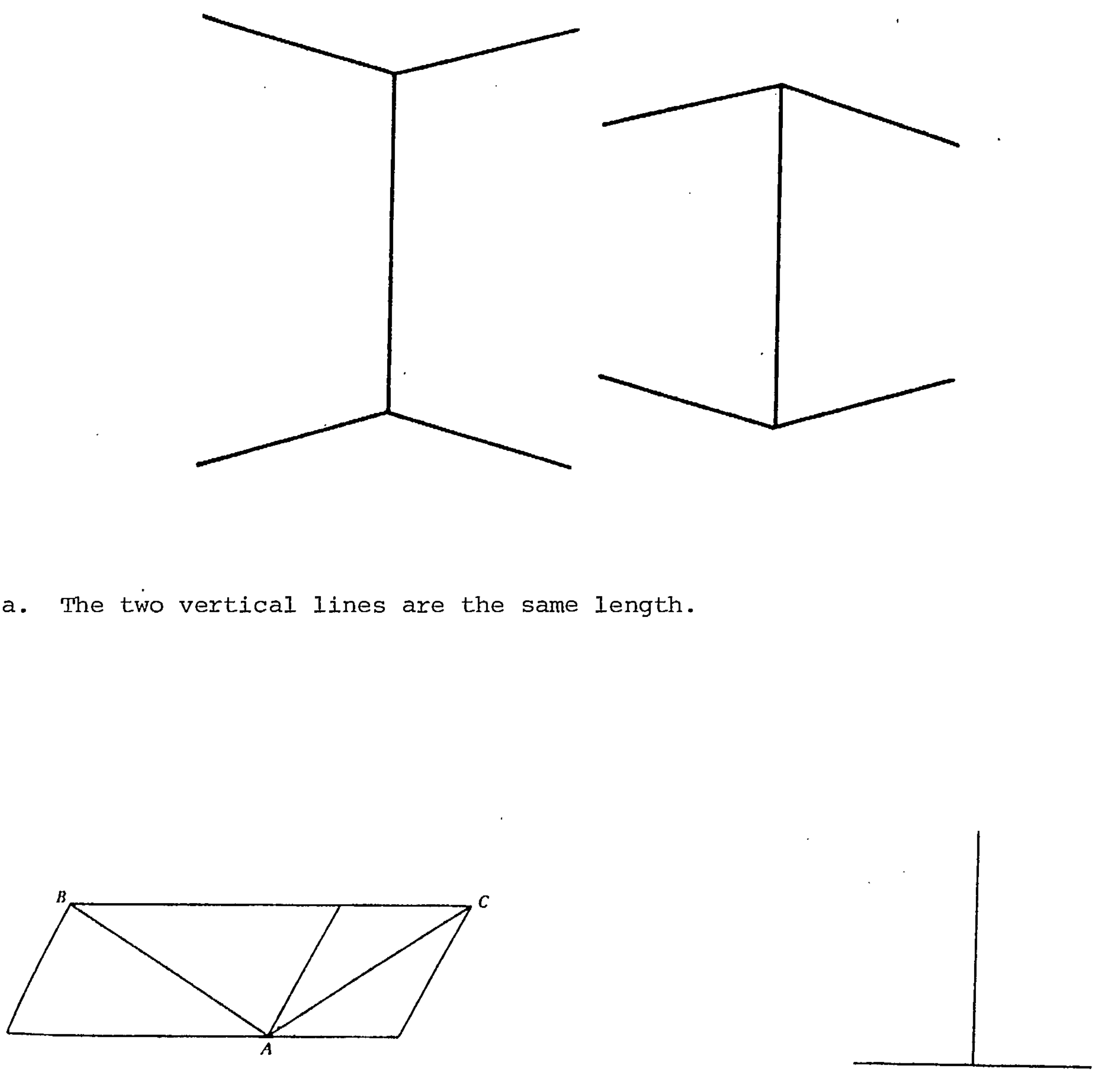


Figure 7. Some Classic Illusions.



7a. The two vertical lines are the same length.

7b. The distance from A to B is the same as the distance from A to C.

7c. The vertical line and the horizontal line are the same length.

Figure 8. Plot of STRESS as a Function of the Dimensionality of the Solution.

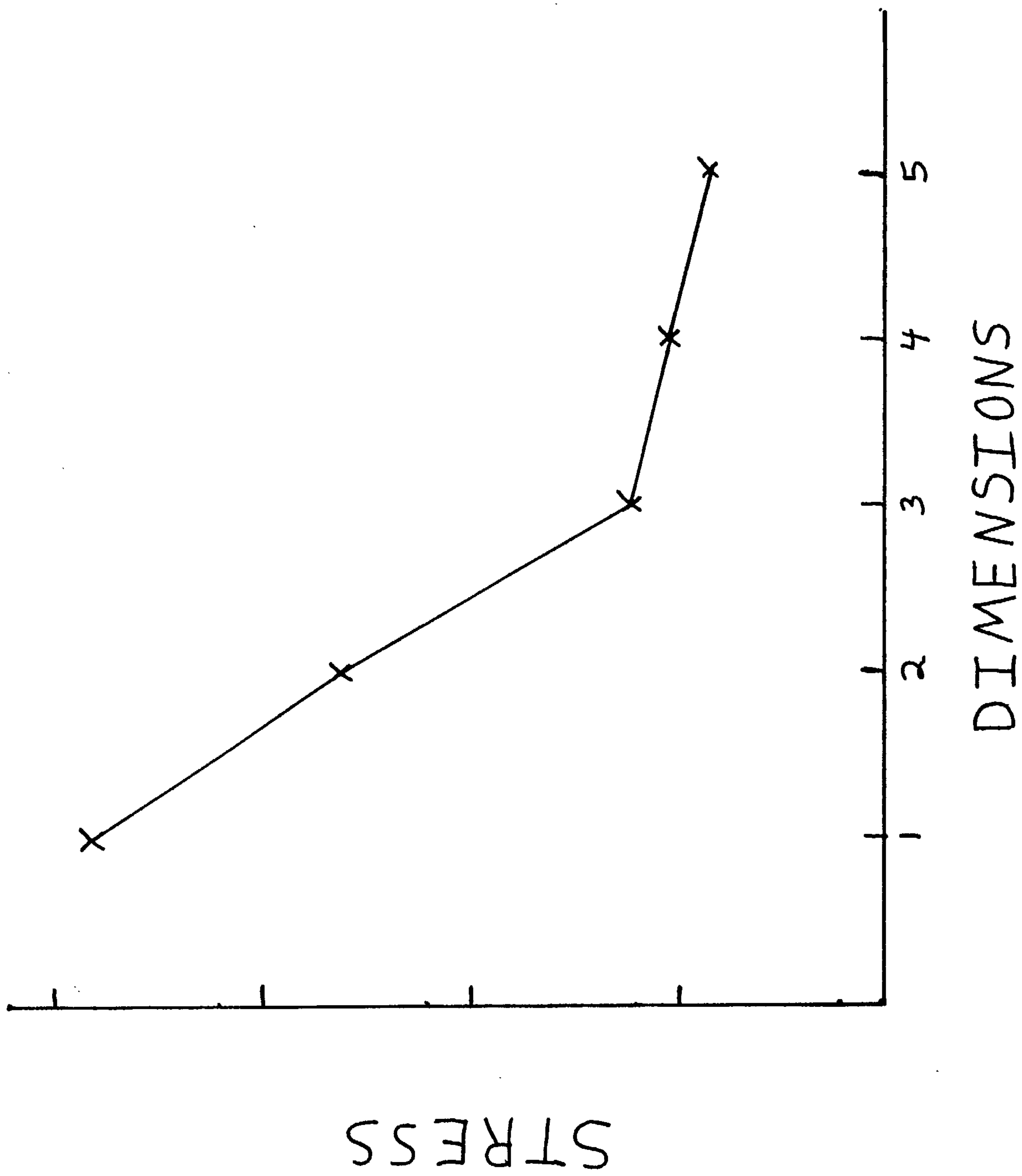


FIGURE 9 GENERAL SCHEMATIC OF THE MODEL

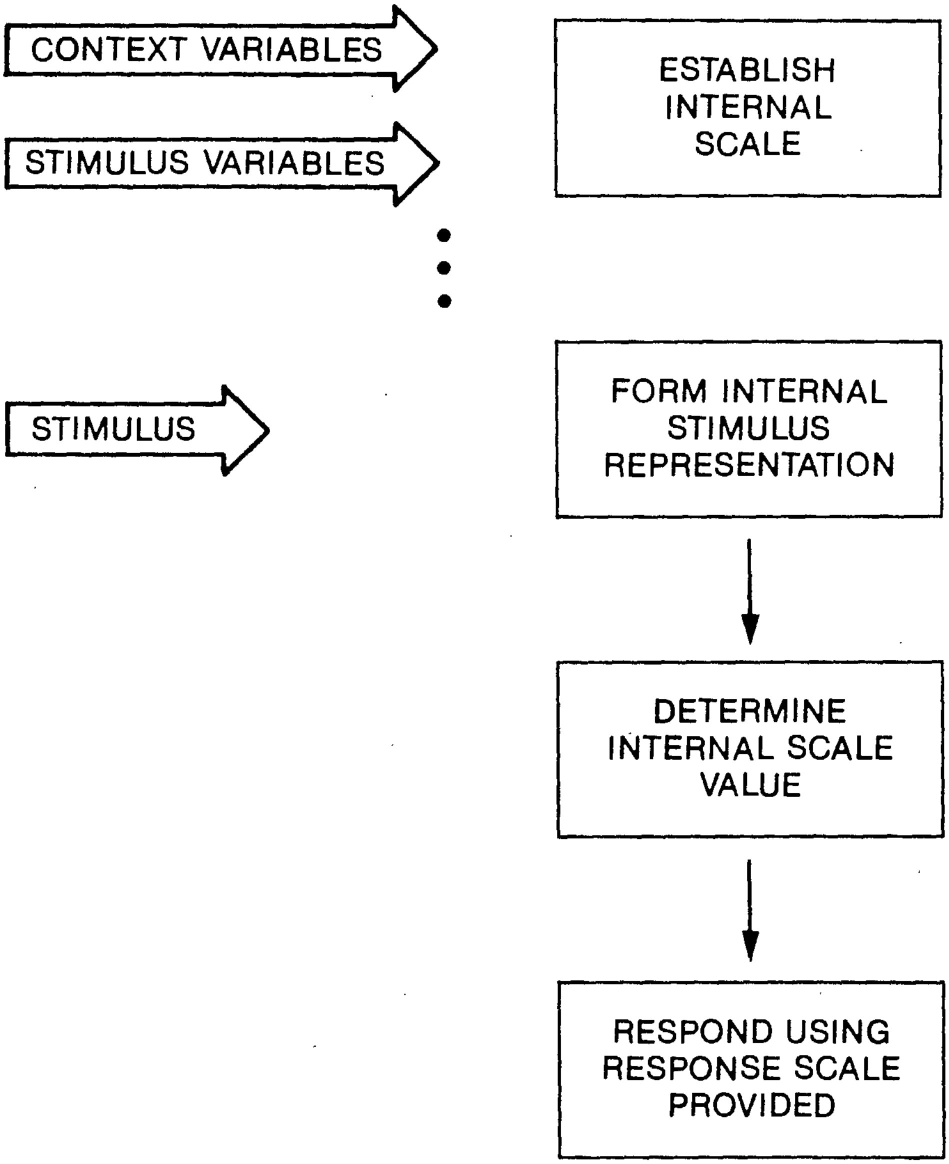
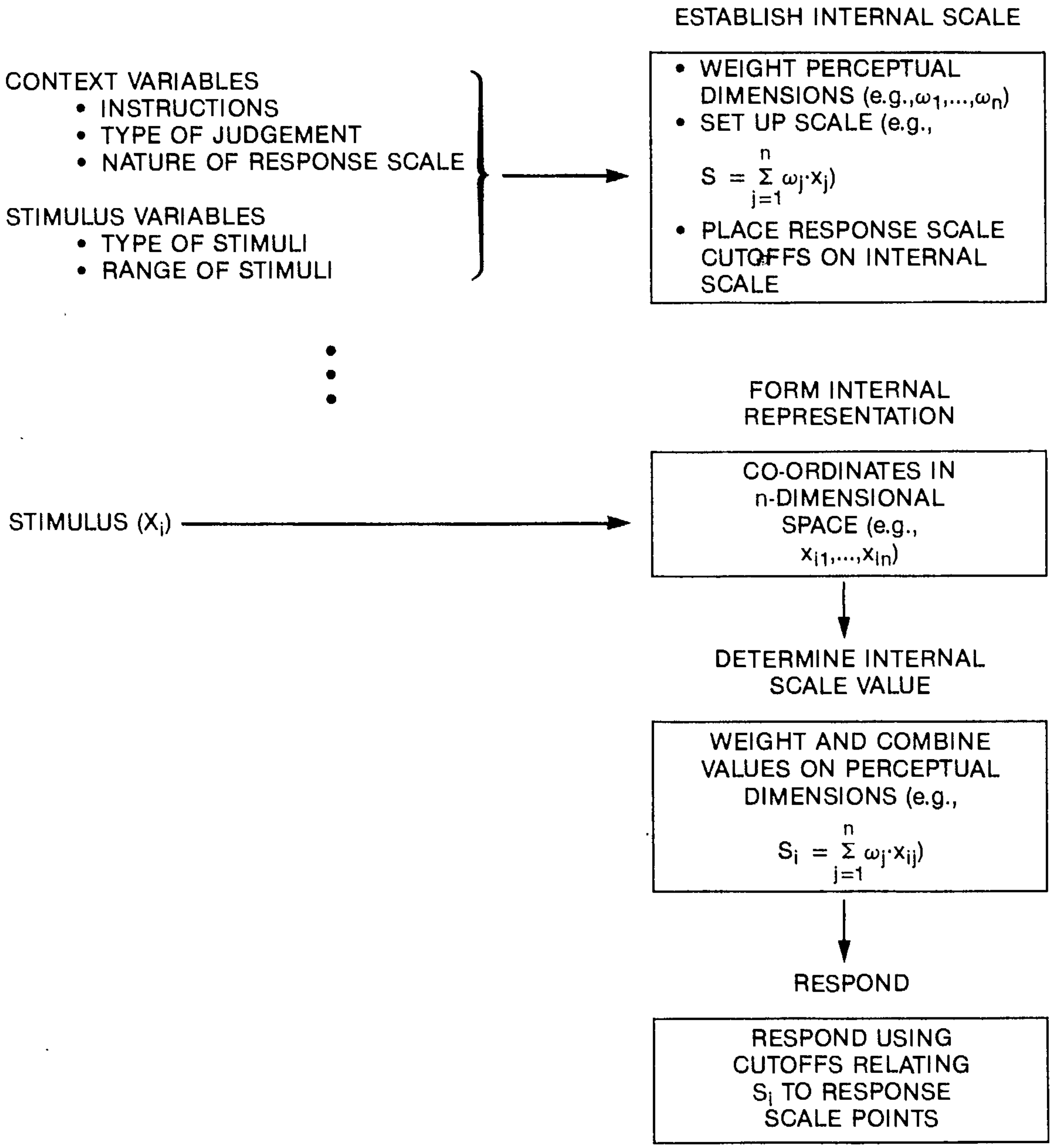

CONTEXT VARIABLES
STIMULUS VARIABLES
STIMULUS
ESTABLISH INTERNAL SCALE
FORM INTERNAL STIMULUS REPRESENTATION
DETERMINE INTERNAL SCALE VALUE
RESPOND USING RESPONSE SCALE PROVIDED

FIGURE 10 MORE DETAILED REPRESENTATION OF THE MODEL


ESTABLISH INTERNAL SCALE
• WEIGHT PERCEPTUAL DIMENSIONS (e.g.,$\omega_1,\ldots,\omega_n$)
• SET UP SCALE (e.g., $S = \sum_{j=1}^{n} \omega_j \cdot x_j$)
• PLACE RESPONSE SCALE CUTOFFS ON INTERNAL SCALE
CONTEXT VARIABLES
• INSTRUCTIONS
• TYPE OF JUDGEMENT
• NATURE OF RESPONSE SCALE
STIMULUS VARIABLES
• TYPE OF STIMULI
• RANGE OF STIMULI
FORM INTERNAL REPRESENTATION
CO-ORDINATES IN n-DIMENSIONAL SPACE (e.g., $x_{i1},\ldots,x_{in}$)
STIMULUS (X_i)
DETERMINE INTERNAL SCALE VALUE
WEIGHT AND COMBINE VALUES ON PERCEPTUAL DIMENSIONS (e.g., $S_i = \sum_{j=1}^{n} \omega_j \cdot x_{ij}$)
RESPOND
RESPOND USING CUTOFFS RELATING S_i TO RESPONSE SCALE POINTS