AD-A016 282

ASSESSING THE REALIBILITY AND VALIDITY OF MULTI-ATTRIBUTE UTILITY PROCEDURES: AN APPLICATION OF THE THEORY OF GENERALIZABILITY

J. Robert Newman

University of Southern California

Prepared for:

Office of Naval Research
Advanced Research Projects Agency

1 July 1975

001597-2-T

USC

UNIVERSITY OF SOUTHERN CALIFORNIA

social science research institute

TECHNICAL REPORT

ASSESSING THE RELIABILITY AND VALIDITY OF MULTI-ATTRIBUTE UTILITY PROCEDURES: AN APPLICATION OF THE THEORY OF GENERALIZABILITY

J. ROBERT NEWMAN

SPONSORED BY:
ADVANCED RESEARCH PROJECTS AGENCY
DEPARTMENT OF DEFENSE

MONITORED BY:
ENGINEERING PSYCHOLOGY PROGRAMS
OFFICE OF NAVAL RESEARCH
CONTRACT NO. N00014-75-C-0487, ARPA
ORDER # 2105

APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED
REPRODUCTION IN WHOLE OR IN PART IS PERMITTED
FOR ANY USE OF THE U.S. GOVERNMENT

JULY 1975

SSRI RESEARCH REPORT 75-7

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the Advanced Research Projects Agency of the U.S. Government.

44

Social Science Research Institute
University of Southern California
Los Angeles, California 90007
213-746-6955

The Social Science Research Institute of the University of Southern California was founded on July 1, 1972 to permit USC scientists to bring their scientific and technological skills to bear on social and public policy problems. Its staff members include faculty and graduate students from many of the Departments and Schools of the University.

SSRI's research activities, supported in part from University funds and in part by various sponsors range from extremely basic to relatively applied. Most SSRI projects mix both kinds of goals — that is, they contribute to fundamental knowledge in the field of a social problem, and in doing so, help to cope with that problem. Typically, SSRI programs are interdisciplinary, drawing not only on its own staff but on the talents of others within the USC community. Each continuing program is composed of several projects; these change from time to time depending on staff and sponsor interest.

At present (Spring, 1975), SSRI has four programs:

Criminal justice and juvenile delinquency. Typical projects include studies of the effect of diversion on recidivism among Los Angeles area juvenile delinquents, and evaluation of the effects of decriminalization of status offenders.

Decision analysis and social program evaluation. Typical projects include study of elicitation methods for continuous probability distributions and development of an evaluation technology for California Coastal Commission decision-making.

Program for data research. A typical project is examination of small-area crime statistics for planning and evaluation of innovations in California crime prevention programs.

Models for social phenomena. Typical projects include differential-equation models of international relations transactions and models of population flows.

SSRI anticipates continuing these four programs and adding new staff and new programs from time to time. For further information, publications, etc., write or phone the Director, Professor Ward Edwards, at the address given above.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
001597-2-T

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Assessing the Reliability and Validity of Multi-Attribute Utility Procedures: An Application of the Theory of Generalizability

5. TYPE OF REPORT & PERIOD COVERED
Technical

6. PERFORMING ORG. REPORT NUMBER
None

7. AUTHOR(s)
J. Robert Newman

8. CONTRACT OR GRANT NUMBER(s)
N00014-75-C-0487

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Social Science Research Institute
University of Southern California
Los Angeles, California 90007

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
ARPA Order No. 2105

11. CONTROLLING OFFICE NAME AND ADDRESS
Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, Virginia 22209

12. REPORT DATE
1 July 1975

13. NUMBER OF PAGES
44

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)
Engineering Psychology Programs
Office of Naval Research
Arlington, Virginia 22217

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release; Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Decision Analysis
Multi-Attribute Utility
Reliability
Validity
Generalizability

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report presents a theoretical rationale for assessing the reliability, validity, and dependability of multi-attribute utility models and techniques. If an investigator is advocating the use of a MAU model or procedure he or she is interested in generalizing from observations at hand to a universe or domain of observations that are members of that same universe. The universe must be unambiguously defined but it is not necessary to assume that universe as having any statistical properties such as uniform variances or covariances. A study of generalizability is conducted by taking measurements on persons, stimuli, tasks, etc. that are assumed to be randomly representative of a universe an investigator wishes to generalize to. The ratio of an estimate of the universe "score" variance to an estimate of the observed score variance is the coefficient of generalizability. This is estimated by the intra-class correlation coefficient. ANOVA and the Expected Mean Square paradigm of Cornfield and Tukey is used to obtain the appropriate variance estimates.

The theory dispenses with unnecessary and unwarranted assumptions, and eliminates the distinction between reliability and validity. Any generalizability study can be conducted without reference to having a parallel measure of the MAU instrument or some external criterion of "success". If a MAU technique is compared to some non-MAU technique for doing the same thing then it is possible to calculate the coefficient of generalizability for both methods thus allowing the investigator to decide which is best for his or her purposes. Three numerical examples are given of the theory. Preliminary investigations have indicated that MAU models and techniques based on such models may be "better" than non-MAU models since the former have a tendency to reduce the interaction between judges and the thing being judged when such interaction represents inconsistency of judgment.

ASSESSING THE RELIABILITY AND VALIDITY
OF MULTI-ATTRIBUTE UTILITY PROCEDURES:
AN APPLICATION OF THE THEORY OF GENERALIZABILITY

Technical Report
1 July 1975

J. Robert Newman
Social Science Research Institute
University of Southern California

This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Engineering Psychology Porgrams, Office of Naval Research under Contract No. N00014-75-C-0487, ARPA, order # 2105.

Approved for Public Release
Distribution Unlimited

SSRI Research Report 75-7

Assessing the Reliability and Validity of Multi-Attribute Utility Procedures: An Application of the Theory of Generalizability

J. Robert Newman
Social Science Research Institute
University of Southern California

Most important decisions involve choosing among alternatives with multiple value characteristics. For example, in deciding what home to buy, some of the value relevant characteristics might be: number of rooms, price, location, potential as an investment, and so on. A set of multi-attribute utility (MAU) models and procedures have been proposed as an aid in making such decisions (Edwards, 1971; Raiffa, 1968, 1969).

The basic idea of a MAU procedure is to "Divide and Conquer" (Raiffa, 1968, p.271). There are three basic steps in this process. First, the decision problem is broken up into little pieces (attributes) along natural lines depending upon the nature of the task. Second, separate judgments are made about each of the component pieces. As a rule, there are two such judgments, numerical judgments about the importance of each attribute relative to each other and numerical judgments about the "worth" or utility of each attribute to each of the competing decision alternatives. Finally, these separate judgments are aggregated using some formal algebraic rule and this is used as an aid to the final decision.

Advocates of MAU procedures have offered them as a replacement for "wholistic" procedures in which the decision maker forms an overall intuitive evaluation of a decision alternative. Such advocates have also argued that MAU procedures are "better" in the sense that they are more reliable and valid than wholistic procedures. It is this comparison that forms the motivation behind this paper which stems from a disillusionment with the way multiattribute utility (MAUM) models and procedures are being assessed as to their reliability and validity. I would like to suggest a re-formulation thay may resolve a good many conceptual and practical difficulties. The reformulation is based on some work of Lee Cronbach and his Associates (Cronbach, et al., 1972) and is called the Theory of Generalizability. Before describing what this theory is all about and why I think it may be of considerable use in studies of MAU models, let me first give an example of the conceptual difficulties that the classical theory can lead to.

Consider the concept of validity. Under the classical theory validity (sometimes called predictive validity, concurrent validity or convergent validity) is defined as the relation between the measuring instrument and other criteria of "success". If, for example, you had a method for assessing the amount of anxiety in a person via a paper and pencil test and could demonstrate that such scores correlated highly with an independent physiological measure of anxiety (e.g., palmer sweating) then you could argue that the paper and pencil test did indeed have validity from a psychometric standpoint which also makes sense from a behavioral standpoint. The basic references on the classical theory of validity are Gulliksen (1950) and Lord and Novick (1968). Now consider how Mau techniques have been validated. Fisher (1972), Huber, Daneshgar, and Ford (1971), and more recently Gardiner (1974) used as the validating criterion for various MAU techniques wholistic judgment in certain decision situations. Each of these investigators demonstrated that decomposed additive utility models (MAU techniques) correlated highly with intuitive wholistic judgments in judgment tasks and used these results to argue that the additive utility models were therefore valid in the sense that they were capable of predicting a criterion, namely, the wholistic judgments. There is an obvious error in logic here. If the decomposed additive utility models are supposed to be "better" than wholistic judgments, why use wholistic judgment as the validating criteria for the MAU technique?

There are other difficulties. It can be demonstrated that wholistic judgments are not as reliable as decomposed judgments (Fisher, 1972; Gardiner, 1974). Therefore we are using a less reliable criterion to provide evidence for predictive validity for a more reliable MAU technique. Aside from an apparent error of logic here, consider what the classical theory of reliability and validity has to say about such a situation. There are two cases to consider:

Correction for an unreliable criterion

If, according to the classical model, you have a reasonably reliable measuring technique and you are using a fallible criterion to assess its validity then it is logically unfair to make it appear that the measuring technique is less valid than it really is. It is desirable, therefore, to correct predictive validity coefficients for attenuation in the criterion measurement (Gulliksen, 1950). The formula for such correction is

$$r'_{xy} = \frac{r_{xy}}{\sqrt{r_{yy}}} \qquad (1)$$

where r'_{xy} is the corrected validity coefficient, r_{xy} is the correlation between the measuring instrument x and the criterion y, and r_{yy} is the reliability of the criterion y. (Note: the expression $\sqrt{r_{yy}}$ is referred to as the index of reliability.)

Gardiner (1974, p. 111) presents test-retest correlations (index reliabilities) for wholistic judgments as having an average of .75 and validity coefficients having an average of .66. This latter was the largest average of the two calculations he did in correlating his Multi-Attribute Rating Scale (MARS) with the intuitive wholistic judgments. Thus applying the above formula using his results we have:

$$r'_{xy} = \frac{.66}{.75} = .88$$

This, of course, according to the classical theory, makes Gardiner's MARS technique have appreciably higher validity than he reported.

Shrinkage of validity coefficients

Any validity coefficient, whether it be a correlation between a predictor (measurement) and a criterion or whatever, is designed to yield the best possible prediction for the sample of data on which it was developed. If the prediction equation which the coefficient represents actually was applied to a new sample of data the predictions invariably would be worse and the resulting validity coefficient lower. This phenomenon is called shrinkage of the validity coefficient. The amount of shrinkage is an indication of how much biased upwards the original coefficient was. The classical theory has a formula for correcting the original validity coefficient for this upward bias. The formula is:

$$\hat{r}_{xy} = [1 - (1 - r_{xy}{}^2)(\frac{N - 1}{N - 2})]^{\frac{1}{2}} \qquad (2)$$

where $\hat{r}_{xy}$ is the estimated corrected coefficient; r_{xy} is the original coefficient, and N is the number of observations in the original sample. When we apply this formula to the results presented by Gardiner who used a sample of size N = 14 subjects, and whose average validity coefficient was .66 we have

$$\hat{r}_{xy} = [1 - (1 - .66^2)(\tfrac{13}{12})]^{\frac{1}{2}}$$

$$= .58$$

Thus we see the classical theory tells us to correct the coefficient upwards to provide for a fallible criterion and to correct it downward to provide for the over estimate due to sampling errors and capitalizing on chance in the validating sample. The Classical theory does not tell us which of these coefficients is "best".

One possible solution is to use the classical theory to first reduce the over estimated validity coefficient using formula (2) and then apply the correction for attenuation formula (1) to adjust for a fallible criterion. If we do this with Gardiner's data the obtained estimate is now .77.

All of these coefficients make sense, if we adopt the classical theory and use it as a guideline to assess the validity of the measuring instrument. I think you will agree, however, that things can be a little confusing. There should be a better way, and Cronbach and his associates have indicated that there is indeed a better way which they call a Theory of Generalizability.

The Theory of Generalizability

Basic Concepts

To ask the question of how reliable or dependable a measure is, is to ask how well one can generalize from the observation at hand to some universe or domain of observations to which it belongs. To ask about rater agreement in MAU studies is to ask how well we can generalize from one set of ratings to ratings from all possible raters who might have been chosen to actually do the rating in the particular study. The theory requires the investigator to specify the universe of conditions of observation over which he wishes to generalize. Conditions is a generic term referring to observers (raters), forms of stimuli, occasions, etc. In addition to generalizing to a universe of raters, for example, we may also wish to generalize to a universe of situations in which the ratings were made. Miller, Kaplan, and Edwards (1968) studied the efficacy of a Utility model performing under four military logistical situations. It may be of interest to know how well one could generalize from these four situations to all possible situations which the particular four represented. Gardiner (1974) used 15 "typical" housing development permit requests in his application of MAU techniques to Coastal Zone Management Decision Making. It is of interest to know how representative these 15 permit requests were and therefore how well one can generalize to the universe of such permit requests.

Questions concerning generalizability are substantive not just methodological. They require thinking about the class of observations and not just the measuring technique which gathered the observations at hand.

The following are requirements or assumptions of the theory:

(a) The universe is defined unambiguously. It must be clear what conditions fall within the universe.

(b) Conditions are experimentally independent. For example, a person's score or rating in one condition does not depend on the fact that he or she has or has not been previously observed under other conditions.

(c) Conditions are randomly selected from the universe of conditions.

This assumption is crucial but no assumptions are made about the content of the universe or about the statistical properties of the conditions within the universe. The restrictive and unnecessary assumptions of the classical theory such as uniform variances and co-variances of two or more samples of items, persons, etc. are eliminated.

If we wish to generalize to persons (raters) then for each person p, the universe score M_{pi} is defined as the expected value $E(X_{pi})$ of the observed score X_{pi} over all conditions in the universe. If we wish to generalize to situations then a universe situation mean is defined in a similar fashion. If we define generically, X_c as the sample observation of some condition c and M_c as its expected value in the population, then we can define the squared correlation $G^2_{X_cM_c} = \frac{\text{estimated universe score variance}}{\text{estimated observed score variance}}$ as the coefficient of generalizability which indicates how well one can generalize from the observed data to the universe score. This definition requires X_c and M_c to be random variables. We will see shortly when we discuss estimates of $G^2_{X_cM_c}$ that the intra-class correlation coefficient (Haggard, 1958) is a lower bound of $G^2_{X_cM_c}$ and can be easily estimated from analysis-of-variance (ANOVA) designs.

Note that this definition does not require an outside or independent criterion against which to assess the dependability of the measuring technique. Any study of the measuring technique will have its own generalizability. This is equivalent to what some investigators have called the "external validity" of the study (Campbell and Stanley, 1963). When a study has been completed and a relation found between some independent and dependent variable then questions of external validity refer to what populations can this relation be generalized to, and how extensive is this generalization? This as contrasted to internal validity which refers to how precise (reliable) the study was in the first place. It is possible, of course, to have highly precise experiments that have little generality. The converse is not true. Experiments with low internal validity are highly imprecise and thus cannot have much generality. Campbell and Stanley note, and correctly so, that internal validity is the sine qua non of a good research design but unless special cautions are taken, the results of a carefully designed study are not representative and hence not generalizable. The ideal design should be high in both internal and external validity. The theory of generalizability meets this problem head on by requiring the investigator to be very explicit about what universe he wishes to generalize to and thus forcing him to design "representative" experiments with the zeal advocated by Egon Brunswik (1956).

G and D Studies

The theory makes the distinction between generalizability studies (G study) and decision studies (D study). The D study provides information from which decisions about individuals, groups, and/or situations are made, while the G study is used to assess the actual measuring technique. The design of G and D studies may be one and the same but they are often different. The distinction between G and D studies is more than a mere recognition of the fact that certain studies are carried out during the development of a measuring instrument and then the instrument is utilized in other studies for practical purposes. The distinction is particularly crucial for anyone who advocates the use use of certain techniques which are claimed to be better than others. Such claims usually can be demonstrated in laboratory-like studies but these techniques may then be used to make very important and practical social judgments such as Coastal Zone Management decisions. The distinction is particularly important for MAU studies to clarify analyses of just how ratings are assessed and used. In Gardiner's study of Coastal zone decisions each subject (rater) gave a utility judgment on all attributes and an importance weight on all attributes. The intra-class correlation among raters (coefficient of generalization) if it were calculated would ignore differences in rater bias. This would be the appropriate coefficient in a subsequent D study if the raters used in that study also made utility and importance weight judgments on all the attributes. However, if the raters in a subsequent D study differed on what attributes they judged or, as might well be the case in practical situations, different persons provided the utility and importance weight judgments, then one would need to know the intra-class correlation that treats such things as rater leniency, possible differences in giving utility judgments versus importance weight judgments, and so on.

Generalizability and Construct Validity

We began this paper by being concerned about the reliability and validity of multi-attribute utility techniques. We then argued that the classical theories of reliability and validity are not satisfactory. Validity in particular is suspect primarily if, by validity, one means how well a measuring instrument correlates or predicts some external criterion. It is often the case that this criterion is itself suspect either because of doubtful relevance to what is really intended to be measured, or the criterion itself may be unreliable. Because of these difficulties, psychometricians have introduced another definition of validity called construct validity. The basic idea of construct validity is that any measuring instrument should have behavioral or psychological meaning in terms of some useful psychological construct that it purports to measure. The construct of anxiety, for example, is a useful and highly valid construct since it can be demonstrated that several different ways of measuring that construct (e.g., Manifest Anxiety scales, palmar sweating) all correlate reasonably well, and what is of even more importance, the measurement of anxiety in people enables you to make differential predictions about other behaviors for people who are located on different scale locations of the anxiety scale. High anxiety individuals, for example, perform quite differently in learning tasks than low anxiety individuals. Closer to home, the construct of utility can also be demonstrated to have high construct validity since when utilities are measured in both animals and humans one can make differential predictions based on these measurements in a wide variety of situations (Greeno, 1968, Ch. 2).

The theory of generalizability has implications for construct validity. The theory requires an investigator to conduct a G study by defining a universe of interest to him and then make observations under two or more selected conditions within that universe. The calculations yielding one or more coefficients of generalizability tell the investigator how well the observed scores represent the universe scores. The universe can be considered as a construct that he introduces since he thinks it has explanatory or predictive power. Thus the investigation of generalizability can be seen to be an investigation of construct validity. Thus, it is not necessary within this framework to make a distinction between reliability and validity. This notion has been recognized before by Tryon (1957) who introduced the idea of a domain sampling model in which a sample of items in any test could be considered a random sample of all items in the domain or universe of items. Tryon also pointed out that if the sampled items were tapping some interesting domain of behavior then reliability could also be considered as behavior domain validity and it is not necessary to distinguish between the concepts of reliability and validity.

Although they use different philosophical reasons, Miller, Kaplan and Edwards (1967) have also recognized that the distinction between reliability and validity is useless especially when one is concerned with decision making systems. They introduce the concept of intellectual coherence which they equate with construct validity for decision systems. To quote these authors:

> Validation is simply establishing the coherence of a procedure, or several procedures. Thus no sharp line separates the concept of reliability from that of validity; both concepts refer to agreements among measures, and a continuum exists from cases in which the measures essentially repeat the same procedure (reliability) to cases in which rather different procedures seem to measure the same thing (validity). (Miller, et al., 1967, p.48)

And further on, ...

> We assert that no external measure of the performance of a judgment-based decision-making system is possible. Any such measure would have to compare the decisions the system made with decisions made some other way, and there would have to be some good reason to suppose that the decisions made the other way were right ones. But if we reject the idea that the business of a decision-making system is to imitate some individual's decisions (in which case the only point of building the system would be to save the individual the trouble of making those decisions himself), then no basis remains for asserting that the decisions made by one procedure (e.g., by the commander) are inherently appropriate simply because they were made by that procedure, regardless of their content. An examination of the merit of decisions in terms of their content is a matter of intellectual coherence or reliability, not validity.
>
> We assert also that intellectual coherence or reliability is very measurable and is in fact what we want the output of a

decision-making system to have.

We are in strong agreement with these statements. Also, we believe that "intellectual coherence or reliability" can be demonstrated using the Theory of Generalizability.

Analysis of Variance (ANOVA) and Variance Components

The theory of generalizability in the conduct of both G and D studies makes extensive use of ANOVA models which are more general and incorporate as special cases the familiar correlational designs utilized by the classical approach. ANOVA designs distinguish between random and fixed factors. A random factor is one in which the levels of the factor are considered a sample from a universe of all possible levels whereas a fixed factor exhausts all levels of interest for that factor. It should be obvious from the previous discussion that a coefficient of generalization for any condition of a G or D study makes sense only if the levels of the factor for that condition are random. A fixed factor in an ANOVA design exhausts all the levels of interest for that factor and there is nothing to generalize to. However, many G and D studies may employ both fixed and random factors (mixed designs).

The general procedure in conducting either a G or D study is to utilize an ANOVA design which will then yield the familiar sums-of-squares and mean squares. The conduct of F tests, however, is rarely done since one is not usually interested in testing hypotheses but rather in estimating various expected values of the mean squares in the fashion suggested by Cornfield and Tukey (1956). These estimates are then used to report the results of the study in terms of the components of variance accounted for and an estimate of the coefficients of generalizability. We will dispense with the formal theory which is well presented in Cronbach, Gleser, Nanda and Rajaratnam (1972) and any good ANOVA book such as Winer (1970), or Kirk (1968) and proceed to give several numerical examples.

Examples

The first example uses fictitious data in a simple study of how to analyze judgments of the importance of attributes as they might be obtained in a typical MAU study. The second two examples are more complicated and use data from actual experiments.

Example 1: Analysis of raters making importance judgments about attributes.

In MAU studies one task for the "expert" subjects is to make judgments of importance for each of the attributes under consideration for the decision. Suppose we have four experts rate each of six attributes on importance on a 10 point scale (1 = least important; 10 = most important). The result might be like that reported in Table 1.

Insert Table 1 about here

Since we are interested in how well the rater might be doing at this task we ask questions about the generalizability of the measuring instrument, i.e., the raters judging importance of attributes. The data in Table 1 are easily analyzed by ANOVA with the results given in Table 2.

Insert Table 2 about here

The expected mean squares E(MS) in the last column of Table 1 are the population values of the sample variances (mean squares). For those readers not familiar with expected mean squares the following intuitive explanation is offered: each expectation consists of an error variance component, since all

Table 1

Importance Ratings Given by 4 Raters
on Each of 6 Attributes

		Attribute						
		1	2	3	4	5	6	Mean ($\bar{R}$)
Rater	1	2	5	1	7	2	6	3.84
	2	4	7	3	9	4	8	5.84
	3	3	5	1	9	6	8	5.34
	4	3	6	2	8	1	4	4.00
Mean ($\bar{A}$)		3.00	5.75	1.75	8.25	3.25	6.50	4.75

Table 2

Analysis of Variance of Ratings

Source of Variation	Degree of Freedom	Sums of Squares	Mean Square	Expected Mean Square E(MS)
A (Between Attributes)	5	122.50	24.50	$\sigma_e^2 + 4\sigma_A^2$
B (Between Raters)	3	17.50	5.83	$\sigma_e^2 + 5\sigma_B^2$
AB (Residual)	15	18.50	1.23	σ_e^2
TOTAL	23	158.50		

sample measurements are infested with error plus a "weighted" variance component of the assumed treatment effect, in this case the effect due to the rows (raters) and that due to columns (attributes). Each treatment variance is weighted by the number of sample values that contributed to each treatment level mean. For example, the sum-of-squares due to attributes in Table 1 came from the fact that each column (attribute) mean is deviating from the grand mean, with each deviation being squared and summed. Each of these squared values, however, are representing the four values that went in to calculate each mean. Thus the expected value component for the effect due to attributes (σ_A^2) gets weighted by 4. In a similar fashion six values contributed to the mean for each row effect and thus the expected value component for the effect due to raters gets weighted by 6.

The general method for obtaining the expected mean squares E(MS) for ANOVA designs is straight forward but involves tedious algebra. Fortunately, Cornfield and Tukey (1956) have provided a convenient algorithm which enables one to set down the appropriate expected mean squares for any ANOVA design. This algorithm is explained in Kirk (1968) and Winer (1970). Since the calculation of expected mean squares are so crucial for the theory of generalizability a version of the Cornfield and Tukey algorithm is given in the appendix to this paper.

Once the expected mean squares are set down it is a matter of simple algebra to estimate the variance components in the experiment. For example, the variance component for factor A (attributes) is

$$\hat{\sigma}_A^2 = \frac{MS_A - MS_e}{4}$$

$$= \frac{24.50 - 1.23}{4}$$

$$= 5.82$$

Where $\hat{\sigma}_A^2$ is an estimate of the variance component σ_A^2 and MS_A and MS_e are the mean squares for the A factor and error respectively. Since this was an ANOVA with only one observation per cell the MS for interaction is the error variance component. Incidentally, while variance components cannot be negative, estimates of such components can be negative due to sampling error. If any variance estimate is negative, it should be set equal to zero. (An example of such a possibility is given in the third example.)

The estimate of each variance component and the percent of variance accounted for in the experiment can be listed in the following diagram:

Variance Component	A	B	AB(Res)	Total
Variance Estimate	5.82	.77	1.23	7.81
Percent of Variance	74	10	16	100

This is a fairly precise experiment with about 16% of the variance due to "error" and thus 84% is predicted variance.

Now, we can estimate the coefficient of generalizability (G^2) for the experiment as follows:

$$\hat{G}^2 = \frac{\text{Estimated Universe Score Variance}}{\text{Estimated Observed Score Variance}}$$

$$= \frac{\hat{\sigma}_A^2}{\hat{\sigma}_A^2 + \hat{\sigma}_e^2}$$

$$= \frac{5.82}{5.82 + 1.23}$$

$$= .83$$

This is a version of the intra-class correlation coefficient and Cronbach, Ikeda, and Avner (1964) illustrate that the intra-class correlation coefficient is an estimate of the lower bound of G^2.

An interpretation of $\hat{G}^2$ is that it is the proportion of universe score variance accounted for by knowledge of the observed score variance.

Another interpretation is as follows: If the experiment were to be repeated with another random sample of four raters from the same universe of raters who then rate the same attributes, the squared correlation between the mean ratings for the two sets of raters would be .83. The closer $\hat{G}^2$ is to one the more representative is the sampled data of the universe of interest. Although a $\hat{G}^2$ of .83 is quite respectable, it is well to ask of the G study why it isn't higher? For the data given in Table 1 the answer lies in the presence of interaction between the raters and the thing being rated (attribute importance). A glance at the numbers in Table 1 indicates that there is excellent agreement between raters 1 and 2; rater 3 is for the most part in agreement but deviates somewhat with the first two raters. Rater 4 is different from raters 1, 2, and 3. Any inconsistency in such raters shows up in the interaction term and therefore tends to inflate the error variance component. In order to move the coefficient of generalizability (G^2) closer to one, a reduction in this interaction would have to occur.

Example 2: The JUDGE Experiment. (A Decision (D) Study)

As our second example we use the experimental results of a somewhat novel application of decision theory to a Tactical Air Control System provided by the work of Miller, Kaplan and Edwards (1967, 1968). They have proposed a system called Judged Utility Decision Generator (JUDGE) for allocating air strike missions to requests in tactical air control environments. A key concept of JUDGE is that value judgments (estimation of military worth associated with requests) can be made directly and in real time by appropriately trained personnel, and that the system should, in principle, maximize the aggregate utility over all the dispatching decisions it makes. The actual details of how JUDGE works is not important for this discussion. What is important is that the system makes extensive use of human judgments which are computer assisted and the advocates of this system claim that it might be a better system than that presently being used by the Air Force in tactical situations. In order to lend support to this notion these investigators conducted a laboratory study comparing JUDGE against a version of the currently used system called Direct Air Support Center (DASC). We have taken the liberty of using the data reported for this experiment to illustrate the basic concepts and interpretations of the theory of generalizability and what is presented below should not be construed as an interpretation of the Miller et al. experiment. In the experiment, each of 14 subjects participated in both the JUDGE and DASC modes of tactical command, in four simulated air tactical command stations, and there were two replications of the experiment. Thus from an ANOVA design viewpoint this can be considered as a three factorial design with Systems, Situations, and Subjects being the three factors of interest or, more specifically, a 2 (systems) x 4 (situations) x 14 (subjects) design with two (2) replications per cell (each subject performed twice). Two dependent variables were used in this study: a measure of efficiency and a measure of effectiveness.[1]

Since the main purpose of this study was to decide which system, JUDGE or DASC, was the best, we will first consider the analysis of the results of the study as a D (decision) study. Table 3 presents the ANOVA summary of the results using the effectiveness measure as the dependent variable. (The results for the efficiency dependent variable are similar and will not be presented.)

Insert Table 3 about here

The calculation of the Expected Mean Squares E(MS) assumes the A and B factors are fixed and the C (subjects) and Within Replicates are random factors.

Again, once the expected mean squares are set down it is a matter of simple algebra to estimate the variance components of the factors and their interactions in the experiment. For example, the variance component for factor A (Systems, JUDGE vs. DASC) is estimated by

$$\hat{\sigma}_A^2 = \frac{MS_A - MS_{AC}}{112} = \frac{1834.33 - 9.89}{112} = 16.29$$

[1] The data are presented in Appendix B of Miller, Kaplan and Edwards (1968).

Table 3

ANOVA Summary Table for the Judge Experiment

Source of Variation	Degree of Freedom	Sums of Squares	Mean Squares	Expected Mean Square E(MS)
A (Systems)	1	1834.33	1834.33	$\sigma_e^2 + 8\sigma_{AC}^2 + 112\sigma_A^2$
B (Situations)	3	21.70	7.23	$\sigma_e^2 + 4\sigma_{BC}^2 + 56\sigma_B^2$
C (Subjects)	13	63.89	4.91	$\sigma_e^2 + 16\sigma_C^2$
AB	3	65.82	21.94	$\sigma_e^2 + 2\sigma_{ABC}^2 + 28\sigma_{AB}^2$
AC	13	128.57	9.89	$\sigma_e^2 + 8\sigma_{AC}^2$
BC	39	114.94	2.95	$\sigma_e^2 + 4\sigma_{BC}^2$
ABC	112	91.51	0.82	σ_e^2
TOTAL	223	2320.76		

where $\hat{\sigma}_A^2$ is an estimate of the variance component σ_A^2 and MS_A and MS_{AC} are the mean squares for the A factor and the AC interaction respectively.

This estimate of each variance component and the percent of total variance accounted for in the experiment can be listed in the following diagram:

Variance Component	A	B	C	AB	AC	BC	ABC	Error	Total
Variance Estimate	16.29	.07	.25	.68	1.13	2.13	.93	.82	22.30
Percent of Variance	73	00	01	03	05	09	04	04	100%

Note that this was a very precise experiment; only 4% of the total variance is due to error and thus 96% is predicted variance. However, the variance due to the different systems (JUDGE vs. DASC) dominates the picture. The JUDGE system was considerably more effective than DASC. (The mean effective performance was 18.77 for JUDGE as contrasted with 13.05 for DASC.) However, this factor was so strong, i.e., accounted for such a large percentage of variance we can say very little about the other factors in the experiment. Incidentally, expressing the results of the ANOVA in a diagram such as the one above is much more revealing than just reporting F ratios which can be very misleading. For example, the F ratio testing the significance of the ABC trial interaction against the error term yielded an F of 3.26 which, for 39 and 112 degrees of freedom, is highly significant ($P < .001$) but this interaction only accounts for 4% of the variance, which is hardly of practical significance.

Since this particular experiment indicates that JUDGE is a more effective system in a simulated tactical air control environment than DASC, we can turn our attention to determining the generalizability of JUDGE as contrasted to DASC. To do this, we will reanalyze the data for each system separately. This is valid since the initial design was completely crossed, i.e., all subjects participated in all conditions of the experiment.

Analysis of the JUDGE Experiment: A Generalizability (G) Study

As an illustration of how a generalizability study may proceed, we have reanalyzed the JUDGE experiment as two 2-factor experiments; one with the subjects operating under the DASC system and the other with the same subjects operating under the JUDGE system. Two separate ANOVAs were carried out with the results displayed in Tables 4 and 5. In Tables 4 and 5, the A factor (situations) is considered fixed and the B factor (subjects) and within replicates are random effects. The appropriate expected mean squares are given in the last column of each table and from these estimates of the variance components and the percent of total variance can be calculated and are given under each table. Now since we wish to generalize to the population of subjects we can estimate a coefficient of generalization for each of the two systems as follows:

Insert Tables 4 and 5 about here

$$\hat{G}^2 = \frac{\hat{\sigma}_B^2}{\hat{\sigma}_B^2 + \hat{\sigma}_{AB}^2 + \hat{\sigma}_e^2} \text{ , A factor fixed}$$

$$\hat{G}^2 \text{ (DASC)} = \frac{1.27}{1.27 + 1.80 + 1.22} = .30$$

$$\hat{G}^2 \text{ (JUDGE)} = \frac{.37}{.36 + .20 + .42} = .37$$

TABLE 4

ANOVA Summary: DASC System

Source	Degree of Freedom	Sums of Squares	Mean Square	Expected Mean Square
A (Situations)	3	78.25	26.08	$\sigma_e^2 + 2\sigma_{AB}^2 + 28\sigma_A^2$
B (Subjects)	13	148.46	11.42	$\sigma_e^2 + 8\sigma_B^2$
AB	39	187.70	4.81	$\sigma_e^2 + 2\sigma_{AB}^2$
Within replicates	56	68.10	1.21	σ_e^2
Total	111	482.51		

Variance Component	A	B	AB	Error (e)	Total
Variance Estimate	.76	1.27	1.80	1.22	5.05
Percent of Total Variance	15	25	36	24	100

TABLE 5

ANOVA Summary: JUDGE System

Source	Degree of Freedom	Sum of Squares	Mean Square	Expected Mean Square
A (Situations)	3	9.26	3.09	$\sigma_e^2 + 2\sigma_{AB}^2 + 28\sigma_A^2$
B (Subjects)	13	42.99	3.31	$\sigma_e^2 + 8\sigma_B^2$
AB	39	31.75	.81	$\sigma_e^2 + 2\sigma_{AB}^2$
Within replicates	56	23.42	.42	σ_e^2
Total	111	107.42		

Variance Component	A	B	AB	Error (e)	Total
Variance Estimate	.08	.36	.20	.42	1.06
Percent of Total Variance	7	34	19	40	100

Thus we see that the JUDGE system has a higher coefficient of generalizability and thus can be considered as more dependable (reliable and valid) than the DASC system. We need not seek nor rely on some outside independent criterion to help us reach this decision. We can also see that the DASC system is not that much worse than JUDGE with respect to its generalizability for subjects, being only seven percent "poorer". This was due primarily to the fact that the interaction of subjects x situations variance component was higher for the DASC system than for the JUDGE system. This interaction term gets included in the estimate of the total observed score variance. Also, what it means from a behavioral standpoint, is that the subjects when in the DASC mode were not being as consistent in their responses to the four situations as when performing in the JUDGE mode. It should be remembered that the tasks for the two subjects are different in the two systems. In DASC the subjects are asked to make dispatching-like decisions in the simulated tactical situations whereas in JUDGE they are making value judgments (estimation of military worth associated with requests for a mission). These value judgments are expressed numerically combined with an estimated probability of "kill" along with certain constraints such as the availability of aircraft,and the dispatching design is made automatically by a computer generated dispatching rule. The data presented in Tables 4 and 5 indicate that when subjects are asked to make value judgments and these judgments are then used in an automatic algorithm then the entire system responds more consistently to various situations. With this design the presence of interaction effects tends to reduce the generalizability over any set of conditions.

One final point before leaving this example. Although the situations factor was fixed in this analysis, it does appear that this was a stronger independent variable when the subjects were performing under the DASC system (15% of the total variance) than when they were performing under the JUDGE (7% of the total variance). This might suggest that in a future G study in which generalizability to situations might be a desired goal that the DASC system might fare better than JUDGE. Actually no such prediction can be made until the actual G study is performed with the situations variable being included in the design of the study as a random factor. However, it is often the case that generalizability to situations, stimuli, tasks, and so on are as important, if not more important, as generalizations about people. An illustration of such a case is given in the next numerical example.

Example 3: Analysis of the Gardiner Study on Coastal Zone Management Decisions

Gardiner (1974) in his study applying multi-attribute utility techniques to Public Policy decision making had his subjects make judgments about whether certain permit requests for various developments along the Southern California Coast should be approved or disapproved by a Coastal Commission which has the authority to approve or deny such requests. The subjects, some of whom were actual commission members, made intuitive wholistic judgments and also made value or worth judgments about the worth of each permit along eight different attributes characterizing each permit request. The attributes, which included such things as the height of the proposed development, distance from the water's edge, and amount of parking space, were those that are actually used in making such decisions. Gardiner took special pains to select a sample of 15 permits that were "typical" of the kind that usually come before the Coastal Commission. He also was interested in comparing two sub-groups of his subjects who described themselves as "Developers", i.e., generally leaning toward development of the coastal line, and "Conservationists", who were generally opposed to developments that might destroy the natural coastal line. One phase of his analysis utilized a two factor ANOVA, with the groups (Developers vs. Conservationists), and permits being the two factors. He had 15 permits and 7 subjects in each group, thus this was a 2 x 15 factorial design with 7 replications per group. The results are given in Tables 6 and 7. Table 6 is the result for the wholistic evaluation of permit worth and Table 7 is for the MAU evaluation of permit worth.

Insert Tables 6 and 7 about here

In these two tables the group factor (Developers vs. Conservationists) is considered fixed and the permit factor and within replicates are considered to be random.

The coefficients of generalizability (G^2) for the two analyses given in Tables 6 and 7 are:

$$G^2 = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_{AB}^2 + \sigma_e^2} \quad \text{, A factor fixed}$$

$$G^2 \text{ (wholistic)} = \frac{274.58}{274.58 + 157.94 + 431.35} = .32$$

$$G^2 \text{ (MAU)} = \frac{124.78}{125.78 + 0 + 106.09} = .54$$

Thus, the MAU technique is more dependable (generalizable) than its wholistic counterpart. In other words, if one were to generalize to the domain or population of all possible permits (of the kind investigated in Gardiner's study) then one would be in a better position using MAU than wholistic judgments.

TABLE 6

ANOVA Summary: Wholistic Evaluation of Permit Worth

Source	Degree of Freedom	Sum of Squares	Mean Square	Expected Mean Square
A (Group)	1	13,675.24	13,675.24	$\sigma_e^2 + 7\sigma_{AB}^2 + 105\sigma_A^2$
B (Permits)	14	59,845.38	4274.67	$\sigma_e^2 + 14\sigma_B^2$
AB	14	21,516.18	1536.87	$\sigma_e^2 + 7\sigma_{AB}^2$
Within replicates	180	77,643.00	431.35	σ_e^2
Total	209	172,679.80		

Variance Component	A	B	AB	Error (e)	Total
Variance Estimate	115.60	274.53	157.94	431.35	979.42
Percent of Total Variance	12	28	16	44	100

TABLE 7

ANOVA Summary: MAU Technique Evaluation of Permit Worth

Source	Degree of Freedom	Sums of Squares	Mean Square	Expected Mean Square
A (Group)	1	2128.51	2128.51	$\sigma_e^2 + 7\sigma_{AB}^2 + 105\sigma_A^2$
B (Permits)	14	25,942.00	1853.00	$\sigma_e^2 + 14\sigma_B^2$
AB	14	1086.82	77.63	$\sigma_e^2 + 7\sigma_{AB}^2$
Within replicates	180	19,096.20	106.09	σ_e^2
Total	209	48,253.53		

Variance Component	A	B	AB[1]	Error (e)	Total
Variance Estimate	19.54	124.78	0	106.09	250.41
Percent of Total Variance	08	50	0	42	100

[1] The mean square estimate ($\hat{\sigma}_{AB}$) for the AB interaction was negative due to sampling error and was set to zero.

Note that the distribution of predicted variance is quite different for the two methods with the most dramatic difference being the elimination of the interaction component when the subject's responses are used under the MAU technique. The statements made earlier are worth repeating here: When human judgment is being used in any scientific or practical study, any interaction between the human judges and the objects, conditions or persons being judged may be a form of inconsistency and lowers the dependability (generalizability) of those judgments. There may be an important principle here, one of considerable theoretical and practical importance. Any "divide-and-conquer" technique such as a MAU technique may minimize or at least reduce substantially interaction sources of variance due to inconsistency thus making any study or application of the technique easier to interpret. This is not to say that components of variance due to interaction should always be reduced. There are certainly situations in which individual differences represent valid differences in judgments about utilities. Arch conservationists may have quite different ideas about what is "best" for the California coastline than arch developers. We certainly would not want a technique that blurs or reduces such differences. The theory described in this paper must be applied to such situations in laboratory and "field" studies to see how useful the theory is in such situations.

Comment on Random Sampling

The theory of generalizability makes one powerful assumption: any sample of observations must be a representative random sample from the universe or population one wishes to generalize to. The question immediately arises as to whether one should truly use the operation of complete random sampling of conditions from some universe in order to generate the set of conditions to be used in any G study. Presumably Brunswik (1956) would argue yes but I would argue that it is not a necessity. The choice of the levels of a factor in an ANOVA design must rest with the investigator and it is the responsibility of that investigator to state whether that factor is random or fixed for any given situation and give his reasons, which other investigators may or may not agree with. It may be that the use of random sampling may be the best way to choose the levels as, for example, in the study of form perception using methods suggested by Attneave (1954). In selecting what development permits to be used in his study, Gardiner could have used some random sampling scheme such as going to the files of proposed permits in the California Coastal Commission's office and by using some random plan select his 15 permits. However, this runs the risk, with a fairly high probability, of yielding a set of 15 permits that were not as representative of the universe of permits as it should be. What Gardiner did was to rely on expert judgment (his own) to select a list of permits that would be useful in his study. This list covered a broad range of typical permits that included almost all of the kinds of proposed developments of interest to the study and the practical application in mind.

The assumption of random sampling in any G or D study should remain that: an assumption on the part of the investigator. Of course all the principles involved in good experimental design should be employed to make that assumption reasonable and plausible.

Summary

This paper has presented a theoretical rationale for assessing the dependability, validity, reliability or intellectual coherence of multi-attribute utility models and techniques. If an investigator is advocating the use of a MAU model or procedure he or she is interested in generalizing from observations at hand to a universe or domain of observations that are members of that same universe. The universe must be unambiguously defined but it is not necessary to assume that universe as having any statistical properties such as uniform variances or covariances. A study of generalizability (G study) is conducted by taking measurements on persons, stimuli, tasks, etc. that are assumed to be randomly representative of a universe an investigator wishes to generalize to. The ratio of an estimate of the universe "score" variance to an estimate of the observed score variance is the coefficient of generalizability. This is estimated by the intra-class correlation coefficient. ANOVA and the Expected Mean Square paradigm of Cornfield and Tukey is used to obtain the appropriate variance estimates.

The theory dispenses with unnecessary and unwarranted assumptions, and eliminates the distinction between reliability and validity. Any G study can be conducted without reference to having a parallel measure of the MAU instrument or some external criterion of "success". If a MAU technique is compared to some non-MAU technique for doing the same thing then it is possible to calculate the coefficient of generalizability for both methods thus allowing the investigator to decide which is best for his or her purposes. Preliminary investigations have indicated that MAU models and techniques based on such models may be "better" than non-MAU models since the former have a tendency to reduce the interaction between judges and the thing being judged when such interaction represents inconsistency of judgment. The extent of this principle, if indeed it is true at all, needs further work.

REFERENCES

Attneave, J. Some informational aspects of visual perception. *Psychological Review*, 1954, *61*, 183-193.

Brunswik, E. *Perception and the Representative Design of Experiments*. Berkeley: University of California Press, 1956.

Campbell, D. and Stanley, J. *Experimental and Quasi-Experimental Designs for Research*. Skokie, Illinois: Rand McNally, 1963.

Cornfield, J. and Tukey, J.W. Average values of mean squares in factorials. *Annals of Mathematical Statistics*, 1956, *27*, 907-949.

Cronbach, L.J., Gleser, G., Nanda, H. and Rajaratnam, N. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: Wiley & Sons, 1972.

Cronbach, L.J., Ikeda, H. and Avner, R.A. Intraclass correlation as an approximation to the coefficient of generalizability. *Psychological Reports*, *15*, 1964, 727-736.

Edwards, W. Social utilities. *The Engineering Economist*, Summer Symposium Series, *VI*, 1971.

Fisher, G.W. Multi-dimensional value assessment for decision making. The University of Michigan Engineering Psychology Laboratory, Technical Report 037230-2-T, June, 1972.

Gardiner, P.C. Public Policy Decision Making: The Application of Decision Technology and Monte Carlo Simulation to Multiple Objective Decisions - A Case Study in California Coastal Zone Management, Ph.D. dissertation, Urban Studies, University of Southern California, 1974.

Greeno, J. Elementary Theoretical Psychology. Reading, Mass.: Addison-Wesley, 1968.

Gulliksen, H. Theory of Mental Tests. New York: Wiley & Sons, 1950.

Haggard, E. Intraclass Correlation and the Analysis of Variance. New York: Holt, Rinehart, and Winston, 1958.

Huber, G.P., Daneshgar, R., and Ford, D.L. An empirical comparison of five utility models for predicting job preferences. Organizational Behavior and Human Performance, 6, No. 3, 1971, 267-282.

Kirk, R. Experimental Design Procedures for the Behavioral Sciences. Belmont, California; Wadsworth, Inc., 1968.

Lord, F.M., and Novick, M. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Miller, L., Kaplan, R., Edwards, W. JUDGE: A value-judgment-based tactical command system. The RAND Corp., RM-5147-PR, March, 1967.

Miller, L., Kaplan, R., Edwards, W. JUDGE: A laboratory evaluation. The RAND Corp., RM-5547-PR, March, 1968.

Raiffa, Howard. Decision Analysis: Introductory Lectures on Choices Under Uncertainty. Reading, Mass.: Addison-Wesley, 1968.

Raiffa, Howard. Preferences for Multi-Attributed Alternatives. Santa Monica: Rand Memorandum RM-5868-DOT/RC, April, 1969.

Tryon, R.C. Reliability and behavior domain validity: reformulation and historical critique. Psychological Bulletin, 1957, 54, 229-249.

Winer, B.J. Statistical Principles in Experimental Design. 2nd ed. New York: McGraw-Hill, 1970.

Appendix A

An Expected Mean Square Algorithm

The theory of generalizability requires an investigation to estimate variance components. This in turn requires the calculation of the expected value of the mean squares (MS) generated by an analysis of variance (ANOVA) study. The calculation of these expected mean squares E(MS) is straightforward but involves tedious algebra. Fortunately, Cornfield and Tukey (1956) have provided a convenient algorithm for generating the E(MS)s for any ANOVA design. This procedure is explained in standard texts such as Winer (1971), and Kirk (1968).

The procedure is illustrated below by following a set of rules. This is a modification of that provided by Kirk (1968, p. 209-210).

Rule 1. Write the linear model for the design. If, for example, there are two factors A, B and n replications the model is:

$$Y_{ijm} = M + A_i + B_j + AB_{ij} + e_{ijm}$$

Rule 2. Construct a two way table such as Table A as follows:

(a) The rows of the table are labeled as the factor effects excluding the general mean. The columns of part 1 of the table are labeled with the subscripts and the limit of the subscript (number of levels of each factor).

(b) Part 2 of the table is labeled as E(MS)

TABLE A
Expected Values of Mean Squares for a Two Factor ANOVA Design

	1			2
	i	j	m	E(MS)
	a	b	n	
A_i	$1 - \frac{a}{A}$	b	n	$\sigma_e^2 + n(1 - \frac{b}{B})\sigma_{AB}^2 + nb\sigma_A^2$
B_j	a	$1 - \frac{b}{B}$	n	$\sigma_e^2 + n(1 - \frac{a}{A})\sigma_{AB}^2 + na\sigma_B^2$
AB_{ij}	$1 - \frac{a}{A}$	$1 - \frac{b}{B}$	n	$\sigma_e^2 + n\sigma_{AB}^2$
e_{ijm}	1	1	$(1 - \frac{n}{N})$*	σ_e^2

Rule 3. Each entry below each column in part 1 is determined as follows:

(a) If the column heading appears as a subscript of a row tern enter the sampling fractions appropriate for that column $1 - \frac{a}{A}$, $1 - \frac{b}{B}$, etc., where a and b are the levels of each factor and A and B are the total number of possible levels.

(b) If the column heading does not appear as a subscript of a row term enter the appropriate letter for that column, e.g., a, b, n, etc. in the row.

(c) The last row should contain all ones under each column heading. For most designs there is no sampling fraction for the replicates effect, since the n replicates for any experiment is usually very small relative to the total number of possible replicates, i.e., $(1 - \frac{n}{N}) = 1$ for large N.

* $(1 - \frac{n}{N}) = 1$ since in almost all applications the number of replications n is considered very small relative to all possible replication N.

Rule 4. For each row in part 2 of the table (E(MS)) list the variance of the linear model term that contain all the subscripts of the row term, for example, the subscript of the first row is i. Variances in terms of the linear model that contain subscript i are σ_e^2, σ_{AB}^2, and σ_A^2.

Rule 5. The coefficients of the variance for each E(MS) are obtained by covering up the columns headed by the subscripts that appear in a row and multiplying each row E(MS) variance by the remaining terms in part 1 of the table. For example, the coefficients in the first row for $\sigma_A 2$ are n and b which are found in the first row of the table. The coefficients for $\sigma_{AB} 2$ are n and $(1 - \frac{b}{B})$ which are found in the third row of the table. The coefficient for σ_e^2 is always 1.

The E(MS) for any main effect always includes the error variance σ_e^2 plus all variance terms in which it is included. In other words the E(MS) is a weighted sum of all the variance components that contain the subscripts of the main effects.

Rule 6. The sampling fractions $(1 - \frac{a}{A})$, $(1 - \frac{b}{B})$, etc. tend to reduce the variance term for which they are coefficients and suppress them completely when the factor is fixed. For example, if factor A is fixed and thus a exhausts all levels of interest a = A and $(1 - \frac{a}{A}) = 0$. If the factor is considered random and a is small relative to A the sampling fraction is one. There may be practical situations in which values for the sampling fractions between 0 and 1 may be appropriate but these two values are most often used.