USING GENERALIZABILITY THEORY TO ASSESS THE DEPENDABILITY OF DIRECT MAGNITUDE SEPARATION ESTIMATES (GALILEO DATA)[1]

M. Mark Miller

In recent years a number of communication researchers have been gathering data by means of a pairwise magnitude estimation task. In general, this method takes any two concepts in the domain under study, assigns an arbitrary unit size to the difference between them, and asks respondents to estimate the differences between all possible pairs of concepts under consideration as ratios of the arbitrarily designated standard. Typically, questions take the form: "If A and B are x units apart, how far apart are C and D?"

Data from such questions usually are averaged across a sample of respondents and these averages are submitted to a metric multidimensional scaling program available under the tradename "Galileo." The Galileo$_{TM}$ computer software contains a variety of algorithms that (1) provide graphic plots of the derived multidimensional spaces, (2) perform least-squares rotation to congruence of two or more separately derived spaces facilitating intergroup and over-time comparisons, and (3) recommend message strategies for meeting persuasive goals.

Proponents of the Galileo approach have applied it in a variety of contexts including planning of marketing and political campaigns (Serota et al., 1977), studies of cultural convergence among immigrant groups (Kincaid et al., 1981), studies of accuracy of translations between languages, (Barnett, et al., 1984) and tests of various theoretic processes (Cody, 1980; Marlier, 1983). Details of the method and its applications are beyond the scope of this presentation; however, Woelfel and Fink (1980) provide a detailed exposition, and Barnett (1980, Appendix) provides a bibliography of several dozen papers.

Researchers who use Galileo methods frequently face skeptical reviewers. These reviewers doubt the quality of the data because the scaling task presented to respondents is both unusual and demanding. Further, these critics argue that subjecting poor data to the elaborate mathematical algorithms of the Galileo$_{TM}$ computer programs might capitalize on error and lead to erroneous conclusions.

Faced with such criticism the Galileo researcher can argue that prior studies of the measurement qualities of the data have indicated that they are sound. Such an argument is weak, however, because it requires the assumption that earlier studies generalize to the immediate context. Often there are appreciable differences between procedures, populations, and concepts of a study and those used in prior research that could result in differential data quality. In short, it is reasonable to require the Galileo researcher to report explicitly the quality of the data under consideration.

The Galileo researcher can assess the quality of his or her data by replicating the procedures used in past research. These procedures, however, are arduous and unconventional. Also, such

procedures either disregard useful information in the data, or subject the data to elaborate mathematical manipulation before examining measurement quality, or both.

What is needed is a procedure for assessing the quality of the data (1) that is easy (mitigating the temptation to assume quality rather than examine it), (2) that is conventional (making the procedure acceptable to the community of scholars), and (3) that does not disregard available information. This presentation describes such procedures and provides empirical examples.

Prior Research on Data Quality

Correlations Between Aggregated Means

Several Galileo researchers have reported test-retest or split-sample reliability coefficients. In an elaborate study of Galileo methods, Gordon (1977, Chapter 10) reported the correlations among the means of pair separation estimates among nine groups of about 100 college students. Gordon studies perceived distances among six popular television shows, six television show types and the concept "me." He used a no-concept pair condition telling respondents to keep the number 10 in mind while making evaluations, varied the concepts used for comparison, and varied the number of units between anchor concepts. He reported product moment correlations among the nine groups' mean estimates of the separations among the 78 pairs ranging between .933 and .988.

Gillham and Woelfel (1977) focused on test-retest correlations among means in a study of Galileo of 24 graduate students and five faculty members evaluations differences among 19 professors in a sociology department. They administered questionnaires to respondents at three time intervals averaging 16.3 weeks between the first and second administrations and 6.3 weeks between the second and third administrations. They reported a correlation of .71 between time 1 and time 2 and between time 2 and time 3, and .65 between time 1 and time 3. While these correlations are moderately high, Gillham and Woelfel argue that they are indicative of a substantial amount of change in mean perceptions of differences among the faculty members. This argument is based in part on the fact that nine-point ordinal scales asking the same respondents to evaluate the same professors yielded test-retest correlations of more than .97. Thus, Gillham and Woelfel favor an argument that magnitude separations are precise and sensitive to subtle changes rather than holding that the moderate correlations are due to unreliability.

Craig (1977), in a study of message effects on 64 students perceptions separations among countries, also reports test-retest correlations. The correlation for all distances in Craig's data (N=105) was .87; for distances not manipulated by messages (N=36), .91, and for manipulated distance (N=69), .84. This pattern is what would be expected under Galileo theory with unmanipulated concepts having higher test-retest correlations than concepts subjected to manipulations.

In a study of cultural convergence of Korean immigrants, Kincaid et al. (1983) report more conventional test-retest correlations. They apparently repeated five pair comparison questions in a questionnaire containing 55 such pairs concerning cultural concepts that was administered to members of five different cultural groups. They note that the reliability coefficient increases as the level of aggregation increases. This finding is to be expected because it is well known that errors of estimates about a mean decrease in proportion to the reciprocal of the sample size. The test-retest correlation at the individual level for all five groups taken together was .65. When computed separately for each group the correlations ranged between .90 and .97. When the means for each group (including from 30 to 48 members) was used, the reliability coefficients ranged from .975 to .996. This pattern is indicative of general agreement concerning the separations across the cultural groups, higher agreement within groups, and increased reliability with aggregation.

Correlations Between Eiegenvectors (Factor Loadings)

Other research has focused on the test-retest or split sample correlations among the eigenvectors (concept loadings) of items on each of the dimensions derived by the multidimensional scaling procedure. Such correlations generally are calculated following a least-squares rotation of two or more spaces to congruence. Obviously such rotation procedures maximize the similarities between spaces in a manner that could capitalize on error. To the degree that capitalization on error inflates the correlations, the derived reliability coefficients are over estimated.

Barnett (1972) offered the first examination of test-retest reliability of the eigenvalues. He had 300 undergraduates at a large midwestern university provide direct magnitude separation estimates of the distances among 16 environmental concepts on two occasions separated by two weeks. To ascertain the number of raters needed to yield reliable data, Barnett calculated the canonical correlations among the eigenvectors for randomly selected subsamples of 25, 50, 75, and 100 individuals. As expected, the reliabilities as indicated by the canonical correlations increased with sample size. These correlations stabilized around .9 at 75 cases and up.

Craig (1977) reported correlations among eigenvectors under two different rotation procedures. The values ranged from .52 to .99 with seven of the ten values reported exceeding .90. The coefficients varied as much as .12 depending on the rotation procedure. Such a difference demonstrates that the coefficients are sensitive to the rotational procedures.

Even assuming that the rotation procedure provides an adequate basis for such correlations, problems remain. This approach results in a number of reliability coefficients, rather than a single overall measure. After multidimensional spaces have been rotated to congruence, correlations are computed between respective columns of the coordinate matrices to assess

the reliability of each dimension. Also correlations are computed for each of the concepts to assess the degree of reliability of its location in the space. Thus, the procedure can result in twice as many reliability coefficents as there are concepts under consideration.

Much of the attention to correlations among eigenvecotrs has focused on the reliability of the "imaginary dimensions" derived by the metric multidimensional scaling procedures used in the Galileo$_{TM}$ computer package. Such dimensions occur because observed separation estimates often violate the "triangle inequality" rule; that is, two sides of a triangle formed by three points do not sum to at least the length of the third side. When this occurs, some of the eigenvalues extracted from the scalar products matrix of the separation estimates are negative. Most Galileo theorists interpret negative eigenvalues as imaginary dimensions and they consider the derived space to conform to Riemann rather that Euclidian geometry.

Most nonmetric multidimensional scaling algorithms assume a Euclidian space and take triangle inequalities to be indicative of measurement error. In fact, in many nonmetric algorithms tranformation of the data to eliminate such inequalities is considered a key step. Most Galileo theorists oppose such transformations on the grounds that the imaginary dimensions contain reliable information. Thus the reliability of such dimensions becomes a key issue.

Craig (1977) noted that correlations between eigenvectors tend to be largest for dimensions with the largest absolute eigenvalues and reports only five of 15 such correlations that would have been available in this data. Cody (1980) argues that imaginary dimensions should be eliminated from consideration for mathematical reasons and that only reliable dimensions should be retained. Cody does, however, report all eigenvector correlations between a control group and several experimental groups. These correlations conform to the pattern noted by Craig (1977) with the largest coefficients tending to be associated with dimensions having the largest eigenvalues, and with coefficients for dimensions with small eigenvalues having small or even negative correlations.

Woelfel and Barnett (1982) disregard Cody's mathematical argments concerning the utility of imaginary dimensions and argue that such dimensions should be used because Cody's data reveal acceptably high correlations among the loadings on them. Woelfel and Barnett also present time-series data from 20 subjects over 12 time points. The concepts scaled included political candidates, parties, and issues and the data were gathered during a political campaign. Woelfel and Barnett reason that correlations among dimensions would decrease as the lag interval increases if change were due to true change rather than unreliability. This proposition is confirmed through an autoregression procedure. Further, the authors state the regression intercept, .63, "provides the best available estimate of the 'true' reliability of the loadings on the largest imaginary eigenvalue."

It is beyond the scope of this presentation to resolve the controversy concerning the utility of retaining imaginary dimensions. However, it should be noted that if they are to be retained, the low correlations of dimensions with low eigenvalues are of less importance than some researchers seem to think. Often the dimensions with low correlations account for less than one percent of the overall variance so their inclusion or exclusion is of trivial concern in most contexts.

The Dependability Coefficient and Galileo Data

While the above approaches to provide substantial insight into the quality of the data and analysis procedures, they either ignore important information (variance around the mean), or subject the data to mathematical manipulations that preclude an intuitive explanation of the coefficients derived and may inflate reliability coefficients by capitalizing on chance. Further, they require substantial amounts of data. Sample sizes have to be large enough to divide into reasonably sized groups for split sample coefficients, or respondents have to be interviewed twice for test-retest coefficients.

The researchers who use direct magnitude separation estimation measures emphasize that their concern is with cultural processes -- the perceptions and attitudes of homogeneous groups. They also note that averages across a sample of individuals statistically tend to converge on the mean of the population from which the sample was drawn. But neither of these points can justify disregarding variance about the means of the separation estimates.

Because concern centers on aggregate level measures under which individuals are sampled and the mean of their evaluations is used as an estimate of a characteristic of the population under consideration, individuals are treated as raters of stimuli. That is, the method seeks to detect systematic differences among stimuli rather than among individuals. Most discussions of reliability assume that the goal is to distinguish among individuals rather than stimuli.

Techniques for assessing the degree to which raters detect systematic differences among stimuli are well developed in the psychometric literature and data back to Horst's (1949) and Ebel's (1951) classic expositions. These techniques, often labeled intraclass correlation coefficients, are covered in such popular statistics textbooks as Hays (1981) and Winer (1971). An influential monograph by Cronbach et al. (1972) stimulated substantial interest in the techniques which have been discussed by Bartko (1976), Strout and Fleiss (1979), and O'Brien (1984) among others.

Cronbach et al. focus attention on situations in which raters evaluate properties of other individuals; however, as O'Brien (1984) points out, their procedures are applicable when stimuli other that humans are being considered. With reference to Galileo measurement methods, the stimuli are the pairs of concepts and the property being evaluated is the perceived differences among the pairs.

Cronbach et al. develop their own terminology to distinguish their conceptualization from that of classical test theory. Where classical test theorists use the term "true score" to describe the theoretical value of an observation made with no errors of measurement, Cronbach et al. prefer the term "universe score." The notion of universe score emphasizes that the investigator is making an inference from a sample of observed data. Further, it emphasizes that any observations can be used to generalize to a variety of populations in the sense that a single case can be a member of many sets. For example, a single person could be considered to be representative of life forms, mammals, persons, Republicans, etc. Cronbach et al. stress that conditions, occasions, and methods of observation are essential to a complete definition of the universe to which generalizations are attempted. This emphasis on conditions of measurement dispenses with the notion of one true score that is assumed to exist independent of measurement efforts. This conceptualization fits well the relativistic notions of Galileo theorists who also hold that measurements have no meaning outside of their procedural contexts.

Each of the conditions of observation Cronbach et al. consider to be a facet that may account for some portion of the variance of observation and can be estimated via analysis of variance. Thus, their theory is an application of multivariate analysis of variance and interpretations of ratios of variance components. The ratios of concern in this presentation are the dependability coefficients which estimate the percentages of variance in a set of observations that is systematic or reliable. These coefficients are readily interpretable in three mathematically equivalent ways: (1) as the percentage of variance in the observed mean scores due to the variance in the universe score; (2) as the estimated correlation between the observed scores and hypothetical observed scores that could be obtained from another sample using the same procedures, and (3) the square root of the correlation between the observed scores and their universe scores.

Cronbach et al. describe methods of deriving a number of dependability coefficients depending on (1) the universe being generalized to, (2) the number of facets under consideration, and (3) the analysis of variance design. Different coefficients can be derived depending on whether the analysis of variance design is fully crossed (all stimuli are judged by all raters) or is nested (different stimuli are judged by different raters) or is nested (different stimuli are judged by different raters).

The typical Galileo study is fully crossed so only two dependability coefficients will be considered in detail here. Nested designs are possible in Galileo studies; the ramifications of this possibility are considered below. Also, the geometric interpretation of multidimensional spaces used by some Galileo procedures blur the usual distinctions typically made between the two dependability coefficients under consideration. This point is clarified below.

As noted above, the typical Galileo study can be construed to be a fully crossed analysis of variance designed. That is, all

persons sampled estimate the differences between all possible pairs of concepts judged to belong to a cognitive domain. This results in an n by m repeated measures analysis of variance where n is the number of persons (raters or judges) and m is the number of pairs (the product of the number of concepts times the number of concepts minus one divided by two).

Under such a design, three components of variance can be estimated: variance due to persons, variance due to pairs, and residual variance. The residual variance could be considered to contain two components: person-by-pair interaction, and random error. However, since the universe scores of concern are the population means of the pair separations, this distinction is of no importance. That is, although the person-by-pair interaction may indicate a systematic characteristic of an individual, it also indicates the degree to which that individual deviates from the pattern of response of the population. Therefore, it is a form of error that contaminates the observed measures. The random error, of course, is also a source of contamination reducing the dependability of the observed measures.

The key concern is the degree to which variance due to pairs is systematic. Variance due to persons might be considered a form of error in some contexts and not in other contests -- a distinction that yields different dependability coefficients.

Typically the decision whether to consider variance due to persons as a form of error depends upon whether the persons are considered to be a sample from a universe of possible persons (a random effect) or are considered to be the universe to which generalizations re to be made (a fixed effect). In educational, personnel and psychotherapeutic contexts in which raters evaluate other persons, differences among raters can result in misclassification if different samples of raters are used on different occasions. However, if the same raters are used on all occasions, differences among them will average out and no inequity will result. (Of course if raters vary across occasions, a new facet is introduced that should be accounted for.)

Clearly in most Galileo studies persons are considered to be a sample suggesting that random effects models are appropriate. However, in most Galileo studies, point estimates of the mean difference among concepts may not be of concern even though persons constitute a sample. Most uses of Galileo data depend on shape formed by the positions of the concepts in the multidimesnional spaces rather than on the sizes of the Galileo studies, point estimates of the mean differences among concepts may not be of concern even though person constitute a sample. Most uses of Galileo data depend on shape formed by the positions of the concepts in the multidimensional spaces rather than on the sizes of the spaces. These properties are not affected by simple differences among persons -- that is non person-by-pair interaction differences in which persons assign proportional but different values across pairs. For example, if one person assigns three different pairs the values of 3, 4, and 5, and another person assigns the same respective pairs the values of 6, 8, and 10, both persons responses can be represented as a right

triangle differing only in size. On the other hand, if the second person assigned the values of 6, 7, and 11, this triangle would be of a different shape. In this case, in addition to a person main effect, there would be a person-by-pair interaction that would deflate the dependability coefficient. Therefore, the random effects dependability coefficient, which treats only the residual term as error, is usual appropriate for Galileo data.

O'Brien (1984) provides the formuli for calculating the dependability coefficients from the output of standard analysis of variance programs. The values needed are: (1) the mean square between pairs, BMS: (2) the mean square within raters, RMS; (3) the error mean square, EMS, and (4) the number of raters, n. The fixed effects dependability coefficient is:

$$Df = (BMS - EMS) / BMS \qquad (1)$$

The random effects dependability coefficient is:

$$Dr = (BMS - EMS) / (BMS + (RMS - EMS)/n) \qquad (2)$$

Empirical Examples

Hurd (1985) gathered data concerning perceptions of political information available during the 1984 presidential election campaign. Concepts scaled included respondents characterizations of election news (e.g. mudslinging, poll results, voting records, campaign promises) derived from open-ended interviews with a random sample of adults. Such concepts along with television news and newspapers, vote decisions, and the self were used to construct a Galileo- type questionnaire covering 11 concepts (55 pairs). The questionniares were administered by telephone to a representative sample of 83 Knoxville, Tenn., adults approximately one month before the 1984 election.

Data from these questionnaires were submitted to a standard statistical computer program for a repeated measures analysis of variance.[2] The results of this analysis are shown in Table 1.

Table 1
Repeated Measure Analysis of Variance
Of Perceptions of Political News

Source	DF	Sums of Squares	Mean Squares	F-Ratio*
Model	136	2759.90	20.29	11.47
Error	4329**	7658.99	1.77	
Pairs	54	292.46	5.42	3.06
Raters	82	2467.45	30.09	17.01

*All F values significant at the .0001 level.
**Error degrees of freedom are reduced because of missing values.

Using the appropriate values from Table 1 yields a fixed effects dependability coefficient of .66 and a random effects coefficient

of .63 for Hurd's data. Many researchers would find these levels of reliability marginal, but probably would proceed with analysis interpreting results with caution.

Another empirical example comes from data presented by Fink et al. (1986) in a study of effects of mere exposure on affect. They asked 153 students to estimate the separations between two graphic stimuli, two affective terms, and nine descriptive concepts. Unlike the previous studies, the Fink et al. study was an experiment which used relatively simple stimuli and allowed for highly controlled data collection. The analysis of variance for these data are shown in Table 2.

Table 2
Repeated Measures Analysis of Variance
Of Perceptions of Shapes and Symbols

Source	DF	Sums of Squares	Mean Squares	F-Ratio*
Model	258	1,720,496	6668.59	32.55
Error	13,908**	2,846,058	204.63	
Pairs	77	170,420	2,213.24	10.82
Raters	181	1,550,076	8,563.93	41.86

*All F values significant at the .0001 level.
**Error degrees of freedom are reduced because of missing values.

These data yield a fixed effects dependability coefficient of .91 and a random effects coefficient of .87. These figures indicate that the direct magnitude separation estimation can yield high quality data. The high figures in the Fink et al. data are probably attributable not only to the procedure, but to (1) the relatively large number of raters, (2) the degree of control in the data gathering procedures, and (3) the relative simplicity of the concepts being scaled.

Direct magnitude separation estimates are known to yield skewed data.[3] For example, the Hurd data has a skewness coefficient of 11.64 and a kurtosis coefficient of 281.5. These are extreme figures that clearly affect the variance estimates used in the dependability coefficients. While the Hurd data show moderate dependability in the raw form it was decided to examine what the dependability in the raw from it was decided to examine what the dependability would be if they were transformed to a closer approximation of the normal distribution. When the Hurd data were transformed by taking their natural log, skewness dropped to 2.38 and kurtosis to 10.10. While these are moderately high figures, they represent a substantial improvement over the raw data. The analysis of variance for the log transformed Hurd data are shown in Table 3.

Table 3
Repeated Measures Analysis of Variance
Of Perceptions of Political News
(Log Transformed Data)

Source	DF	Sums of Squares	Mean Squares	F-Ratio*
Model	136	286.35	2.10	18.42
Error	4,329**	494.82	.11	
Pairs	54	51.95	.96	8.42
Raters	82	234.39	2.86	25.81

*All F values significant at the .0001 level.
**Error degrees of freedom are reduced because of missing values.

Both the fixed and random effects dependability coefficients for the transformed Hurd data round to .88 -- a substantial improvement over the figures for the untransformed data of .63 and .66.

The James data had skewness of 7.14 and kurtosis of 73.31 before transformation. These figures fell to skewness, 1.68 and kurtosis, 5.32 after the log transformation. Analysis of variance of the logged James data is shown in Table 4. The fixed effects dependability coefficient derived from this table is .95 and the random effects coefficients is .92 -- again a notable improvement.

Table 4
Repeated Measures Analysis of Variance
Of Separations of College Choice Concepts
(Log Transformed Data)

Source	DF	Sums of Squares	Mean Squares	F-Ratio*
Model	105	191.39	1.80	15.94
Error	2,809**	309.29	.11	
Pairs	54	106.32	1.97	17.21
Raters	51	85.07	1.66	14.58

*All F values significant at the .0001 level.
**Error degrees of freedom are reduced because of missing values.

The Fink et al. data had skewness, 4.75 and kurtosis, 21.64 before transformation, and skewness, -.20 and kurtosis, 1.04 after log transformation. These latter figures generally would be considered to be a close approximation of the normal distribution. The analysis of variance for the Fink et al. data (shown in Table 5) yield a fixed effects coefficient of .97 and a random effects coefficient of .95 indicating very high quality data.

Table 5
Repeated Measures Analysis of Variance
Of Perceptions of Shapes and Symbols
(Log Transformed Data)

Source	DF	Sums of Squares	Mean Squares	F-Ratio*
Model	258	22,399	86.81	48.11
Error	13,908**	25,100	1.80	
Pairs	181	11,743	64.88	36.00
Raters	77	10,656	138.39	76.68

* All F values significant at the .0001 level
** Error degrees of freedom are reduced because of missing values.

Discussion

The above analysis reveals that, even without transformation, direct magnitude separation estimates yield data of sufficient reliability to be acceptable to most communication researchers. When log transformation is applied to the data, the dependability coefficients range from good to excellent by most standards. It should be noted that such levels of reliability are obtained enven though the data are much finer grained than those used in most social research. As Woelfel and Fink (1980) state, direct magitude separation estimates contain far more information than conventional measures because respondents are not confined to a very limited number of rank ordered response categories.

The dependability coefficients have substantial advantages over the techniques used to assess reliability of Galileo data used in prior research. Perhaps the most important of these is that the coefficients are based on a well developed psychometric literature. Thus, the coefficients are conventional, readily interpretable, and should be acceptable to the general community of social science researchers. The dependability coefficients also avoid the problems associated with disregarding available information (variance around the mean) and subjecting the data to mathematical manipulations that may capitalize on chance (rotation of spaces to congruence).

The dependability coefficient makes less arduous demands of the researcher in terms of the amount of data needed to assess data quality. Test-retest correlation procedures double the data collection efforts and run the risk that sensitization effects or that true change will occur in the interval between administrations.

When test-retest data are available, generalizability theory provides a far more detailed method for understanding data qualities than do correlational methods. Introduction of time as an explicit facet in the design provides an additional main effect and three two-way interactions each of which are substantively interpretable. The time main effect is indicative

of the amount of change in the average of the separations (in Galileo terms, whether the overall space is shrinking or expanding). The pair-by-time interaction is indicative of whether the pairs are differentially subject to change. The pair-by-person interaction is perhaps the most interesting because it is indicative of the degree to which individuals have ideosyncratic responses that deviate from group norms. A large person-by-pair-by-time interaction might indicate non homogeniety among the raters and reveal hidden variables that should be taken into consideration. The person-by-pair-by-time interaction is the residual term indicative of random error.

Split sample procedures require a large enough number of cases to allow for groups of adequate size for calculation of reasonable coefficients. Also, a large group could be divided into a numerous split halves that would yield differing coefficients of reliability. Like coefficient alpha, the dependability coefficient is interpretable as the average of all possible split halves and solves this problem.

When samples can be divided into meaningful groups (e.g., men and women, Republicans and Democrats), generalizability theory again provides a more detailed understanding of the data. The group main effect is indicative of overall differences in estimates (whether one group's space in larger than another). The pair-by-group interaction is indicative of differential assessments of differences between the groups. Because persons are nested in groups, there is no person-by-group interaction. The person-by-pair interactions cannot be estimated and remains part of the residual term.

Given the output from standard statistical packages, the dependability coefficients are easily derived. Therefore, calculation and reporting of such coefficients place no major burden on the researcher.

The availability of fixed and random effects dependability coefficients provides the researcher with a more detailed understanding of the data and the choice of a coefficient appropriate to the inferences under consideration. Depending upon whether the researcher is concerned with the shape of the space (when the fixed effects coefficient is appropriate), or the size and shape when the random effects coefficient is appropriate), a correct coefficient can be chosen. It should be noted that past approaches to reliability estimation are roughly equivalent to the fixed effects coefficient because they are based on correlational rather than covariance procedures. Therefore, past coefficients reported have not been appropriate in settings in which the size of the space was under consideration.

The empirical examples reported above demonstrate that there is substantial variation in the dependability of directmagnitude separation estimates. Therefore, it is essential that data quality be assessed for each data gathering effort. The quality of data can not be assumed on the basis that prior research has yielded reliable data.

It is not possible to assess exactly why the dependability coefficients were substantially different in the empirical examples provided. However, several differences between the studies could account for the observed differences. These include the conceptual domain under consideration, the population being sampled, and the interview procedures. Such differences could provide a basis for further research into the quality of data. For example, it would be relatively easy to design a study in which data on the same conceptual domain from equivalent samples were gathered via telephone interviews or via self-administered questionnaires. Such research would introduce a new facet to the analysis variance design, but the procedures for deriving dependability coefficients for such designs are explicated in Cronbach, et al., 1972.

The use of the log tranformation to mitigate skewness in the data results in a substantial improvement in the dependability coefficients. This improvement was highest in data that dealt with the most abstract concepts and had the least rigorous data gathering procedure (i.e., the Hurd data) and was lowest in the data that dealt with the most concrete concepts and had the most rigorous data gathering procedures (i.e., the Fink et al. data). The effects of concept domain and data gathering procedures on the distributional properties of the data need more investigation as do the effects of skewness on the dependability coefficient.

In situations in which researchers have gathered data and found the dependability of their data insufficient, the results of the dependability analysis can be used to estimate the number of persons that would be needed to be interviewed to achieve acceptable dependability levels. This would be done by inserting the dependability coefficient into the Spearman-Brown Prophecy Formula.

In contexts in which the researcher wishes to examine a large number of concepts, it would be necessary to have evaluations of a large number of pairs of concepts. This could make it desirable to mitigate respondent burden by not requiring all respondents to evaluate all possible pairs. This would result in a nested analysis of variance design. Procedures for dealing with such designs are explicated in Cronbach, et al., (1972) and O'Brien (1984).

In summary, generalizability theory provides convenient, rigorous, and general techniques that solve numerous problems associated with evaluation of the quality of direct magnitude separation estimates.

NOTES

1. An earlier version of this paper was presented to the annual meeting of the International Communication Association, 1986. The author is indebted to George A. Barnett, Edward L. Fink, Ralph O'Brien, and John W. Lounsbury for their comments on an earlier draft of this manuscript.

2. The data were processed using the SAS statistical package, version 82,4 on the IBM 360 computer at the University of Tennessee, Knoxville. All distances were arrayed as a single

vector with associated vectors indicating person and pair identification with missing values excluded. The SAS GLM (general linear models) procedures was run with pair distance as the dependent variable and person and pair designations as classification variables. Computer code examples are available from the author upon request.

3. Because the data are skewed, some researchers advocate that the data be transformed prior to submission to the Galileo program for derivation of multidimensional spaces. Generally the raw data are log tansformed; the mean of the logged data is calculated, and the antilog of this mean is submitted to the Galileo program for analysis (Fink, 1979).

REFERENCES

Barnett, G.A. (1972) Reliability and metric multidimensional scaling. Unpublished manuscript, Michigan State University, East Lansing.

Barnett, G.A. (1980) Bibliography of Galileo materials. Paper presented to the International Communication Association, Acapulco, Mexico, May. (Appendix).

Barnett, G.A., Palmer, M.T., & Noor Al-Deen, H. (1984) Translation accuracy using multidimensional scaling. In R. Bostrom (Ed.) Communication Yearbook 6. Beverly Hills: Sage.

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

Cody, M.J. (1980) the validity of experimentally induced motions of public figures in multidimensional scaling configuration. In D. Nimmo (Ed.). Communication Yearbook 4. New Brunswick, N.J.: Transaction Books.

Craig, R.T. (1977) Limiting the scope of the spacial model of communication effects. Human Communication Research, 3. 310-325.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam (1972) The Dependability of Behavioral Measurement: Theory of Generalizability for Score Profiles. New York: Wiley.

Ebel, L. (1951) Estimation of the reliability of ratings. Psychometrica, 16: 407-423.

Fink, E.L. (1979) Data transformation in multidimensional space. Paper presented to the International Communication Association, Philadelphia, May.

Fink, E.L., Monahan, J.L., & Kaplowitz, S.A. (1986) Mere exposure and concept trajectory in a multidimensional space. Paper presented to meeting of the International Communication Association, Chicago.

Gillham, J., & Woelfel, J. (1977) The Galileo system of measurement: preliminary evidence for precision stability, and equivalence to traditional measures. Human Communication Research, 3: 222-234.

Gordon, T.F. (1977) Subject abilities to use metric MDS: effects of varying criterion pair. Paper presented at the Association for Education in Journalism, College Park, Maryland, August.

Hays, W.L. (1981) Statistics. New York: Holt, Rinehart, and Winston.

Horst, P. (1949) A generalized expression for the reliability of measures. Psychometrica, 14:21-31.

Hurd, R.E. (1985) The publics' perceptions of its information needs and the utility of news media content during the 1984 presidential election campaign. Paper presented at the Association for Education in Journalism and Mass Communication, Memphis, August 1985.

James, A.A. (1985) A Method for Identifying Effective Messages for University Recruitment Communication Strategies. Unpublished Masters Thesis, College of Communications, University of Tennessee, Knoxville.

Kincaid, D.L., Yum, J.O., Woelfel, J., & Barnett, G.A. (1983) The cultural convergence of Korean immigrants in Hawaii: an empirical test of a mathematical theory. Quality and Quantity, 18: 59-78.

Marlier, J. (1983) Respondent self concept and responses to scaling tasks. In R. Bostrom (Ed.) Communication Yearbook 7. Beverly Hills: Sage.

O'Brien, R.M. (1984) Using generalizability theory to estimate the dependability of aggregate-level varibles. Quality and Quantity, 18: 193-205.

Serota, K.B., Cody, M.J., Barnett, G.A., & Taylor, J.A. (1977) Precise procedures for optimizing campaign communication. In B. Ruben (Ed.) Communication Yearbook I. New Brunswick, N.J.: Transaction Books.

Strout, P.E., & Fleiss, J.L. (1979) Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86: 420-428.

Winer, B.J. (1971) Statistical Principles in Experimental Design (2nd ed.) New York: McGraw-Hill.

Woelfel, J., & Barnett, G.A. (1982) Multidimensional scaling in Riemann space. Quality and Quantity, 16: 469-491.

Woelfel, J., & Fink, E.L. (1980) The Measurement of Communication Processes: Galileo Theory and Method. New York: Academic Press.