PSYCHOMETRIKA—VOL. 59, NO. 2, PAGES 203–216
JUNE 1994

RELIABILITY AND EXPECTED LOSS: A UNIFYING PRINCIPLE

BRUCE COOIL AND ROLAND T. RUST

OWEN GRADUATE SCHOOL OF MANAGEMENT
VANDERBILT UNIVERSITY

We provide a unified, theoretical basis on which measures of data reliability may be derived or evaluated, for both quantitative and qualitative data. This approach evaluates reliability as the "proportional reduction in loss" (PRL) that is attained in a sample by an optimal estimator. The resulting measure is between 0 and 1, linearly related to expected loss, and provides a direct way of contrasting the measured reliability in the sample with the least reliable and most reliable data-generating cases. The PRL measure is a generalization of many of the commonly-used reliability measures.

We show how the quantitative measures from generalizability theory can be derived as PRL measures (including Cronbach's alpha and measures proposed by Winer). For categorical data, we develop a new measure for the general case in which each of N judges assigns a subject to one of K categories and show that it is equivalent to a measure proposed by Perreault and Leigh for the case where N is 2.

Key words: alpha, kappa, agreement, intercoder reliability, decision rule, generalizability theory, test theory.

1. Introduction

The issue of data reliability is critical in most research settings, and many types of reliability measures are available. For example, these measures are important in the psychometric field of test theory where the data are generally of a quantitative form (i.e., interval or ratio-scaled). Also, many reliability measures are available for categorical data (nominal or ordinal scaled). In each case, a reliability measure is intended to provide an assessment of data accuracy, but the appropriate choice of a measure depends on more than just the type of data available: it also depends in part on the nature of the characteristic measured, on the type of accuracy sought, and what this implies about the optimal way of using the data to estimate its real value.

In selecting a measure of test reliability, the primary goal is to measure the degree to which test questions accurately measure an examinee attribute. In this setting, the choice of a reliability measure typically reflects the fact that the researcher is using an examinee mean score to measure the true value of an attribute, and wishes to distinguish among examinees. In other research settings, one may need to measure the reliability of judges or raters who classify subjects into preselected categories. Here the data are of a categorical nature. Examples include measuring the reliability of coders who classify answers to open-ended interview questions (i.e., intercoder reliability), or even the reliability with which physicians diagnose patients (Shouten, 1986). In these cases, the objective is to measure how accurately judges (coders) classify (code) subjects (data items). The degree to which coders agree is often referred to as "intercoder

Bruce Cooil is an Associate Professor of Statistics, and Roland T. Rust is a Professor and area head for Marketing. The authors thank three anonymous reviewers and an Associate Editor for their helpful comments and suggestions. This work was supported in part by the Dean's Fund for Faculty Research of the Owen Graduate School of Management, Vanderbilt University.

Requests for reprints should be sent to Bruce Cooil, Owen Graduate School of Management, Vanderbilt University, 401 21st Avenue South, Nashville, TN 37203.

0033-3123/94/0600-92079$00.75/0

reliability'', and is implicitly a proxy for the degree to which a coder consensus (plurality) is repeatable, and (by extension) the degree to which it is likely to be correct. Thus, reliability measures for categorical data are used to infer how good the consensus judgments are likely to be, and thus how much an analysis based on the data can be trusted.

We propose that proportional reduction in loss (PRL) be used as a consistent theoretical basis to derive, justify, and interpret reliability measures for any possible measurement scenario, whether the data are quantitative or categorical. A loss function not only provides a way of defining accuracy, but it also forces the researcher to consider the optimal way of constructing an estimator from the data. By measuring reliability as the proportional reduction of loss, one has a natural was of gauging reliability on a zero-to-one scale, depending on the reduction in loss that has occurred relative to the least reliable data-generating situation. For quantitative reliability measures, the main advantage of this approach is that it simplifies the interpretation of existing measures (e.g., generalizability-theory measures in general, including Cronbach's alpha). In the case of categorical data, the results are even more interesting: The PRL approach leads to a new general measure that is not equivalent to any existing measure when there are more than two judges.

Section 2 describes reliability from an expected loss perspective, and presents our general approach to constructing reliability measures. Section 3 is a brief discussion of existing quantitative reliability measures in the test-theory context, and shows how they are PRL measures. Section 4 discusses categorical measures, and a general reliability measure is derived for a commonly-encountered data-generating scenario. Section 5 concludes and summarized.

2. Reliability and Expected Loss

We begin by assuming that there is a true (unknown) value for an examinee or subject. It is either a quantitative level (e.g., the true value of an attribute that we attempt to measure by testing) or a category (e.g., the correct diagnosis of a patient). We refer to the true value for examinee or subject i as θ_i. Based on the data, we also obtain the estimated value $\hat{\theta}_i$. In a testing framework, $\hat{\theta}_i$ will generally represent the mean score of an examinee. If the data are categorical, $\hat{\theta}_i$ will generally represent the modal classification (of a subject) that is selected by a group of judges. We assume that the loss, $L_i \equiv L(\theta_i, \hat{\theta}_i)$, is a function of θ_i, and $\hat{\theta}_i$, which is always nonnegative and equal to zero if $\hat{\theta}_i = \theta_i$.

We further assume that the purpose of a reliability measure is to reflect expected loss. That is, when $\hat{\theta}_i$ is perfectly reliable, the expected loss should be at its minimum, and the corresponding reliability measure should be at its maximum. This reflects how researchers generally think about reliability measures, at least implicitly. We base our reliability measure on the explicit decision theoretic construct of expected loss, $E[L_i]$.

In a testing framework where $\hat{\theta}_i$ is an examinee mean score, squared error loss is usually an appropriate measure of accuracy because the test theory models (e.g., see (4)) are linear functions of parameters that are defined strictly in terms of their means and variances. Even without assuming that the error terms in these models have normal distributions, the examinee mean score is an optimal estimator in the sense that it has the minimum variance among unbiased linear estimators.

In the categorical case, $\hat{\theta}_i$ is the estimate of the correct category. A natural choice for loss is an indicator function that takes on the value 1 if $\hat{\theta}_i$ is an incorrect category and 0 if it is the correct category. Expected loss is then simply the probability that $\hat{\theta}_i$ is not the correct category. We argue that $\hat{\theta}_i$ should be chosen to minimize this prob-

ability and that the corresponding reliability measure should measure the performance of $\hat{\theta}_i$ in terms of this probability.

To normalize the reliability measure to the zero-one range, we construct a "proportional reduction in loss" (PRL) measure. The PRL that occurs when $\hat{\theta}_i$ is used to estimate θ_i is defined as:

$$\text{PRL}(\hat{\theta}_i) = \frac{\{E_{\text{SUP}}[L_i] - E[L_i]\}}{\{E_{\text{SUP}}[L_i] - E_{\text{INF}}[L_i]\}}, \tag{1}$$

where $E_{\text{SUP}}[L_i]$ and $E_{\text{INF}}[L_i]$ are the lowest upper bound and the greatest lower bound, respectively, of the expected loss that can occur when the estimator $\hat{\theta}_i$ is used to predict or estimate θ_i. As illustrated in sections 3 and 4, these bounds are taken across the appropriate family of distributions for $\hat{\theta}_i$ and in many cases $E_{\text{INF}}[L_i] = 0$. Thus, when the expected loss is as large as possible, PRL = 0, and when the expected loss is as small as possible, PRL = 1. Also, the range between PRL = 0 and PRL = 1 is of constant scale (i.e., it does not depend on the units of measurement), and is proportional to changes in expected loss. We will see how this conceptual framework is fully general, and subsumes several prominent reliability measures as special cases.

In our discussion of quantitative reliability measures (section 3) we will consider test designs where N test items are used to measure the true scores of M examinees. When considering categorical reliability measures (section 4), we will refer to the generic situation where N judges classify each of M subjects. In either context, θ_i represents the "true" code or value of examinee or subject i, $1 \leq i \leq M$, and the data are used to construct an estimate $\hat{\theta}_i$. To evaluate the overall reliability of a test that measures the true scores of M examinees, or the degree to which judges reliably classify M subjects, we use the across-examinee/subject PRL of the estimates $\hat{\boldsymbol{\theta}} \equiv (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_M)$,

$$\begin{aligned}\text{PRL}(\hat{\boldsymbol{\theta}}) &\equiv \frac{\sum_{i=1}^{M} \{E_{\text{SUP}}[L_i] - E[L_i]\}}{\sum_{i=1}^{M} \{E_{\text{SUP}}[L_i] - E_{\text{INF}}[L_i]\}} \\ &= \frac{M^{-1} \sum_{i=1}^{M} \{E_{\text{SUP}}[L_i] - E[L_i]\}}{M^{-1} \sum_{i=1}^{M} \{E_{\text{SUP}}[L_i] - E_{\text{INF}}[L_i]\}},\end{aligned} \tag{2}$$

which can be thought of interchangeably as the proportional reduction in total loss, or the proportional reduction in average loss, across M examinees/subjects. Measures have been proposed that average pairwise or multiple-judge indices of reliability across all judge subgroups of a given size (e.g., Conger, 1980). In contrast, the measure in (2) represents a single reliability index for the entire group of N judges.

This across-examinee/subject PRL is a reliability measure that reflects the reliability of the data, through the specific choice of the estimator $\hat{\theta}_i$. The estimator $\hat{\theta}_i$ is simply a decision rule that summarizes how the data should be used to measure the quantity or attribute θ_i. Different reliability measures arise by considering different

decision rules. We are of course interested primarily in the reliability measures that correspond to optimal decision rules, where optimality depends on the type of data (quantitative or qualitative) and on the experimental design. We argue that the most appropriate reliability measures correspond to these optimal decision rules. Generally we will be considering balanced designs where the same basic model and decision rule are used for each examinee/subject. In this case, the PRL for examinee/subject i, $\text{PRL}(\hat{\theta}_i)$, is the same for all examinees or subjects i, $1 \leq i \leq M$, and is therefore the same as the across-examinee/subject PRL defined in (2).

The general approach of proportional reduction in loss is similar to other measures in various domains of science, including most notably the concept of "proportional reduction in error" (PRE) in psychometrics. Costner (1965) argues that the PRE framework provides a useful way of delineating the differences among various measures of association and by so doing, simplifies the interpretation of various measures. Hildebrand, Laing, and Rosenthal (1977, pp. 39–40) also show how prediction rules can be developed and studied in terms of PRE. We argue that similar advantages apply when interpreting and developing reliability measures as proportional reduction in loss measures.

3. Quantitative Measures

Generalizability theory (Cronbach, Gleser, Nanda & Rajaratnam, 1972) already provides a unifying framework for most quantitative reliability measures. In this section we briefly review generalizability theory and show how it relates directly to the PRL paradigm. We do this by looking explicitly at three reliability measures in a test-theory context, beginning with Winer's reliability measure for the mean in a one-way analysis of variance design. We then conclude this section by looking at the PRL interpretation of two measures (including Cronbach's alpha) in a two-way ANOVA framework.

An advantage of the PRL approach is that it provides a direct way of contrasting the measured reliability in the sample with the least reliable and most reliable data-generating situations. In this way the PRL framework shows explicitly how each reliability measure gauges the accuracy (in terms of expected loss) of the data in a specific decision-making context. Also, it provides a simple and important way of thinking about each measure, and therefore simplifies the choice and clarifies the interpretation of the appropriate measure for a specific research design.

Generalizability Theory and the Proportional Reduction in Loss

Consider the general situation where each of a group of examinees answers a fixed number of test-items. The examinee mean scores are reliable only to the degree to which they differentiate among the true scores of the examinees. Generalizability theory (G-theory) examines how the variance of the observed examinee mean scores can be partitioned into two components, a component due to the variance among examinees (the differentiation variance) and a component of absolute error variance,

$$\text{Variance of mean measurement} = [\text{differentiation variance}] + [\text{absolute error variance}].$$

How this is done depends on the specific experimental design. In some designs, absolute error variance can be further decomposed into a relative component of error that does not include certain biases in the mean measurement,

$$\text{absolute error variance} = [\text{relative error variance}] + [\text{remaining error variance}].$$

G-theory posits that an appropriate measure of reliability (or dependability) is simply the ratio of the differentiation variance to the sum of the differentiation variance and an appropriate component of error variance (either relative or absolute),

$$\text{Reliability Measure (or Dependability Index)} = \frac{[\text{differentiation variance}]}{\{[\text{differentiation variance}] + [\text{absolute or relative error variance}]\}}. \quad (3)$$

Relative error is used when part of the absolute error is not deemed important in a particular application (the measure G_2^*, described below, provides an example). When absolute error is used in (3), the denominator is the total variance of the mean examinee measurement, and the reliability of this mean is simply the proportion of this total variance that is due to the differentiation variance (the variance among examinee true values). The formula in (3) can also be used to assess the average reliability of a single test-item by defining the components of differentiation and error variance in formula (3) as those components present in a single test-item measurement.

As a specific example, consider the case where each of M examinees answers N test items. If we assume that the N test items are not always the same for each examinee, then it is appropriate to use a one-way random effects ANOVA design (i.e., a randomized block design). If y_{ij} represents the ith examinee's score on item j, then y_{ij} is the sum of two components: the true score for examinee i, θ_i, and the within item random error $\varepsilon_{ij}(\theta)$,

$$y_{ij} = \theta_i + \varepsilon_{ij}(\theta), \qquad 1 \le i \le M, \qquad 1 \le j \le N. \quad (4)$$

Here we assume that the true score for examinee i, θ_i, is itself a random variable with mean μ_θ and variance σ_θ^2, while the random errors $\varepsilon_{ij}(\theta)$ have mean zero and variance $\sigma^2_{\varepsilon(\theta)}$. Equation (4) implies that the mean score for examinee i, $\bar{y}_i$, is the true examinee score θ_i, up to an error term $\bar{\varepsilon}_i$, $\bar{y}_i = \theta_i + \bar{\varepsilon}_i$. To measure the reliability of $\bar{y}_i$ as an estimate of θ_i, one uses the ratio of the random-effect variance of the examinee true scores θ_i, to the total variance of the mean examinee score (Winer, 1971, p. 285):

$$G_1 \equiv \frac{\sigma_\theta^2}{\left[\sigma_\theta^2 + \left(\frac{\sigma^2_{\varepsilon(\theta)}}{N}\right)\right]}. \quad (5)$$

G_1 is a specific example of G-theory formula of (3), where the random-effect variance, σ_θ^2, is the differentiation variance (the variance among examinee true scores), and absolute error variance, $\sigma^2_{\varepsilon(\theta)}/N$, is used as part of the denominator.

Now consider the proportional reduction in loss interpretation of G_1. In a linear model like (4), squared error loss provides a natural way of evaluating the accuracy of an estimator for θ_i, and the optimal unbiased estimator with respect to this loss function is $\bar{y}_i$. So it is appropriate to consider the proportional reduction in the squared error loss function evaluated at $\bar{y}_i$,

$$L_i \equiv (\bar{y}_i - \theta_i)^2.$$

Equation (4) implies that the mean evaluation of examinee i, $\bar{y}_i$, is the true score up to an error term $\bar{\varepsilon}_i$, $\bar{y}_i = \theta_i + \bar{\varepsilon}_i$, so that the expected loss that occurs when $\bar{y}_i$ is used to estimate θ_i is simply the variance of the average error $\bar{\varepsilon}_i$, or $\sigma^2_{\varepsilon(\theta)}/N$:

$$E[L_i] = \frac{\sigma^2_{\varepsilon(\theta)}}{N}. \tag{6}$$

The test-items are least reliable when they fail to differentiate among the θ_i (the true scores of the examinees). Using model (4), this occurs when all of the variance of the test-item measurements y_{ij} is due to the error term $\varepsilon_{ij}(\theta)$. Thus, if we let $y^\dagger_{ij}$ represent the least reliable test-item measurements, the $y^\dagger_{ij}$ are generated from the model

$$y^\dagger_{ij} = \mu_\theta + \varepsilon_{ij}(\theta). \tag{7}$$

In this case the mean score for each examinee is $\bar{y}^\dagger_i = \mu_\theta + \bar{\varepsilon}_i$ and the expected maximum loss of $\bar{y}^\dagger_i$, written $E_{\text{SUP}}[L_i]$, is the expectation of L_i with respect to model (7):

$$E_{\text{SUP}}[L_i] = E[(\bar{y}^\dagger_i - \theta_i)^2] = E[(\mu_\theta + \bar{\varepsilon}_i - \theta_i)^2]$$

$$= \sigma^2_\theta + \left(\frac{\sigma^2_{\varepsilon(\theta)}}{N}\right). \tag{8}$$

By (5), (6) and (8), $G_1 = (E_{\text{SUP}}[L_i] - E[L_i])/E_{\text{SUP}}[L_i]$ (and $E_{\text{INF}}[L_i] = 0$), so that G_1 is simply the PRL of the ith examinee's mean score, $\bar{y}_i$, that occurs when the mean score, of the ith examinee, is generated from model (4) instead of from model (7). Model (7) represents the least reliable limiting case of (4) where test-items are not able to differentiate among the true scores of the examinees, but do generate unbiased measurements with the same total variance as in (4).

By interpreting G_1 as a PRL, we are contrasting the two measurements $\bar{y}_i$ and $y^\dagger_i$:

$$\bar{y}_i = \theta_i + \bar{\varepsilon}_i \text{ versus } \bar{y}^\dagger_i = \mu_\theta + \bar{\varepsilon}_i.$$

The estimator $\bar{y}^\dagger_i$ is the least reliable limiting case of $\bar{y}_i$ because it does not differentiate among the true scores θ_i. On the other hand, $\bar{y}_i$ becomes a more accurate estimator for θ_i as either the number of items, N, increases or as the error variance $\sigma^2_{\varepsilon(\theta)}$ decreases. Consequently, $G_1 \to 1$ as either $N \to \infty$ or as $\sigma^2_{\varepsilon(\theta)} \to 0$.

The PRL Interpretation of Other Quantitative Measures

The G-theory formulation (3) relates directly to a general proportional reduction in loss interpretation for quantitative reliability measures. The differentiation variance in the numerator represents the reduction in expected loss that occurs when we use a specific mean score $\bar{y}_i$ to measure the true value θ_i of examinee i, while the denominator of the generic measure in (3) represents the maximum loss that occurs when measurements do not differentiate among examinees. These measures differ only because the decision context, in which $\hat{\theta}_i = \bar{y}_i$ is evaluated, differs in each case.

Now consider the two-way design where each examinee answers the same N test-items. If y_{ij} represents the ith examinee's score on item j, then y_{ij} is the sum of three components: the true value for examinee i, θ_i, the component of error due to test item j, δ_j, and the remaining random error $\varepsilon_{ij}(\theta)$,

$$y_{ij} = \theta_i + \delta_j + \varepsilon_{ij}, \qquad 1 \le i \le M, \qquad 1 \le j \le N. \tag{9}$$

As in the one-way model, the true value for examinee i, θ_i, is itself a random variable with mean μ_θ and variance σ^2_θ. Also, the component of error due to test item j, δ_j, is

random with mean zero and variance σ_δ^2, while the random errors $\varepsilon_{ij}(\theta)$ have mean zero and variance $\sigma_{\varepsilon(\theta \times \delta)}^2$. In this setting, the reliability measure G_2 is defined as (Cronbach et al., 1972),

$$G_2 \equiv \frac{\sigma_\theta^2}{\left[\sigma_\theta^2 + \left(\frac{[\sigma_\delta^2 + \sigma_{\varepsilon(\theta \times \delta)}^2]}{N}\right)\right]}. \tag{10}$$

As in the case of G_1, the denominator represents the total variance of the mean score $\bar{y}_i$, so that the two measures are algebraically equivalent. In this design, the mean measurement is of the form

$$\bar{y}_i = \theta_i + \bar{\delta} + \bar{\varepsilon}_i, \qquad 1 \le i \le M,$$

so that the expected squared-error loss from using the examinee mean score $\bar{y}_i$ to estimate the true value θ_i, is

$$E[L_i] = \frac{[\sigma_\delta^2 + \sigma_{\varepsilon(\theta \times \delta)}^2]}{N}. \tag{11}$$

The expected maximum loss, $E_{\text{SUP}}[L_i]$, is the expectation of L_i with respect to the model

$$y_{ij}^\dagger = \mu_\theta + \delta_j + \varepsilon_{ij}, \qquad 1 \le i \le M, \qquad 1 \le j \le N.$$

For the 2-way design, this represents the worst case scenario in which test items are completely unable to differentiate among the examinees, but instead simply generate unbiased measurements with the same total variance. Thus the mean measurement would be of the form

$$\bar{y}_i^\dagger = \mu_\theta + \bar{\delta} + \bar{\varepsilon}_i, \qquad 1 \le i \le M,$$

so that

$$E_{\text{SUP}}[L_i] = E[(\bar{y}_i^\dagger - \theta_i)^2] = E[(\mu_\theta + \bar{\delta} + \bar{\varepsilon}_i - \theta_i)^2]$$

$$= \sigma_\theta^2 + \left(\frac{[\sigma_\delta^2 + \sigma_{\varepsilon(\theta \times \delta)}^2]}{N}\right). \tag{12}$$

The PRL derivation of G_2 now follows from (11) and (12) (here $E_{\text{INF}}[L_i] = 0$).

The advantage of the 2-way design (9) is that each of the estimators $\bar{y}_i$ has the same component of average test-item error $\bar{\delta}$,

$$\bar{\delta} \equiv \sum_{j=1}^{N} \frac{\delta_j}{N}.$$

Therefore, G_2 underestimates the actual reliability of the $\bar{y}_i$ in any framework where the constant stochastic bias $\bar{\delta}$ is not important. This would be true if the $\bar{y}_i$ were used in any location-invariant analysis, including covariance or correlation statistics, a factor analysis based on covariances or correlations, or rank-order statistics. In these settings, it is appropriate to use G_2^*, the reliability measure with respect to relative error (Winer, 1971, p. 289):

$$G_2^* \equiv \frac{\sigma_\theta^2}{\left[\sigma_\theta^2 + \left(\frac{\sigma_{\varepsilon(\theta \times \delta)}^2}{N}\right)\right]}. \tag{13}$$

G_2^* is the reliability measure estimated by Cronbach's alpha (Brennan, 1983, p. 18; Cronbach, 1951) and has also been referred to by Winer (1971, p. 289) as the reliability of the mean that "adjusts for anchor points" (i.e., it adjusts for the constant average test-item error $\bar{\delta}$). The PRL interpretation of G_2^* is virtually the same as that of G_2. The only difference is that we now define proportionate reduction in loss with respect to the nonzero minimum loss,

$$E_{\text{INF}}[L_i] = \frac{\sigma_\delta^2}{N},$$

that occurs when mean measurements are completely accurate except for the average test-item error $\bar{\delta}$,

$$\bar{y}_i = \theta_i + \bar{\delta}, \qquad 1 \leq i \leq M.$$

Then G_2^* of (13) is simply the PRL measure,

$$G_2^* = \frac{\{E_{\text{SUP}}[L_i] - E[L_i]\}}{\{E_{\text{SUP}}[L_i] - E_{\text{INF}}[L_i]\}},$$

where $E[L_i]$ and $E_{\text{SUP}}[L_i]$ are defined in (11) and (12) as they were for G_2.

The Equivalence of PRL When Measured Across Examinees

Each of these reliability measures assumes a balanced experimental design where the expected loss per examinee is constant across examinees so that for any examinee i,

$$E[L_i] = E\left[M^{-1} \sum_{i=1}^{M} L_i\right], \tag{14}$$

and

$$E_{\text{INF}}[L_i] = E_{\text{INF}}\left[M^{-1} \sum_{i=1}^{M} L_i\right]. \tag{15}$$

The expected loss that occurs when items do not differentiate among examinees is also constant across examinees,

$$E_{\text{SUP}}[L_i] = E_{\text{SUP}}\left[M^{-1} \sum_{i=1}^{M} L_i\right]. \tag{16}$$

Consequently, by (14), (15) and (16), the PRL for examinee i, PRL($\bar{y}_i$), is equal to the total PRL across all M examinees, PRL($\bar{\mathbf{y}}$), as defined in (2), and each of the measures G_1, G_2, and G_2^* can be interpreted as either the PRL per examinee or the total PRL.

4. Categorical Measures

The PRL paradigm also provides a natural way of developing and comparing reliability measures when the data are of a categorical nature. In this section we begin with a brief discussion of three reliability measures that have been proposed for categorical data. We then introduce a general decision theoretic framework for evaluating categorical reliability measures. Using this framework, we present what we argue is the most natural measure of reliability ("PRL reliability", written Δ_{PRL}) and show that it is equivalent to a measure proposed by Perreault and Leigh (1989) when there are only 2 judges.

All of the basic categorical measures were originally developed for the case where there are 2 judges ($N = 2$). If we assume that these judges classify each of M subjects into one of K mutually exclusive categories, the simplest measure of reliability is to calculate the proportion of times (across the M subjects) that two judges agree that a subject should be classified into the same category. Although proportional agreement has a certain intuitive appeal as a reliability measure, it is clearly a function, at least in part, of the number of possible categories K, since a certain amount of agreement would occur by chance even when judges classify each subject randomly. Cohen (1960, 1968) developed the kappa index to correct for the degree of random agreement:

$$\kappa \equiv \frac{\left[\sum_{k=1}^{K} f(k, k) - \sum_{k=1}^{K} f(\cdot, k) f(k, \cdot)\right]}{\left[1 - \sum_{k=1}^{K} f(\cdot, k) f(k, \cdot)\right]}, \tag{17}$$

where

$f(k, k)$ = the proportion of subjects that both judges put into category k,

$$f(\cdot, k) \equiv \sum_{j=1}^{K} f(j, k), \qquad f(k, \cdot) \equiv \sum_{j=1}^{K} f(k, j),$$

so that $f(\cdot, k)$ and $f(k, \cdot)$ represent the proportion of times that Judge 1 and Judge 2 classify a subject to category k, respectively. Note that $\sum_{k=1}^{K} f(k, k)$ represents the proportional agreement between judges, and if judges make their classifications independently, $\sum_{k=1}^{K} f(\cdot, k) f(k, \cdot)$ represents the amount of agreement that would occur by chance. Thus, kappa is the ratio of *nonrandom* agreement that occurs, $\sum_{k=1}^{K} f(k, k) - \sum_{k=1}^{K} f(\cdot, k) f(k, \cdot)$, relative to the maximum possible nonrandom agreement, $1 - \sum_{k=1}^{K} f(\cdot, k) f(k, \cdot)$. Of course, Cohen (1960, p. 39) proposed kappa as a "coefficient of agreement" rather than as a reliability measure per se.

As an alternative approach, Perreault and Leigh (1989) propose that reliability actually be defined as the probability with which two judges make "reliable judgments", and they show how one can estimate this probability from the observed level of agreement between two judges. If each judge makes reliable judgments with probability Δ_I, and if judges act independently, then one would expect both judges to agree on a total of $M\Delta_I^2$ reliable classifications (since a total of M classifications are made by each judge). Among the other $M(1-\Delta_I^2)$ classifications, at least one judge has made an unreliable classification, and yet in each of these cases the two judges will still agree by chance with probability $1/K$ (because there are K categories) and so there will be

another $M(1-\Delta_I^2)\times(1/K)$ classifications on which both judges agree. Consequently, one would expect the total number of agreements among M joint classifications, $M \times E[\sum_{k=1}^{K} f(k, k)]$, to be:

$$ME\left[\sum_{k=1}^{K} f(k, k)\right] = M\Delta_I^2 + M(1 - \Delta_I^2)\left(\frac{1}{K}\right),$$

which implies

$$\Delta_I \equiv \left\{\frac{\left[E\left[\sum_{k=1}^{K} f(k, k)\right] - K^{-1}\right]}{[1 - K^{-1}]}\right\}^{1/2}, \tag{18}$$

whenever $E[\sum_{k=1}^{K} f(k, k)] > 1/K$ (otherwise $\Delta_I \equiv 0$). On this basis, a reasonable sample estimate of the proportion of reliable classifications made by each of 2 judges becomes

$$\hat{\Delta}_I \equiv \left\{\frac{\left[\sum_{k=1}^{K} f(k, k) - K^{-1}\right]}{[1 - K^{-1}]}\right\}^{1/2}, \quad \text{if } \sum_{k=1}^{K} f(k, k) > \frac{1}{K},$$

$$\equiv 0, \text{ otherwise.} \tag{19}$$

Note that the reliability measure Δ_I is based on the assumption that judge classifications are independent and equally reliable. This in turn implies that the judges make classifications to each category with the same expected frequencies, that $E[f(\cdot, k)]=E[f(k, \cdot)]$, $1 \leq k \leq K$. In contrast, Cohen's κ does provide a measure of agreement under more general conditions, although κ can only achieve the value of 1 when the observed marginal frequencies are equal (Cohen, 1960, p. 43).

A Decision Theoretic Framework for Categorical Measures

We now consider a framework that allows us to develop a general PRL measure for the case where there are an arbitrary number of judges, $N \geq 2$, and K categories. In classifying each subject, we will assume that each judge acts independently, chooses the correct category with probability $p \geq 1/K$, and makes incorrect classifications randomly to each of the other $K - 1$ categories with probability $(1 - p)/(K - 1)$. Let θ_i represent the correct category for subject i, and let $x_i(k)$, $1 \leq k \leq K$, represent the category counts, that is

$$x_i(k) = \text{the number of judges that classify subject } i \text{ into category } k.$$

Then under these assumptions the $x_i(k)$ are realizations of a multinomial random vector $\mathbf{x}_i \equiv (x_i(1), \ldots, x_i(K))$ with probability function

$$P[X_i(k) = x_i(k), \quad 1 \leq k \leq K | \theta_i]$$

$$= \left[\frac{N!}{\left(\prod_{k=1}^{K} x_i(k)!\right)}\right] p^{n_i^*}\left[\frac{(1 - p)}{(K - 1)}\right]^{[N - n_i^*]}, \tag{20}$$

where $n_i^* \equiv x_i(\theta_i)$ represents the count of the correct category θ_i. Here we are assuming that $p \geq 1/K$ is an unknown constant and that the correct categories θ_i, $1 \leq i \leq M$, are random effects with the probability distribution

$$P[\theta_i = k] \equiv r_k, \qquad 1 \leq k \leq K. \tag{21}$$

In this way the θ_i are direct analogues of the treatment effects used to interpret the quantitative reliability measures (see (4) and (9)). The probability function in (21) provides a way of modeling the specific realizations of θ_i, $1 \leq i \leq M$, as a sample from an arbitrary and completely general distribution. This generality also allows the investigator to incorporate prior information on the relative frequencies of the K categories when it exists.

For an arbitrary estimator $\hat{\theta}_i$ of the correct category θ_i, a reasonable loss function would be:

$$\begin{aligned} L_i \equiv L(\theta_i, \hat{\theta}_i) &= 1, \qquad \hat{\theta}_i \neq \theta_i \\ &= 0, \qquad \hat{\theta}_i = \theta_i, \end{aligned} \tag{22}$$

so that the expected loss associated with θ_i is simply the probability that $\hat{\theta}_i \neq \theta_i$,

$$E_p[L_i] = P[\hat{\theta}_i \neq \theta_i] = 1 - P[\hat{\theta}_i = \theta_i]. \tag{23}$$

Here both the expectation and probability calculations are performed with respect to the predictive distribution of $\mathbf{X}_i \equiv (X_i(1), \cdots, X_i(K))$,

$$\begin{aligned} P[\mathbf{X}_i = \mathbf{x}_i] &= \sum_{k=1}^{K} P[\mathbf{X}_i = \mathbf{x}_i | \theta_i = k]\, P[\theta_i = k] \\ &= \sum_{k=1}^{K} P[\mathbf{X}_i = x_i | \theta_i = k]\, r_k, \end{aligned} \tag{24}$$

where we use the likelihood function (20) with the prior distribution (21). The predictive expected loss in (23) is sometimes referred to as Bayes risk. (Implicitly, the expected loss derivations in section 3 were also with respect to the predictive distribution of the data). If judges are unable to differentiate among subjects and are equally likely to classify a given subject i into any one of the K categories, the category frequencies $x_i(k)$ are generated from the multinomial distribution (20) with $p = 1/K$, and the maximum expected loss becomes:

$$E_{\text{SUP}}[L_i] \equiv E_{p=1/K}[L_i] = 1 - P\left[\hat{\theta}_i = \theta_i; \quad p = \frac{1}{K}\right]. \tag{25}$$

Thus, by (1) and (25), the PRL that occurs when we predict that $\hat{\theta}_i$ is the correct category is

$$\begin{aligned} \text{PRL}(\hat{\theta}_i) &= \frac{\{E_{\text{SUP}}[L_i] - E[L_i]\}}{E_{\text{SUP}}[L_i]} \\ &= \frac{\left\{P[\hat{\theta}_i = \theta_i; p] - P\left[\hat{\theta}_i = \theta_i; p = \frac{1}{K}\right]\right\}}{\left\{1 - P\left[\hat{\theta}_i = \theta_i; p = \frac{1}{K}\right]\right\}}. \end{aligned} \tag{26}$$

(Here $E_{\text{INF}}[L_i] = E_{p=1}[L_i] = 0$.)

An obvious estimator for θ_i is the category most frequently selected by the N judges. (This is, in fact, what practicing researchers virtually always choose.) Denote this modal category as MODE_i. If there is more than one such modal category, it will be defined to represent a random selection from among the most frequently selected categories (a slightly modified definition of MODE_i that incorporates estimates of the prior probabilities r_k, $1 \leq k \leq K$, is also possible). Whenever $p \geq 1/K$, MODE_i is the maximum likelihood estimator (MLE) of θ_i (and it is the unique MLE whenever there is a unique mode). If $p \geq 1/K$ and we have a uniform discrete prior in (21) with $r_k = 1/K$ $(1 \leq k \leq K)$, MODE_i also minimizes the expected loss function of (23) and is the Bayes estimator. In fact, MODE_i also minimizes expected loss under more general conditions. The corresponding decision-theoretic reliability measure, Δ_{PRL}, is simply the proportional reduction in the loss (i.e., the probability of making the wrong choice among categories) that occurs when one selects $\hat{\theta}_i = \text{MODE}_i$ (a category selected by the largest number of judges) relative to the worst case scenario where $p = 1/K$ and judges are randomly classifying subject i. We are implicitly assuming that the likelihood function (20) and the prior distribution (21) remain constant across subjects so that the average across-subject reliability is the same as the reliability for each individual subject i. Thus, we will define the across-subject reliability Δ_{PRL} to be the reliability of MODE_i as an estimator of θ_i, $1 \leq i \leq M$, and using (26) this becomes

$$\Delta_{\text{PRL}} \equiv \text{PRL}(\text{MODE}_i)$$

$$= \frac{\left\{P[\text{MODE}_i = \theta_i; p] - P\left[\text{MODE}_i = \theta_i; p = \frac{1}{K}\right]\right\}}{\left\{1 - P\left[\text{MODE}_i = \theta_i; p = \frac{1}{K}\right]\right\}}$$

$$= \frac{[P[\text{MODE}_i = \theta_i; p] - K^{-1}]}{[1 - K^{-1}]}, \qquad p \geq K^{-1} \tag{27}$$

The last equality follows because $P[\text{MODE}_i = \theta_i; p = 1/K] = 1/K$, even for the general prior distribution (21), since a judge is equally likely to choose any category when $p = 1/K$, so that MODE_i will be the correct category with only a probability of $1/K$. If each individual judge selects category θ_i with a probability smaller than $1/K$, then we define Δ_{PRL} as 0,

$$\Delta_{\text{PRL}} \equiv 0, \qquad p < K^{-1}. \tag{28}$$

If there were reason to believe that p, the individual judge reliability, was not constant across subjects (or that some other component of the likelihood function (20) or the prior distribution (21) varied across subjects), then Δ_{PRL} would be defined as in (27) with $M^{-1} \sum_{i=1}^{M} P[\text{MODE}_i = \theta_i; p]$ used in place of $P[\text{MODE}_i = \theta_i; p]$.

Equivalence of the Perreault and Leigh Measure Δ_I and Δ_{PRL} when $N = 2$

Whenever there are 2 judges and an arbitrary number of categories K, we will show that the Perreault and Leigh measure Δ_I in (18) is equivalent to the proposed measure Δ_{PRL} in (27). If we consider the general multinomial model of (20) with $N = 2$, then MODE_i will be the correct category θ_i with probability,

$$P[\text{MODE}_i = \theta_i]$$

$$= P[\text{both judges select } \theta_i] + 0.5 \cdot [\text{only one judge selects } \theta_i]$$

$$= p^2 + 0.5 \cdot [2p(1-p)] = p,$$

because there are only two ways MODE_i can be the correct category θ_i: either both judges make the correct choice (which happens with probability p^2) or only one judge makes the correct choice (this happens with probability $2p(1-p)$) *and* MODE_i is then randomly assigned to the correct category (which happens with probability 0.5 since θ_i is one of the 2 modes). It follows from (27) that

$$\Delta_{\text{PRL}} = \frac{(p - K^{-1})}{(1 - K^{-1})} = \frac{(Kp - 1)}{(K - 1)}, \qquad N = 2. \tag{29}$$

On the other hand, by the definition of $f(k, k)$ following (17) and using the multinomial model (20), when $N = 2$,

$$E\left[\sum_{k=1}^{K} f(k, k)\right] = \sum_{k=1}^{K} (M)^{-1} \sum_{i=1}^{M} P[X_i(k) = 2] = M^{-1} \sum_{i=1}^{M} \sum_{k=1}^{K} P[X_i(k) = 2]$$

$$= \sum_{k=1}^{K} P[X_i(k) = 2] = p^2 + \frac{(1-p)^2}{(K-1)}.$$

Using this last result in (18), we have

$$\Delta_I = \left\{ \frac{\left[p^2 + \dfrac{(1-p)^2}{(K-1)} - K^{-1} \right]}{[1 - K^{-1}]} \right\}^{1/2} = \frac{(Kp - 1)}{(K - 1)}.$$

By comparing this result with (29), we see that Δ_I and Δ_{PRL} are equivalent whenever there are only 2 judges, and they are both linear in p.

Perreault and Leigh (1989, p. 140) refer to Δ_I as the frequency with which each judge makes "reliable judgments." Since $\Delta_I = \Delta_{\text{PRL}}$ (when $N = 2$), we see that when there are only two judges, each measure may be interpreted as the proportional reduction in the probability of incorrect classification that occurs when each judge makes correct classifications with probability p, rather than with probability $1/K$. Alternatively, each measure also represents the proportional reduction in the probability that MODE_i is not the correct category θ, when judges make correct classifications with probability $p \geq 1/K$, rather than with probability $1/K$.

5. Discussion

The proportional reduction in loss (PRL) approach provides a general theoretical basis for the derivation and evaluation of reliability measures for both quantitative and qualitative data. In particular, it provides a new unified way of interpreting the major quantitative reliability measures developed by Cronbach (1951, 1972) and Winer (1971) and a useful way of choosing among them. In the case of categorical data, the Perreault and Leigh measure is shown to be a PRL measure for the case of 2 judges. We also use the PRL paradigm to derive a categorical measure for the general case involving multiple judges and multiple assignment categories.

A major strength of the PRL approach is that it may be used to derive a reliability measure in virtually any setting. Rather than deriving a new measure based on ad hoc criteria, one may apply the consistent theoretical principle of proportional reduction in loss to derive a measure that is tailored to a specific research design and loss function. The resulting measure then provides a direct and easily interpretable assessment of the accuracy with which the characteristic of interest is estimated.

References

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City: American College Testing Program.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*, 687–699.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin, 88*, 322–328.

Costner, H. L. (1965). Criteria for measures of association. *American Sociological Review, 30*, 341–353.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: theory of generalizability for scores and profiles*. New York: John Wiley & Sons.

Hildebrand, D. K., Laing, J. D., & Rosenthal, H. (1977). *Prediction analysis of cross classifications*. New York: John Wiley & Sons.

Hughes, M. A., & Garrett, D. E. (1990). Intercoder reliability estimation approaches in marketing: A generalizability theory framework for quantitative data. *Journal of Marketing Research, 27*, 185–195.

Kassarjian, H. H. (1977). Content analysis in consumer research. *Journal of Consumer Research, 4*, 8–18.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Beverly Hills, CA: Sage Publications.

Peter, J. P. (1977). Reliability, generalizability, and consumer behavior. In W. D. Perreault (Ed.), *Advances in consumer research* (Vol. 4, pp. 394–400). Atlanta: Association for Consumer Research.

Perreault, W. D. Jr., & Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research, 26*, 135–148.

Schouten, H. J. A. (1982). Measuring pairwise agreement among many observers, II: Some improvements and additions. *Biometrical Journal, 24*, 431–435.

Schouten, H. J. A. (1986). Nominal scale agreement among observers. *Psychometrika, 51*, 453–466.

Winer, B. J. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill.

Manuscript received 11/26/92
Final version received 5/4/93