On ensuring a higher level of data quality when documenting human rights violations to support research into the origin and cause of human rights violations

Romesh Silva
School of International and Public Affairs
Columbia University
rs2032@columbia.edu

ABSTRACT

This paper reviews some of the measurement challenges posed when documenting large-scale human rights violations and considers a number of approaches. Practical methods and techniques based on recent field experience in Sri Lanka are presented which, when employed, will significantly improve the quality of human rights violations data. These improvements in data quality can enhance the ability of researchers to analyze the factors, origins and causes of human rights violations. Furthermore, through a review of the current literature on reliability measurement techniques, consideration is given to desirable statistical properties for reliability measures when applied to data on human rights violations.

INTRODUCTION

Human rights researchers collecting data in conflict zones face challenges of accuracy, precision, validity and reliability when documenting violations. These challenges arise from the need for a robust system which accurately records the nature, scope and intensity of violations. However the very nature of a conflict zone where large-scale human rights violations are being perpetrated makes it difficult to meet these demands.

With the establishment of a permanent International Criminal Court and increases in domestic legal human rights prosecutions[1], there is a pressing need for rigorous research into the origins and causes of human rights violations. The standard of evidence demanded by these forums therefore requires human rights researchers to not only explain the methods of data collection and statistical analysis used in amassing statistical human rights evidence but will also require them to provide complimentary scientific measures of the quality of their data. These requirements are now prompting the development of reliability measures for human rights data and statistical methods for operational use in the collection of data in the field.

The conceptualization of human rights is constantly evolving as evidenced by the continual evolution of human rights norms and standards in both customary international law and treaty law both internationally and regionally. However, as Henkin (2000) notes, the modern conception of rights has evolved principally from a legal and political basis through the International Bill of Rights. As a result only over the last ten years have statisticians started to formulate a statistical framework for human rights monitoring and reporting.

The Universal Declaration of Human Rights (UDHR) is a good starting point and provides for a general reference framework, in that it articulates a widely agreed set of standards of civil and political rights and economic and social rights. However, as it stands, the UDHR does not constitute a comprehensive framework for a statistical nomenclature and measurement system.

Based on fieldwork experiences in Sri Lanka and the lessons learned from Ball, Spirer and Spirer (2000) in studying large-scale violations in Guatemala, El Salvador and South Africa, it appears unrealistic to assume one can develop a universal statistical methodology to study human rights violations. Instead any Information System applied in a human rights setting needs to be custom-designed by a multidisciplinary group so as to yield high explanatory power of the nature and causes of large-scale violations while also accommodating local strengths, needs, weaknesses and conditions.

In this paper we consider three inter-related areas concerning the quality of human rights data: (1) the design of Human Rights Information Management Systems, (2) the theoretical concept of data quality and (3) the application of data quality measures to human rights settings.

DESIGN OF AN INFORMATION SYSTEM FOR HUMAN RIGHTS MONITORING

As Ball (1996) has noted there are essentially four basic steps in any Information System for human rights monitoring or human rights documentation project, namely (1) collection of information, (2) data processing, (3) database representation and (4) generation of analytic reports. This paper focuses on human rights data quality issues which arise principally in the data processing stage of an Information System.

According to Spain and Hollenbeck (1975) systematic coding systems provide the most mathematically sound methods for observational research. For the human rights field, such coding systems provide an explicit reference framework for studying violations. In particular they provide a

[1] Refer to Henkin (2000) and United Nations (1996).

rule-based foundation around which analysts can develop a methodology that is underpinned by the principles of accuracy, objectivity, consistency, credibility and security (Refer to Table 1).

Table 1: Basic Principles of Documentation and Reporting of Human Rights Violations

Principle	Basic Requirements for meeting Principle
Accuracy	- Non-extrapolation of data beyond content of raw information source - Non-simplification of data which results in loss of information about violations, victims or perpetrators - Representation of raw information according to Boundary Conditions and Counting Rules of Controlled Vocabulary
Objectivity	- Minimization of influence from ideological, political and/or ethnic prejudice - Avoidance of attempts to support preconceived hypotheses
Security	- Maintenance of confidentiality of all data (especially identities of victims, witnesses and perpetrators) - Data stored and maintained in a way that prevents unauthorized access, copying and tampering
Consistency	- Compliance with standardized data collection guidelines, data coding rules and data storage and security practices - Active improvement of rates of data coding consistency amongst and between different coders
Credibility	- Process through which data is collected and processed is systematic, reliable and secure - HRIMS is able to note different levels of detail of data

To produce meaningful human rights statistics, a controlled vocabulary is a fundamental requirement. It transforms the information on violations, victims and perpetrators into a countable set of data categories without discarding important information and without misrepresenting the collected information. Such a controlled vocabulary then facilitates the calculation of meaningful statistics about violations, perpetrators and victims on the sample data which can help us to answer the insightful question of "Who did what to whom?". As Ball (1996) and Spirer, Spirer and Ball (2000) have shown, the process of answering the question "Who did what to whom?" forces the researcher to decipher the often-complex relationships between violation, victim and perpetrator.[2] Such a system can lead researchers beyond the anecdotally based case-by-case analysis of human rights violations to a more systematic overview of the totality of large-scale human rights violations. Yet the standardized treatment and coding of individual information sources also facilitates movement between the micro and macro level analyses of violations. Furthermore, the attempt to answer the question "Who did what to whom?" via a controlled vocabulary can lessen the likelihood of analysts extrapolating the data beyond its significance.

The power of a controlled vocabulary lies in its ability to transform qualitative information into countable set of data which represent the nature, scope and intensity of human rights violations. Human rights fieldworkers can collect data from a wide range of information sources – ranging from legal case files, newspaper articles, e-mails, faxes, letters, phone conversations, testimonies, interviews, radio and television programs, video clips and photos. Thus the utilization of these wide-ranging sources can increase the coverage of violations reported, but at the same time also lends itself to an increase in the variability in the quality of data and the complexity of the process of coding raw information sources into human rights data. In particular, such a wide range of information sources may entail large variations in the detail, accuracy and verifiability of violations. Such variability points, in turn, to the need for a systematic management system to manage the quality of data in the face of variation in detail and accuracy of source information.

As H.F. Spirer and L. Spirer (2001) have argued, to ensure a high level of data quality every violation definition in a controlled vocabulary must satisfy the following properties:

- Mutually exclusive: no violation (or victim or perpetrator) can fit into any two definitions in the controlled vocabulary.

[2] Ball (1996) notes that the relationships between victim and violation, victim and perpetrator and perpetrator and violation can be one-to-one, one-to-many, many-to-one, or even many-to-many relationship.

- Exhaustive: there must exist a definition for every possible violation that can occur in the situation being studied.
- Distinguished: each definition must have an explicit characteristic which distinguishes this violation/victim/perpetrator from all others in the controlled vocabulary.
- Exemplified: each definition must be accompanied by examples showing how to apply the definition in a specific situation.
- Countable: each definition must contain a counting rule which explicitly states how violations, victims and perpetrators will be enumerated.

These five properties establish a set of measurement standards and rules to which notions of data quality can then be meaningfully applied. Before we consider different contemporary notions of data quality and their applicability to the human rights field, it is useful to examine the above five properties in more detail by way of an example from recent fieldwork conducted in Sri Lanka with the Human Rights Documentation Coalition.

Consider the violation category "Rape", which is characterized as follows:

VIOLATION CATEGORY	Rape
DEFINITION	Forceful/unwilling intercourse/penetration on any individual by another individuals genitals regardless of gender under intimidation, threat, fraud, lies, intoxication, or while in custody. The rape must be committed by a person/s identified on the list of perpetrators.
BOUNDARY CONDITION	Act must consist of vaginal or anal intercourse/penetration. Excludes acts which are covered by "Sexual Abuse (SEAB)" and "Genital Abuse (GEAB)" in the Controlled Vocabulary.
COUNTING RULE	▪ Continuous and one perpetrator = 1 violation (committed by single perpetrator) ▪ Continuous and multiple perpetrators = 1 violation (committed by multiple perpetrators) ▪ Non-continuous and different perpetrators = multiple violations (each separate act of penetration constitutes one violation)
EXAMPLE	A woman is gang-raped by three army personnel behind a security checkpoint

The acts which, when committed, fulfil the violation category of rape are specified in the definition. Furthermore, the definition also stipulates that this violation category is restricted to acts committed by persons on the perpetrators list[3], thus limiting the scope of the violation to actors in the context of the conflict being monitored.

The boundary condition of "Rape" then clearly distinguishes this violation from similar violations such as "sexual abuse", "torture" and "genital abuse".[4] Thus ensuring that any act which constitutes "rape" cannot also be classified under other violation categories in the controlled vocabulary at the same time – as required by the property of mutual exclusiveness.

Whereas, the counting rule assigns an explicit counting methodology so that violations, perpetrators and victims can be unambiguously enumerated in a consistent and accurate manner.

CONCEPTUALIZING A MEANINGFUL NOTION OF DATA QUALITY

Having established the pressing demand for high-quality human rights data when analyzing violations and their origin, cause and effects, we now examine the notion of data quality and, in particular, its several dimensions in relation to human rights information.

Many critics of statistics, such as Barsh (1993), that are based on qualitative research cite the difficulty of ensuring coding replicability as its major weakness. They reason that subjective biases introduced into the coding and analysis of qualitative data have adverse consequences. In a human rights setting, coding biases can lead to misrepresentation of the scale and magnitude of violations as well as lead to ineffective or counterproductive interventions. Hence, failure to

[3] The perpetrator list explicitly names all parties to the conflict including government agents (e.g. police department, military units, special forces, etc), political parties, armed groups and organizations.

[4] In the fieldwork being conducted by Human Rights Documentation of Sri Lanka: "sexual abuse" is defined as "Forceful/unwilling sexual contact on specific sexual areas of the human body (breasts, genitals, buttocks) regardless of gender at any time, on any individual including a minor, by another individual under intimidation, threat, fraud, lies or intoxication. "Genital abuse" consists of squeezing and or assault of the sexual organs through the use of devices and or objects.

strengthen, measure and maintain the quality of data can undermine the credibility, validity and usefulness of quantitative analyses of qualitative data sources.

However, in the social sciences, the notion of data quality is plagued with several problems. The most apparent one is that of a definitional and conceptual nature. As researchers across disciplines interchangeably use such terms as “reliability”, “validity”, “agreement”, “precision”, “accuracy” and “stability”, data quality has evolved as a fuzzy notion which has thus often been either overlooked or addressed in an ambiguous manner. Most notably, the concepts of reliability, accuracy and agreement have been used in loose and contradictory ways.

For example, in their respective studies in which qualitative information was coded into quantitative data:

W.G. Hopkins (2000) referred to reliability as the reproducibility of values of a test, assay or other measurement in repeated trials on the same individual. Better reliability implies better precision of single measurements and better tracking of changes in measurements in research or practical settings.

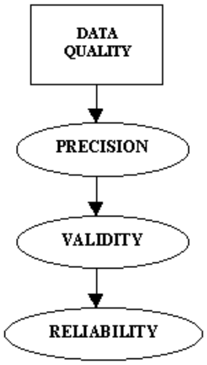
P. Martin and P. Bateson (1986) defined reliability as the extent to which measurement is free from random errors, and accuracy as the extent to which measurement is free from systematic errors.

D.P. Hartmann and D.D. Wood (1982) defined agreement as the degree to which observers assign the same raw score, whereas reliability is the degree to which observers assign the same standard score to an event or person.

For the purposes of collecting and coding human rights data, we propose a hierarchical structure of data quality which encompasses the notions of precision, validity, and reliability (Refer to Figure 1 above). However, such a hierarchical structuring delineates between the ability of data to accurately represent the nature and scope of human rights violations but does not imply a hierarchical ordering of individual human rights themselves. In fact a motivation of such a hierarchy is to guard against the practice of only recording the “most serious” violation. As Ball (1996) and Spirer & Spirer (2000) note, this practice leads to the loss of a wealth of information about multiple violations. It is this information which, although often of varying degrees of detail and accuracy, is of fundamental importance in understanding the different relationships between victims, violations and perpetrators. Furthermore, as noted by Ward (2000) and Ball (1996) the practice of reporting only the “most serious violation” produces a significant underestimation of the nature and intensity of the human rights situation and leads to distortion of the trends and patterns of human rights violations, their victims and perpetrators

We define *reliability* is then the extent to which the standardized coding procedure yields the same results when repeated by a different data coder using the same controlled vocabulary. Validity is the extent to which the coding process and its associated Controlled Vocabulary represent the intended, and only the intended, phenomena. Thus validity is a measure of the extent to which the Information System is free from systematic errors. Precision is the extent to which the actual coded data obtained from applying the standard controlled vocabulary represents exactly what is contained in the raw information source.

Figure 1: Hierarchical Structuring for Quality of Human Rights Data



Hence, based on the above-mentioned hierarchical structuring of data quality, it follows that as the number of violation categories in a controlled vocabulary increases the human rights data will become more precise. However, this increased precision may come with a trade-off in the way of decreased reliability as a result of data coders finding it more difficult to identify and classify violations as the controlled vocabulary becomes larger and more complex. Yet data reliability is influenced not only by the strength of the standardized controlled vocabulary but also a function of the information source (as the level of difficulty to code different information sources can differ) and the particular data coder. For example, as explained in Neundorf (2000), levels of training, experience and fatigue of data coders can influence the data coder’s ability to apply the controlled vocabulary.

The Units of Analysis

Data collected through observational studies often do not necessarily have an intuitive unit of analysis from which data quality and reliability can be analyzed. This is particularly the case for data on human rights violations.

As systemic data quality issues, which are a function of the particular information source, are beyond the scope of this paper, we focus our attention on the primary unit of analysis for monitoring data reliability and overall data quality.

As noted above, the ensuing units of analysis arising from systematic coding of human rights information sources can be derived from the question "Who did what to whom?". Thus the purpose of the analysis defines the units of analysis. Therefore it follows that the basic units of analysis are the violation category, perpetrator identity/affiliation and victim identity/affiliation.[5] However, as the existence of a perpetrator and a victim is contingent on the existence of a violation, it follows that the primary unit of analysis is "violation".

The two primary sources of disagreement between data coders, which may decrease the level of data reliability, are the instances when data coders either (i) disagree on the existence of a violation or (ii) agree on the existence of a violation but disagree on how to classify the violation according to the controlled vocabulary. The existence of variation amongst data coders in classifying perpetrators and victims is not only a second order issue but also, based on field experience in Sri Lanka, more of a systemic function of the information source than a direct effect of misapplication of the controlled vocabulary by data coders.

APPLYING MEASURES OF DATA QUALITY TO DATA ON HUMAN RIGHTS VIOLATIONS

As mentioned above, social scientists may erroneously apply data quality terminology. This problem originates from the mis-specification of the measurement concept and results in the use of a misleading statistical measure. In this section we examine three widely-used measures of data quality in the social and behavioral sciences: the proportion of agreement, the kappa coefficient and the generalizability coefficient. We explore some of their basic properties to assess their applicability to human rights violation data.

As mentioned above, for human rights projects which endeavor to answer the question "Who did what to whom?", the primary unit of analysis is the "violation category" and secondary units of analysis are the identities and affiliations of the perpetrator and victim. However, unlike other fields which utilize agreement and reliability measures, the human rights field has an added complication. Data coders in most fields observe the same number of objects (e.g. patients in a psychiatric hospital, etc...). Yet in the human rights field, data coders scanning raw information sources (e.g. a witness testimony, a newspaper report or legal case file) do not necessarily observe the exact same number of primary objects (i.e. violations). Hence reduced reliability and validity can be caused by not only (i) categorization of an act into an incorrect violation category but also (ii) non-recognition of an act as a violation according to the controlled vocabulary.

Yet, such measures as the proportion of agreement, kappa coefficient and generalizability coefficient necessarily require that all coders observe the same number of objects. Hence, in order to apply these standard data quality measures to the human rights setting, we must first manipulate the data set to ensure that all coders appear to observe the same number of objects. Thus for the purposes of data quality measurement, we add the category "Non-Violation" to the controlled vocabulary. Thus, for example, in the case where one data coder observes the violation "Loss of Livelihood" but others do not, this would be regarded as one data coder observing the violation "Loss of Livelihood" and other observing the violation "Non-Violation".

We note, however, that this practice does not assume that the coding of the violation "Loss of Livelihood" is necessarily correct. Rather, it provides a framework in which disagreement between coders on the existence of a violation can be treated analogously to disagreement between coders on which violation category should be used for a specific act – themselves both sources of unreliability in human rights data.

[5] For example, in our current fieldwork in Sri Lanka thirty five separate violation categories have been defined which encompass violations of (i) property rights, (ii) civil and political rights, (iii) economic and social rights and (iv) legal rights. These violations include destruction of property, forced recruitment into an armed group or organization, restriction on freedom of association and delays in indictment. The set of perpetrators includes the armed forces, police, political parties, prison authorities, militant groups and government's special units. While data on the key distinguishing features of victims are collected. These distinguishing features include sex, religion, ethnicity, place of residence, occupation.

The Proportion of Overall Agreement

The proportion of agreement simply reports the proportion of times observers agree on the assignment of violation classifications from a standardized vocabulary. For the simple case of two coders rating acts into C violation categories (where the C^{th} category is the artificially constructed category “non-violation”), we would obtain the following agreement matrix:

Figure 2: Agreement Matrix for 2 Coders using a controlled vocabulary with C-1 different violation categories (the Cth violation category being the artificially constructed “No violation”)

<table>
<tr><td colspan="8">Ratings by two coders into many categories</td></tr>
<tr><td></td><td colspan="7">Coder 2</td></tr>
<tr><td rowspan="5">Coder 1</td><td></td><td>1</td><td>...</td><td>C-1</td><td>C</td><td><i>total</i></td></tr>
<tr><td>1</td><td>n_{11}</td><td>...</td><td>n_{1C-1}</td><td>n_{1C}</td><td>$n_{1.}$</td></tr>
<tr><td>2</td><td>N_{21}</td><td>...</td><td>n_{2C-1}</td><td>n_{2C}</td><td>$n_{2.}$</td></tr>
<tr><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td><td>...</td></tr>
<tr><td>C-1</td><td>n_{C-11}</td><td>...</td><td>N_{C-1C-1}</td><td>n_{C-1C}</td><td>$N_{C-1.}$</td></tr>
<tr><td></td><td>C</td><td>n_{C1}</td><td>...</td><td>n_{CC-1}</td><td>n_{CC}</td><td>$n_{C.}$</td></tr>
<tr><td></td><td>total</td><td>$n_{.1}$</td><td>...</td><td>$n_{.C-1}$</td><td>$n_{.C}$</td><td>N</td></tr>
</table>

Thus the observed proportion of overall agreement is given by:

$$P_0 = [n_{11} + n_{22} + n_{33} + \ldots + n_{C-1C-1} + n_{CC}] / N$$

$$= [1 / N] \sum_{i=1}^{C} n_{ii} \qquad (1)$$

where n_{ij} represents the number of acts assigned violation category i by Coder 1 and violation category j by Coder 2, with i,j = 1, ...,C. In this notation, “$n_{.j}$” is the sum of n_{ij} for j=1,...C which amounts to the marginal sum for Coder 1 and violation category i. By construction $n_{CC} = 0$.

For K coders formula (1) can be generalized, by way of the manipulations outlined in the Appendix I, to the following:

$$P_0 = \sum_{i=1}^{N} \sum_{k=1}^{K} n_{ik}(n_{ik}-1) \quad / \sum_{k=1}^{K} n_k(n_k-1) \qquad (2)$$

The proportion of overall agreement measure is particularly appealing due to its ease of calculation. A corollary of its ease of computation, as noted by Mitchell (1979), has been its widespread use in observational studies such as in Early Childhood Development and Developmental Psychology. It is this ease of computation and its avoidance of using relatively more sophisticated measures like variance and covariance which make it also appealing for use in human rights field work by NGOs who possess varying degrees of capacity in this area.[6]

However, the proportion of overall agreement does not have a robust mathematical interpretation nor does it possess any insightful metric properties. A number of factors can affect the percentage of agreement, most notably the number of violation categories in the controlled vocabulary. As a smaller number of violation categories are used in a controlled vocabulary, a higher proportion of agreement is expected due to the increased agreement between coders due to chance. Furthermore, the proportion of overall agreement does not isolate the source of disagreement, but merely reports an arbitrary index of agreement. Yet, the magnitude of the overall proportion of agreement has no clear meaning: in particular, there isn’t a threshold at which “acceptable” and “unacceptable” proportions can be distinguished. Without explanation, Krippendorf (1980) and Brouwer et al. (1969) arbitrarily adopted the policy of “reporting on variables only if their reliability was above 0.8 and admitted variables with reliability between 0.67 and 0.8 only for drawing highly tentative and cautious conclusions”.

Correcting for Chance Agreement – the Kappa Coefficient

In the contemporary literature, the main criticism of the overall proportion of agreement has been its inability to correct for chance agreement. In response to this perceived limitation Cohen (1960) developed the kappa coefficient, which amounts to a chance corrected proportion of agreement.

The kappa coefficient, κ, is defined as

$$\kappa = (P_0 - P_E) / (1 - P_E) \qquad (3)$$

Where P_E is the proportion of agreement due to chance alone and P_0, the proportion of overall agreement, is given by equation (2). Mathematically, P_E can be represented by

$$P_E = [1/N^2] \sum_{i=1}^{N} n_{.i}\, n_{i.} \qquad (4)$$

However, the term P_E is relevant strictly only under conditions of statistical independence. Yet, in the human rights setting statistical independence appears to be a questionable assumption, as data coders are not independent in as much as they are using a common controlled vocabulary to assign violation categories.

[6] For a discussion of the statistical competence of human rights NGOs, refer to Spirer (2000).

Other criticisms of kappa are, like the proportion of agreement, its inability to explicitly isolate the sources of disagreement and the lack of coherent meaning of its magnitude. Although, Landis and Koch (1977) have made rule-of-thumb-type suggestions for "acceptable" and "unacceptable" ranges of kappa, these ranges still lack insight into the sources of disagreement and therefore do not lead to any theoretically-based corrective measures.

Therefore we advocate a two-track approach. On the one-hand the calculation of a simple, yet complementary measure to the proportion of overall agreement – the specific proportion of agreement. While on the other hand, further research into the application of robust measures of data quality which go beyond the simple identification of unreliability and also disaggregate the relative source of unreliability. Existing robust measures, from the field of Generalizability Theory are strong candidates for this.

A Simple, Complementary Measure – Proportion of Specific Agreement

However, by analyzing clusters off the diagonal of the agreement matrix (Figure 2) one can identify the codes which are the causes of most disagreement. Such analysis is insightful in identifying areas for re-training of data coders and possible clarifying amendments to the controlled vocabulary. Furthermore they potentially point to the main impediments to high levels of reliability and therefore high levels of data quality.

These clusters off the diagonal can be explicitly represented by the Proportion of Specific Agreement measure. In our simple 2-coder example, the proportion of specific agreement for violation category i equates to

$$P_{specific} = [2 * n_{ii}] / [n_{i.} + n_{.i}] \qquad (5)$$

which for K coders generalizes, via the manipulations in Appendix 2, to the form in equation (6)

$$P_{specific} = \sum_{k=1}^{K} n_{ik}(n_{ik}-1) \;/\; \sum_{k=1}^{K} n_k(n_k-1) \qquad (6)$$

Equation (6) can then be interpreted as the estimated conditional probability that given that one of the coders, randomly selected, assigns violation category i to a certain violation object[7], that the other coder will also assign violation category i to the same violation object.

Hence, by complementing the proportion of overall agreement with proportion of specific agreement, further insight can be gained into which violation categories are the primary sources of disagreement. Hence these simple descriptive statistics can yield valuable insights into the levels of agreement generated by data coders applying a standardized controlled vocabulary.

Towards a more comprehensive approach - Generalizability Theory

The power of generalizability theory is that it provides for a more detailed analysis of sources of variance components, and hence goes beyond the level of descriptive statistics such as the kappa coefficient and proportions of agreement.

Until recently the concept of variance of categorical data had not been thoroughly explored. Even though Gini (1939) conceptualized the concept of variance of categorical data in the first half of the 20th century. His conceptualization can be applied to pair-wise agreement between coders. In particular, if coder k_i and k_j make identical assignments of a violation category to a violation object, then the pair-wise difference is set to zero, otherwise it is set to 1.

Mathematically, the difference in assignment of violation category is represented as:

$$d_{ij} = d(k_i, k_j) = \begin{cases} 0, \text{ if } k_i \text{ and } k_j \text{ assign the same violation category to a violation object} \\ 1, \text{if } k_i \text{ and } k_j \text{ assign the a different violation category to a violation object} \end{cases} \qquad (7)$$

We note that the Gini (1939) methodology simply treats all disagreements equally (in that any pair-wise difference contributes one unit to the variance regardless of "how much" the assigned violations differ – e.g. perhaps it could be argued that the two violation categories "Rape" and "Sexual Abuse" differ less than "Rape" and "Disappearance"), yet Gini's definition of variance would not differentiate between degree of difference but rather treat any difference of category equally.

However, through Gini's conceptualization of variance of categorical data, a meaningful concept of variance partitioning for categorical data can then be applied. By application of a partitioning devised by Light and Margolin (1971), the total

[7] We refer to a "violation object" as an action to which at least one data coder assigns a violation category from the standardized vocabulary. We use the word "violation object" instead of "violation" so an not to erroneously assume that all violation objects assigned a violation category by a data coder do, in fact, constitute a violation according to the standardized controlled vocabulary.

variance (SS_T) can then be partitioned into a between-coder (SS_{BC}) and within-coder (WSS_{WC}) component, such that:

$$SS_T = (nk/2) - (1/2nk) \sum_{k=1}^{K} n^2_{..k} \qquad (8)$$

$$SS_{BC} = (1/2n) \sum_{k=1}^{K} \sum_{n=1}^{N} n^2_{.kn} - (1/2kn) \sum_{k=1}^{K} n^2_{..k} \qquad (9)$$

$$SS_{WC}=(nk/2) - (1/2n)\sum_{k=1}^{K} \sum_{n=1}^{N} n^2_{.kn} = SS_E + SS_{BVO} \qquad (10)$$

Equation (10) represents the pooled "within-coder" effect with the residual term.

From this the "proportion of variation explained", R^2, is thus defined as

$$R^2 = SS_{BC} / SS_T \qquad (11)$$

Analagously, by re-application of Gini's variance measure for categorical data, Light and Margolin (1971) obtained canonical forms for the three above-sums of squares equation, as follows:

$$SS_T = (nk/2) - (1/2nk) \sum_{j=1}^{J} n^2_{.j} \qquad (12)$$

$$SS_{BVO} = (1/2n) \sum_{j=1}^{J} \sum_{n=1}^{N} n^2_{nj} - (1/2kn) \sum_{j=1}^{J} n^2_{.j} \qquad (13)$$

$$SS_{WVO} = (nk/2) - (1/2n)\sum_{j=1}^{J} \sum_{n=1}^{N} n^2_{nj} = SS_E + SS_{BC} \qquad (14)$$

These formulae are essentially a violation-object-by-violation category format, where
SS_{BC} is the sum of squares between data coders,
SS_E is the sum of squares of residuals,
k is the number of data coders,
n is the total number of violation-objects being classified,
n_{nj} is the number of data coders who assign a particular category j (j=1,2,...J) to violation object n (n=1,2,...,n),
n_j is the total number of classifications in the jth violation category,
J-1 is the number of violation categories in the controlled vocabulary (and therefore the Jth violation category is the artificially constructed "No Violation" category).

The above sums-of-squares values can then be converted into their respective variance components, using the standard formula:

$$\sigma^{2 =} (MS - MS_e) / (df) \qquad (15)$$

By utilizing these two canonical forms of variation in the categorical data, we are able to separate the systematic variation between raters from the residual variation. As a result, we can obtain the Generaliability Coefficient which is defined as

$$G = \sigma_{VO}^2 / (\sigma_{VO}^2 + \sigma_C^2 + \sigma_e^2) \qquad (16)$$

By employing generalizability theory, researchers can thus focus on the significance of variance components and measurement error, instead of just calculating a descriptive statistic such as kappa or a proportion of agreement which lacks meaningful interpretation. In particular, the relative amount of variance contributed by the violation-object effect, between-coder effect and residual variation. The interpretation of these variance components, in turn, can then contribute to quality control and improvement of the data collection and data coding process in a human rights documentation project.

However, as Light and Margolin (1971) showed the major limitation of Generalizability theory is that presently there is no method to determine whether the Generalizability coefficient differs significantly from zero. However, even if such a method did exist, the interpretation of the magnitude of the G coefficient would still need to be based on an individual judgement. Furthermore, especially pertinent in the human rights setting, the calculation of the Generalizability coefficient is defined only for when a constant number of data coders is used (as is the Kappa Coefficient). However, in a human rights setting often the number of data coders varies hence making the application of generalizability theory difficult.

CONCLUSION

In this paper we have outlined a statistical framework in which data quality of a HRIMS can be conceptualized. In it we have endeavored to delineate between the different gradations of data quality encompassed by agreement, reliability and validity. However, via a review of present literature we have noted the present theoretical shortcomings which prevent the application of such robust statistical measures as the generalizability coefficient to human rights settings. As an interim measure, we advocate the use of descriptive statistical measures such as the proportion of overall agreement and proportion of specific agreement as proxies for monitoring data quality. These measures, while not providing a robust statistical framework do however provide insight into the sources of disagreement between coders and therefore are useful and practical measures for monitoring data quality in the field. Further research, however, is required to ensure that more robust measures such as the kappa coefficient and generalizability coefficient can be employed in a human rights setting.

REFERENCES

Bakeman, R. and Gottman, J.M. (1997) *Observing Interaction: An Introduction to Sequential Analysis (2nd Edition).* Cambridge University Press (UK).

Ball, P. (1996) *Who did what to whom? : planning and implementing a large scale human rights data project.* Washington, D.C. (USA): AAAS.

Ball, P, Spirer, H.F. and Spirer, L. (2000) *Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis.* Washington, D.C. (USA): AAAS.

Barsh, R.L. (1993) *Measuring Human Rights: Problems of Methodology and Purpose.* Human Rights Quarterly 15(3), 87-121.

Brouwer, M., Clark, C.C., Gerbner, G., and Krippendorf, K. (1969) *The Television World of Violence.* 311-339 and 519-591 in R.K. Baker and S.J. Ball (eds.) Mass Media and Violence: A Report to the National Commission on the Causes and Prevention of Violence. Washington D.C: Government Printing Office.

Carletta, J. (1988) *Assessing Agreement on Classification Tasks: the kappa statistic.* Working Paper. University of Edinburgh (UK).

Cichchetti, D.V. and Sparrow, S.A. (1981) *Developing Criteria for Establishing Interrater Reliability of Specific Items: Applications to Assessment of Adaptive Behavior.* American Journal of Mental Deficiency. 86(2): 127-137.

Cohen, J.A. (1960) *A Coefficient of Agreement for Nominal Scales.* Educational and Psychological Measurement. 20(1):37-46.

Fleiss, J.L. (1981) *Statistical Methods for Rates and Proportions (2nd Edition).* New York: John Wiley.

Gini, C (1939) *Variabilita e concentrazione.* Memorie di metodologia statistica Vol. 1 [Variability and concentration. Vol. 1: Notes on statistical methodology]. Milano, Italy: Guiffre.

Hartmann, D.P. and Wood, D.D. (1982) Observational Methods. In A.S. Bellack, M. Hersen & A.E. Kazdin (Eds.), *International Handbook of Behavior Modification and Therapy* (pp. 109-138). New York: Plenum.

Henkin, L. (2000) *The International Human Rights Movement* in Henry J. Steiner and Philip Alston (eds.), International Human Rights in Context: Law, Politics, Morals: Oxford University Press, Second Edition.

Henkin, L. (2000) *Human Rights and Humanitarian Law* in Henry J. Steiner and Philip Alston (eds.), International Human Rights in Context: Law, Politics, Morals: Oxford University Press, Second Edition.

Krippendorf, K. (1980) *Content Analysis: An Introduction to its Methodology.* Beverly Hills, CA: Sage Publications.

Landis, J.R. and Koch, G.G. (1977) *The Measurement of Observer Agreement for Categorical Data.* Biometrics, 10:133-139.

Light, R.J. and Margolin, B.H. (1971) *An Analysis of Variance of Categorical Data.* Journal of the American Statistical Association (Theory and Methods Section). 66(335): 534-544.

Martin, P. and Bateson, P. (1986) *Measuring Behavior: An Introductory Guide.* London: Cambridge University Press.

Mitchell, S.K. (1979) *Interobserver Agreement, Reliability and Generalizability of data collected in Observational Studies*. Psychological Bulletin, 86:376-390.

Neuendorf, Kimberly A. *The Content Analysis Guidebook.* Thousand Oaks, CA: Sage, 2002.

Shavelson, R.J. and Webb, N.M. (1991) *Generalizability Theory: A Primer.* Newbury Park, California (USA): Sage Publications.

Spirer, H.F. (1988) *Quantitative Measurement of Human Rights.* Paper presented at the Annual Meeting of the Decision Sciences Institute, Las Vegas, Nevada (USA).

Spirer, H.F. (2000) *Meeting the Statistical Needs for Training and Education.* Paper presented at the International Association of Official Statistics Meeting, Montreux (Switzerland).

Spirer, H.F. and Spirer, L. (2001) *Intermediate Data Analysis for Human Rights: A Handbook*. Stamford, Connecticut (USA): AAAS.

United Nations General Assembly (1996), *Establishment of an International Criminal Court* (A/RES/51/207), December 17.

Ward, Ken (2000) *The United Nations Mission for Verification of Human Rights in Guatemala: Database Representation* in *Making the Case: Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis.* Washington, D.C. (USA): AAAS.

APPENDIX I

1. Generalized Case of Proportion of Overall Agreement

Consider k violation objects, k >2, which are coded by a group of data coders.

Then the classifications of violation object k by the group of data coders according to a controlled vocabulary with C-1 different violation categories (the C^{th} violation category being the artificially constructed violation category "No violation") can be represented by the following set:

$\{n_{jk}\}_{(j=1,...,C)} = (n_{1k}, n_{2k}, ..., n_{ck})$

where n_{jk} is the number of times violation category j (j=1,...,C) is assigned for violation object k.

For violation object k, the number of actual agreements on violation category j is given by $n_{jk}(n_{jk}-1)$

The total number of agreements specifically on violation category j across all violation objects is given by

$$SA_j = \sum_{k=1}^{K} n_{jk}(n_{jk}-1)$$

The total number of actual agreements, regardless of violation category, is equal to the sum of SAj across all C violation categories (i.e. the C-1 violation categories in the standardized controlled vocabulary and the C^{th} artificially constructed violation category "No Violation")

$$OTA = \sum_{j=1}^{C} \sum_{k=1}^{K} n_{jk}(n_{jk}-1) \qquad (A1)$$

Now if we let the total number of classifications made on violation object k be denoted by

$$n_k = \sum_{J=1}^{C} n_{jk}$$

Then the total number of possible agreements for all K violation objects is given by

$$OPA = \sum_{k=1}^{K} n_k(n_k-1) \qquad (A2)$$

By dividing (A1) by (A2) we then obtain the overall proportion of observed agreement

$$P_{overall} = OTA/OPA = \sum_{k=1}^{K} n_k(n_k-1) / \sum_{j=1}^{C} \sum_{k=1}^{K} n_{jk}(n_{jk}-1)$$

2. Generalized Case of Proportion of Specific Agreement

To generalize our 2-coder formula for the proportion of specific agreement, which is given by

$P_j(I) = 2 * n_{ii} / n_{i.} + n_{.i}$

We recognize that the total number of agreements on violation category j across all violation objects is

$$STA = \sum_{k=1}^{K} n_{jk}(n_{jk}-1)$$

It follows that the number of possible agreements specifically on violation category j for violation object k is given by $n_{jk}(n_k-1)$

And the number of possible agreements specifically on category j for violation object k is then given by $n_{jk}(n_k-1)$

and the number of possible agreements on violations category j across all K violation objects is

$$SPA = \sum_{k=1}^{K} n_{jk}(n_k-1)$$

Thus the specific proportion of agreement for violation category j is obtained simply by dividing the total number of specific agreements on violation category j by the total number of possible specific agreements on violation category j

$$P_{specific}(j) = STA/SPA$$

$$= \sum_{k=1}^{K} n_{jk}(n_{jk}-1) / \sum_{k=1}^{K} n_{jk}(n_k-1)$$