Intercoder Reliability (ICR) Assessment for the National Human Rights Institution (NHRI) Data Collection Project: Organizational Data

Ryan M. Welch
Department of Political Science, Florida State University

Courtenay R. Conrad
Department of Political Science, University of California, Merced

Jacqueline H.R. DeMeritt
Department of Political Science, University of North Texas

Will H. Moore
Department of Political Science, Florida State University

March 17, 2015

Version 1.1

This project has received support from the Department of Political Science at the University of California, Merced, the Department of Political Science at the University of North Texas, the Department of Political Science at Florida State University, the Kroc Institute for International Peace Studies at University of Notre Dame, and the Department of Political Science and Public Administration at the University of North Carolina at Charlotte.

Contents

1 Introduction **1**

2 Collecting Data and Analysis **1**
2.1 Coders, Training, and Data Collection 1
2.2 Identifying Documents 2

3 Measures of Reliability **2**

4 Intercoder Reliability Scores **3**

1 Introduction

To evaluate the extent to which the Organizational data for the NHRI data collection project are reliable we recruited and trained a second set of coders who coded a common set of documents. This document describes the process for the data collection, a discussion of the Intercoder Reliability[1] (ICR) measures used, and a reporting of the ICR measures follow.

2 Collecting Data and Analysis

2.1 Coders, Training, and Data Collection

We recruited five undergraduate coders identified by the authors from past or current classes as students with the ability to utilize high cognitive ability in a self-motivated, meticulous manner. We used the same process to identify a pool of recruits for the coding of the data. The project manager emailed the prospective coders the coding rules, asked them to review them, and then met with them to explain the project and their role in it. The meeting allowed coders to ask questions about the coding rules.

Training documents were harvested[2] by consulting prior spreadsheets and choosing documents that have been coded in the past. Examples include founding legislation, web pages, annual reports, and NGO reports. The project manager assigned the same training document to each coder. Using the same training document for every coder allows everyone to be on the same page in the next meeting while learning how to navigate the coding rules. It also allows the project manager to better identify which coders are ready to be certified as everyone is working on the same document. The project manager qualitatively identified coders to take the certification test.

The project manager identified different documents to act as testing documents. He then coded the testing document before assigning it to selected coders. If the coder attained at least 80% agreement (s)he moved onto the coding phase. If not, the process was repeated with another

[1] This is also known as Interrater Reliability.

[2] Below we discuss in more detail how we identified documents.

testing document.

Five coding documents were identified in the same manner as the training and certification documents. The coder had a week to code the document. The data resulting from this coding was used to assess the inter-coder reliability.

2.2 Identifying Documents

All three types of documents (training, testing, coding) were chosen in the following manner:

- Randomly [3] select one of the spreadsheets used to create the NHRI Organization Database.
- Randomly select one of the columns within the spreadsheet.
- Use the source URL to obtain source as potential document.
- If source is not usable,[4] keep moving right on spreadsheet until appropriate source is found.

3 Measures of Reliability

In order to test the reliability, we use two measures: overall percentage agreement (OPA) and proportion of overall agreement (P_o). The overall percentage agreement is the mean level of agreement across all pairs of coders (Fleiss, 1971). In order to calculate the OPA, each coder i is paired with each coder $\sim i$ and the percentage of their agreements are measured. Then the mean over each pair is calculated.

We also calculate the Proportion of Overall Agreement (P_o)(Fleiss, 1971, 1981). The proportion of overall agreement gives a simple baseline to assess reliability as it is the total number of actual agreements on categories $J = 1, \ldots, C$ divided by the total number of possible agreements for each case $K = 1, \ldots, k$ across all raters, n:

$$P_o = \frac{\sum_{j=1}^{C} \sum_{k=1}^{K} n_{jk} n_{jk-1}}{\sum_{k=1}^{K} n_{jk} n_{jk-1}} \tag{1}$$

[3] All randomization done with www.random.org

[4] Unusable sources include those that contain scant information or broken links.

One should be aware that for both measures presented, agreement may result from chance. In order to account for this threat, many researchers turn to Krippendorff's α (Krippendorff, 2004). However, if the marginals in the coincidence matrices used to calculate the Krippendorff's α are substantially imbalanced, a paradox results in which high agreement can lead to misleadingly low Krippendorff α values (Feinstein and Cicchetti, 1990). The imbalance in the marginals occurs for our data, thus we do not rely on the Krippendorff α measure.

4 Intercoder Reliability Scores

For both the OPA and P_o a score of at least 0.8 is a common threshold used to consider a variable sufficiently reliable to be used in analyses. Table 1 presents the calculated OPA and P_o for each variable. Variables that cross the 0.8 threshold for both OPA and P_o are bold, and those that cross one are italicized, but we encourage researchers to use the values in the table to decide which variables they deem reliable for their studies.

There are 52 Organizational variables, 32 of which have both an OPA and a P_o score of 0.8 or greater, and eight of which have an OPA or P_o score of at least 0.8.

Table 1: Calculated Intercoder Reliability

Variable	Overall Percentage Agreement	Proportion of Overall Agreement
Begin Information Date	0.34	0.00
End Information Date	0.72	0.40
ICC Status	0.92	0.80
NHRI Office Type	0.72	0.40
Established by	0.70	0.40
Year Formally Established	0.50	0.20
Year First Occupied	0.48	0.20
Worker's Rights	0.84	0.60
Arbitrary Detention	0.92	0.80
Disappearance	0.68	0.40

Extra-judicial Killing	0.76	0.40
Torture	0.92	0.80
Freedom of Speech	0.92	0.80
Freedom of Assembly	0.92	0.80
Freedom of Foreign Movement	0.84	0.60
Freedom of Domestic Movement	0.92	0.80
Electoral Self–Determination	0.92	0.80
Freedom of Religion	1.00	1.00
Women's Economic Rights	0.84	0.60
Women's Political Rights	0.84	0.60
Women's Social Rights	0.84	0.60
Children's Rights	0.92	0.80
Non–Human Rights Objectives	0.44	0.00
Scope of Jurisdiction	0.40	0.00
No Reporting Required	0.92	0.80
Report to Executive	0.80	0.60
Report to Legislature	0.76	0.60
Report to Judiciary	1.00	1.00
Report to International Institution	0.88	0.80
Report to Public	0.68	0.40
Independent	0.60	0.20
Member Appointment	0.62	0.20
Leadership Appointed by Executive	0.72	0.40
Leadership Appointed by Legislature	0.80	0.60
Leadership Appointed by Judiciary	0.92	0.80
Leadership Appointed by UN	1.00	1.00
Leadership Appointed by NHRI	0.76	0.40
Leadership Appointed by Public	1.00	1.00
Leadership Appointed by Other	0.92	0.80
Chairperson Term	0.88	0.80

Donor Source: Government	0.80	0.60
Donor Source: Private	0.92	0.80
Donor Source: IGO	0.92	0.80
Donor Source: NGO	1.00	1.00
Donor Source: Other Country	0.92	0.80
Permitted: Complaints	0.60	0.20
Permitted: Investigations	0.92	0.80
Permitted: Bring Charges	0.48	0.00
Permitted: Compel Testimony	0.60	0.20
Permitted: Visit	0.68	0.40
Permitted: Publish Findings	0.88	0.80
Permitted: Levy Punishment	0.92	0.80
Permitted: Other	0.60	0.20
Relations with NGOs	0.76	0.40

References

Feinstein, Alvin R. and Domenic V. Cicchetti. 1990. "High Agreement but Low Kappa: I. The Problems of Two Paradoxes." *Journal of Clinical Epidemiology* 43(6):543–549.

Fleiss, Joseph L. 1971. "Measuring nominal scale agreement among many raters." *Psychological Bulletin* 76(5):378–382.

Fleiss, Joseph L. 1981. The measurement of interrater agreement. In *Statistical methods for rates and proportions*, ed. J.L. Fleiss, B. Levin and M.C. Paik. New York: Wiley pp. 212–236.

Krippendorff, Klaus. 2004. *Content analysis: An introduction to its methodology*. Thousands Oaks: Sage Publications.