Literature Review of *Inter-rater Reliability*

Inter-rater reliability, simply defined, is the extent to which the way information being collected is being collected in a consistent manner (Keyton, et al, 2004). That is, is the information collecting mechanism and the procedures being used to collect the information solid enough that the same results can repeatedly be obtained? This should not be left to chance, either. Having a good measure of inter-rater reliability rate (combined with solid survey/interview construction procedures) allows project manners to state with confidence they can be confident in the information they have collected. Statistical measures are used to measure inter-rater reliability in order to provide a logistical proof that the similar answers collected are more than simple chance (Krippendorf, 2004a).

Inter-rater reliability also alerts project managers to problems that may occur in the research process (Capwell, 1997; Keyton, et al, 2004; Krippendorf, 2004a,b; Neuendorf, 2002). These problems include poorly executed coding procedures in qualitative surveys/interviews (such as a poor coding scheme, inadequate coder training, coder fatigue, or the presence of a rogue coder – all examined in a later section of this literature review) as well as problems regarding poor survey/interview administration (facilitators rushing the process, mistakes on part of those recording answers, the presence of a rogue administrator) or design (see *Survey Methods*, *Interview/Re-Interview Methods*, or *Interview/Re-Interview Design* literature reviews). From all of the potential problems listed here alone, it is evident measuring inter-rater reliability is important in the interview and re-interview process.

Preparing qualitative/open-ended data for inter-rater reliability checks

If closed data was not collected for the survey/interview, then the data will have to be coded before it is analyzed for inter-rater reliability. Even if closed data was collected, then coding may be important because in many cases closed-ended data has a large amount of possibilities. A common consideration, the YES/NO priority (Green, 2004), requires answers to be placed into yes or no paradigms as a simple data coding mechanism for determining inter-rater reliability. For instance, it would be difficult to determine inter-rater reliability for information such as birthdates. Instead of recording birthdates, then, it can be determined whether the two data collections netted the same result. If so, then YES can be recorded for each respective survey. If not, then YES should be recorded for one survey and NO for the other (do not enter NO for both, as that would indicate agreement). While placing qualitative data into a YES/NO priority could be a working method for the information collected in the ConQIR Consortium given the high likelihood that interview data will match, the forced categorical separation is not considered to be the best available practice and could prove faulty in accepting or rejecting hypotheses (or for applying analyzed data toward other functions). It should, however, be sufficient in evaluating whether reliable survey data is being obtained for agency use. For best results, the survey design should be created with reliability checks in mind, employing either a YES/NO choice option (this is different than what is reviewed above – a YES/NO option would include questions like, “Were you born before July 13, 1979?” where the participant would have to answer yes or no) or a likert-scale type mechanism. See the *Interview/Re-Interview Design* literature review for more details.

How to compute inter-rater reliability

Fortunately, computing inter-rater reliability is a relatively easy process involving a simple mathematical formula based on a complicated statistical proof (Keyton, et al, 2004). In the case of qualitative studies, where survey or interview questions are open-ended, some sort of coding scheme will need to be put into place before using this formula (Friedman, et al, 2003; Ketyon, et al, 2004). For closed-ended surveys or interviews where participants are forced to choose one choice, then the collected data is immediately ready for inter-rater checks (although quantitative checks often produce lower reliability scores, especially when the likert scale is used) (Friedman, et al, 2003).

To compute inter-rater reliability in quantitative studies (where closed-answer question data is collected using a likert scale, a series of options, or yes/no answers), follow these steps to determine Cohen's kappa (1960), a statistical measure determining inter-rater reliability:

1. **Arrange the responses from the two different surveys/interviews into a contingency table.** This means you will create a table that demonstrates, essentially, how many of the answers agreed and how many answers disagreed (and how much they disagreed, even). For example, if two different survey/interview administrators asked ten yes or no questions, their answers would first be laid out and observed:

Question Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Interviewer #1 ➔	**Y**	**N**	**Y**	**N**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**
Interviewer #2 ➔	**Y**	**N**	**Y**	**N**	**Y**	**Y**	**Y**	**N**	**Y**	**N**

From this data, a contingency table would be created:

	RATER #1 (Going across)	
RATER #2 (Going down)	**YES**	**NO**
YES	**6**	**0**
NO	**2**	**2**

Notice that the number six (6) is entered in the first column because when looking at the answers there were six times when both interviewers found a *YES* answer to the same question. Accordingly, they are placed where the two *YES* answers overlap in the table (with the *YES* going across the top of the table representing Rater/Interviewer #1 and the *YES* going down the left side of the table representing Rater/Interviewer #2). A zero (0) is entered in the second column in the first row because for that particular intersection in the table there were no occurrences (that is, Interviewer/Rater #1 never found a *NO* answer when Interviewer/Rater #2 found a *YES*). The number two (2) is entered in the first column of the second row since Interviewer/Rater #1 found a *YES* answer two times when Interviewer/Rater #2 found a *NO*; and a two (2) is entered in the second column of the second row because both Interviewer/Rater #1 and Interviewer/Rater #2 found *NO* answers to the same question two different times.

NOTE: *It is important to consider that the above table is for a YES/NO type survey. If a different number of answers are available for the questions in a survey, then the number of answers should be taken into consideration in creating the table. For instance, if a five*

question likert-scale were used in a survey/interview, then the table would have five rows and five columns (and all answers would be placed into the table accordingly).

2. **Sum the row and column totals for the items.** To find the sum for the first row in the previous example, the number six would be added to the number zero for a first row total of six. The number two would be added to the number two for a second row total of four. Then the columns would be added. The first column would find six being added to two for a total of eight; and the second column would find zero being added to two for a total of two.
3. **Add the respective sums from step two together.** For the running example, six (first row total) would be added to four (second row total) for a row total of ten (10). Eight (first column total) would be added to two (second column total) for a column total of ten (10). At this point, it can be determined whether the data has been entered and computed correctly by whether or not the row total matches the column total. In the case of this example, it can be seen that the data seems to be in order since both the row and column total equal ten.
4. **Add all of the agreement cells from the contingency table together.** In the running example, this would lead to six being added to two for a total of eight because there were six times where the *YES* answers matched from both interviewers/raters (as designated by the first column in the first row) and two times where the *NO* answers matched from both interviewers/raters (as designated by the second column in the second row). The sum of agreement then, and the answer to this step, would be eight (8). The agreement cells will

always appear in a diagonal pattern across the chart – so, for instance, if participants had five possibilities for answers then there should be five cells going across and down the chart in a diagonal pattern that will be added.

***NOTE**: At this point simple agreement can be computed by dividing the answer in step four by the answer in step five. In the case of this example, that would lead to eight being divided by ten for a result of 0.8. This number would be rejected by many researchers, however, since it does not take into account the probability that some of these agreements in answers could have been by chance. That is why the rest of the steps must be followed to determine a more accurate assessment of inter-rater reliability.*

5. **Compute the expected frequency for each of the agreement cells appearing in the diagonal pattern going across the chart.** To do this, find the row total for the first agreement cell (row one column one) and multiply that by the column total for the same cell. Divide this by the total number possible for all answers (this is the row/column total from step three). So, for this example, first the cell containing the number six would be located (since it is the first agreement cell located in row one column one) and the column and row totals would be multiplied by each other (these were found in step two) and then divided by the total: 6 x 8 =48 ➔ 48/10=4.8. The next diagonal cell (one over to the right and one down) is the next row to be computed: 2 x 4=8 ➔ 8/10=0.8. Since this is the final cell in the diagonal, this is the final computation that needs to be made in this step for the sample problem; however, if more answers were possible, then the step would be repeated as many times as there are answers. For a five answer likert scale, for instance,

the process would be repeated for five agreement cells going across the chart diagonally in order to consider how those answers matched up and provide a full account of inter-rater reliability.

6. **Add all of the expected frequencies found in step five together.** This represents the expected frequencies of agreement by chance. For the example used in this literature review, that would be 4.8 + 0.8 for a sum of 5.6. For a five answer likert scale, all five of the totals found in step five would be added together.
7. **Compute kappa.** To do this, take the answer from step four and subtract the answer from step six. Place the result of that computation aside. Then take the total number of items from the survey/interview and subtract the answer from step six. After this has been completed, take the first computation from this step (the one that was set aside) and divide it by the second computation from this step. The resulting computation represents kappa. For the running example that has been provided in this literature review, it would look like this: 8 - 5.6 = 2.4; 10 – 5.6 = 4.4 ➔ 2.4/4.4 = 0.545.
8. **Determine whether the reliability rate is satisfactory**. If kappa is at 0.7 or higher, then the inter-rater reliability rate is generally considered satisfactory (CITE). If not, then it is often rejected.

What to do if inter-rater reliability is not at an appropriate level

Unfortunately, if inter-rater reliability is not at the appropriate level (generally 0.7) then it is often recommended that the data be thrown out (Krippendorf, 2004a). In cases such as these, it is often wise to administer an additional data collection so a third

set of information can compared to the other collected data (and calculated against both in order to determine if an acceptable inter-rater reliability level has been achieved with either of the previous data collecting attempts). If many cases of inter-rater issues are occurring, then the data from these cases can often be observed in order to determine what the problem may be (Keyton, et al, 2004). If data has been prepared for inter-rater checks from qualitative collection measures, for instance, the coding scheme used to prepare the data may be examined.

It may also be helpful to check with the person who coded the data to make sure they understood the coding procedure (Ketyon, et al, 2004). This inquiry can also include questions about whether they became fatigued during the coding process (often those coding large sets of information tend to make more mistakes) and whether or not they agree with the process selected for coding (Keyton, et al, 2004). In some cases a *rogue coder* may be the culprit for failure to achieve inter-rater reliability (Neuendorf, 2002). Rogue coders are coders who disapprove of the methods used for analyzing the data and who assert their own coding paradigms. Facilitators of projects may also be to blame for the low inter-rater reliability, especially if they have rushed the process (causing rushed and hasty coding), required one individual to code a large amount of information (leading to fatigue), or if the administrator has tampered with the data (Keyton, et al, 2004).

References

Capwell, A. (1997). *Chick flicks: An analysis of self-disclosure in friendships*. Cleveland: Cleveland State.

Cohen, J. (1960). Kappa test: A coefficient of agreement for nominal scales. *Education Psychology Measures, 20*, 37-46.

Friedman, P. G., Chidester, P. J., Kidd, M. A., Lewis, J. L., Manning, J. M., Morris, T. M., Pilgram, M. D., Richards, K., Menzie, K., & Bell, J. (2003). Analysis of ethnographic interview research procedures in communication studies: Prevailing norms and exciting innovations. *National Communication Association,* Miami, FL.

Green, B. (2004). Personal construct psychology and content analysis. *Personal Construct Theory and Practice, 1*, 82-91.

Keyton, J., King, T., Mabachi, N. M., Manning, J., Leonard, L. L., & Schill, D. (2004). *Content analysis procedure book*. Lawrence, KS: University of Kansas.

Krippendorf, K. (2004a). Content analysis: An introduction to its methodology. Thousand Oaks, CA: Sage.

Krippendorf, K. (2004b). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*, 411-433.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.