CHAPTER 8

Inter-Rater Reliability: Conditional Analysis

OBJECTIVE

This chapter introduces a number of measures of validity (as opposed to the measures of reliability discussed in the past few chapters), and describes statistical techniques for analyzing the extent of agreement among raters conditionally upon the subject membership in a specific category. The specific category used in the conditioning, could be the subject's "true" category if it exists, or the category into which one rater classified the subject. Conditional analysis offers the advantage of evaluating the extent of agreement among raters for a subgroup of subjects known to belong to a particular category. This analysis reduces the dependency of the agreement coefficient on trait prevalence and on the distribution of subjects across categories, and can help identify a special group of subjects where agreement is hard to reach. Methods for computing the variances associated with these conditional measures are also discussed.

Contents

8.1 *Overview* 241
8.2 *Two-Rater Conditional Agreement for ACM Studies* 245
8.2.1 *Basic Conditional Probabilities for ACM Studies* 246
8.2.2 *Conditional Reliability for 2 Raters in ACM Reliability Studies* . . 249
8.2.3 *Unconditional Validity Coefficient for 2 Raters in ACM Studies* . 258
8.2.4 *Concluding Remarks on Section 8.2* 261
8.3 *Multiple-Rater Coefficients for ACM Studies* 261
8.3.1 *Validity Coefficients for 3 Raters or More* 262
8.3.2 *Conditional Agreement Coefficients for three Raters or More* . . . 270
8.4 *Conditional Agreement in RCM Studies* 278
8.5 *Concluding Remarks* 282

8.1 Overview

The focus of the past few chapters was put on the study of statistical techniques for quantifying the extent of agreement among raters using all subjects that received a rating. However, there are situations in practice where researchers may want to study inter-rater reliability based on a restricted pool of subjects that meet specific characteristics. For example, a panel of experts may provide on each subject, a consensus rating that represents supposedly its "true" category membership. It may then be of interest for the researcher to investigate inter-rater reliability with respect to subjects that belong to a specific category. This allows researchers to identify groups of subjects where agreement is difficult to achieve. In addition to the panel of experts, groups of subjects may be defined using the ratings of one of the judges who participated in the inter-rater reliability study. Fleiss (1971) discussed this issue, and proposed an adaptation of his generalized kappa coefficient to address it. Referring to the study of agreement on a restricted pool of subjects as conditional agreement, Light (1971) pointed out that this technique "can contribute additional insight to a data analysis."

Before introducing conditional agreement coefficients, I need to present the different ways that a subject's category membership can be defined. Category membership in a nominal-scale inter-rater reliability study can be defined in two possible ways:

- A clear-cut operational definition exists. It defines a deterministic relationship between subjects and categories. For example, a thermometer could be used to measure the body temperature of patients before classifying them into one of the three groups labeled as "Low," "Normal," and "High" defined according to pre-specified intervals.

- The rater chooses a category based on personal preferences, which are expected to vary from rater to rater. The subject category membership in this case, can only be determined with respect to one particular rater's preferences.

In studies where an operational definition of category membership exists, a body of experts often reach a consensus around one category that is then labeled as the subject "true" category. Such a consensus allows the researcher to determine whether agreement reached by the raters is associated with the "correct" category also known in the literature as the "Gold Standard." The existence of gold-standard scores also makes it possible to pinpoint problem categories on which raters may have difficulties reaching agreement. Evaluating the validity of agreement and identifying problem categories are the two main goals of this chapter.

Gold-standard scores - when they exist - are seen as the "true" scores of the subjects they are associated with. These subjects are said to have an "Absolute Category Membership" (or ACM), because the "true" category is tied to the subject,

and not to the rater. When categories are tied to the raters rather than to the subjects, then classification depends more on each rater's preferences. No operational definition exists linking subjects to specific categories. The subjects are then said to have a "Relative Category Membership" (or RCM). Marginal probabilities in this case are often seen as fixed since raters generally have known preferences. Inter-rater reliability coefficients for RCM ratings could be further analyzed by considering only subjects that one rater classified into a specific category.

Let us consider an experiment involving the chart review of pregnant women who enter a hospital's emergency service with an abdominal pain or a vaginal bleeding. Two chart abstractors named "Abstractor 1", and "Abstractor 2" must assign 100 patients into one of the following two categories:

- Ectopic Pregnancy (EP),
- Intrauterine Pregnancy (IP).

An experienced chart reviewer also categorizes the same 100 patients into what is considered to be the "True" categories. The results of this experiment are summarized in Table 8.1, which shows the distribution of subjects by rater and by "true" pregnancy type as determined by the expert abstractor. It follows that both abstractors categorized 15 pregnancies as Ectopic, of which 13 are actual "True" Ectopic pregnancies while the other 2 are "True" Intrauterine pregnancies. Moreover, 14 of the 18 pregnancies that abstractor 2 classified as Ectopic are "True" Ectopic pregnancies while the remaining 4 are "True[1]" **IP**s.

It is natural for a researcher to want to know whether abstractors are more likely to agree while rating a "True" Ectopic pregnancy than while rating a "True" **IP**. Agreement in this case must be evaluated conditionally upon the true nature of the pregnancy. The statistical notion of conditioning applies in this case by restricting the pool of female subjects to be rated to those who carry a specific pregnancy type that is of interest. For example, the conditional percent agreement given a true **EP** is $p_{a|\text{EP}} = (13+2)/20 = 0.75$. That is, abstractors agreed to classify 13 of the 20 true **EP**s as EPs, 2 as IPs. The denominator in this case is 20, because the analysis is limited to the 20 true **EP**s in the study group as shown in Table 8.1. If the analysis was unconditional then the denominator would be 100, representing the entire group of subjects that were part of the inter-rater reliability experiment.

Conditional analysis always involves two events:

- The Conditioning Event that must be considered as having been observed.

[1]For convenience, I use the boldface to designate the true category of subjects, and the regular font to designate the category into which a rater chooses to classify a subject.

- The Conditioned Event whose chance of occurrence may have been affected by the conditioning event.

The *Conditioning Event* aims at restricting the list of possibilities, and at specifying the information the researcher should consider as given a priori, before the propensity of occurrence of the *Conditioned Event* can be evaluated. For example, when evaluating the conditional percent agreement given the true **EP**, the conditioning event is that the "true pregnancy type is known to be **EP**," and it is assumed to be given. The objective here is to evaluate the extent to which this conditioning event, if realized, would affect the chance of occurrence of the conditioned event defined as "Abstractors 1 and 2 agree given that the true pregnancy type of the patient being rated is **EP**."

Table 8.1: Distribution of 100 Emergency Room Pregnant Women by Abstractor and Type of Pregnancy

	Abstractor 2								
	True Category								
Abstractor	**EP**			**IP**			All Subjects		
1	EP	IP	Total	EP	IP	Total	EP	IP	Total
EP	13	4	17	2	3	5	15	7	22
IP	1	2	3	2	73	75	3	75	78
Total	14	6	20	4	76	80	18	82	100

While the 2 true **EP**s classified as IPs by both abstractors would increase reliability, they are expected to decrease validity. They should not be considered as agreement if validity is being measured. Validity analysis, which is addressed in section 8.2, deals with research questions such as "*Are abstractors more likely to positively detect true Ectopic pregnancies than to positively detect true **IP**s?*" Being able to identify categories where agreement is more easily reached will also pinpoint other problem categories that should be the focus of further abstractor training. Conditional analysis could also lead to a possible modification of some categories that abstractors deem unclear. This analysis is carried out by breaking down an inter-rater reliability coefficient $\widehat{\kappa}_{\mathrm{X}}$ into 2 components $\widehat{\kappa}_{\mathrm{X}|\mathrm{EP}}$, and $\widehat{\kappa}_{\mathrm{X}|\mathrm{IP}}$ associated with the 2 response categories. These two types of conditional inter-rater reliability coefficients are discussed in greater details in section 8.2.

Let us turn to reliability experiments where the notion of "True" score is nonexistent. Consider Tables 8.2 and 8.3 where two raters classified 100 garments into one

of two categories "Good" (or G) and "Bad" (or B). The rating process in this case depends more on the rater's personal taste than on the very nature of the object being rated. Even though the garment type still affects the rater's choice, the relationship between the two remains under the control of each rater. Consequently, the rater's marginal probabilities can be considered fixed for a given collection of garments, making them sufficiently important to play a pivotal role in the interpretation of the inter-rater reliability magnitude.

Based on the AC_1 coefficient, the extent of agreement between raters A and B is evaluated at 0.597 and that between raters C and D is evaluated at 0.31. Although AC_1 indicates that raters A and B are more in agreement than raters C and D by a ratio of almost 2 to 1, a close look at both Tables 8.2 and 8.3 suggests that given the observed marginal probabilities[2], raters A and B have reached the minimum agreement possible while raters C and D have reached the maximum agreement possible. Therefore, one may argue that raters C and D are more in agreement than raters A and B (in a relative sense) given their respective rating propensities.

Table 8.2: *Distributions of 100 Garments by Rater (A/B) & Quality of Garment*

Rater A's	Rater B's Scores		
Scores	B	G	Total
B	70	15	85
G	15	0	15
Total	85	15	100

Table 8.3: *Distributions of 100 Garments by Rater (C/D) & Quality of Garment*

Rater C's	Rater D's Scores		
Scores	B	G	Total
B	50	40	90
G	0	10	10
Total	50	50	100

One objective of this chapter is to present ways of evaluating the extent of agreement among raters conditionally on their marginal probabilities. Conditional analysis of raters' agreement are generally recommended if the researcher wants to study the effect of categories on the agreement level, or if comparison between groups of raters is of interest and marginal probabilities can be assumed fixed. Readers should know that all numeric examples presented in this chapter are also done in the Excel workbook:

www.agreestat.com/books/cac5/chapter8/chap8calculations.xlsx

In section 8.2, I will introduce the notions of conditional inter-rater reliability coefficient, and that of validity coefficient in the simple context where the number of raters is limited to 2. This section is written as a convenient way to expose the reader

[2] i.e. the marginal probabilities (0.85 and 0.15 for rater A for instance) are considered fixed.

to these new concepts using a simple inter-rater reliability experiment. Ratings from Table 8.1 will be used regularly for illustration purposes. These same notions will be expanded to the more general and complex inter-rater reliability experiments that involve 3 raters or more.

8.2 Agreement Coefficient for two Raters in ACM Studies

Throughout this section, a k-subject refers to any subject whose "True" response category is k, and I assume that each of the two raters who participated in the experiment rated all n subjects. If some subjects were rated by only one of the two raters then the practitioner faces the problem of missing ratings. In this case, I recommend using alternative methods described in section 8.3 and developed for the more general setting involving 3 raters or more.

The rating of subjects is said to be reliable when the raters consistently classify subjects into the same categories; but will be valid only if the subjects are consistently classified into their correct category by the raters. That is,

Validity = Reliability + Exactness.

In statistical jargon, you would say "Validity = Precision + Unbiasedness." A precise process hits its target, whereas a valid process hits the right target.

In this section, I will introduce reliability and validity measures. A measure of reliability in the case of two raters for example, quantifies how often both raters classify subjects into the same category[3] (whether it is the "true" category or not). A measure of validity on the other hand quantifies the extent to which both raters classify subjects into their respective true categories. Because validity is a more stringent condition than reliability, validity coefficients are expected to be smaller than reliability coefficients. When the pool of subjects used to evaluate reliability or validity is restricted to k-subjects only, one obtains conditional reliability and conditional validity coefficients given category k. The use of all subjects for which ratings have been collected would lead to the unconditional coefficients that were extensively discussed in the past few chapters.

Throughout this chapter, p_k represents the probability that the true category of a randomly selected subject is k. For all practical purposes, it represents the relative number of k-subjects in the sample. Referring to the emergency room pregnancy data of Table 8.1, the probability of true Ectopic pregnancy is $p_{\text{EP}} = (14+6)/100 = 0.20$, while that of true Intrauterine Pregnancy is $p_{\text{IP}} = (4+76)/100 = 0.80$. In subsection 8.2.1, I will introduce the basic elements needed for calculating reliability and validity coefficients conditionally upon the true category.

[3]This frequency is possibly adjusted for chance agreement.

8.2.1 *Basic Conditional Probabilities for ACM Studies*

Conditional analysis of ACM data requires the use of various basic probabilities that I will define in this section[4]. A typical conditional agreement coefficient given a true category k will take a general form such as $\widehat{\kappa}_{\mathsf{x}|k_0} = (p_{a|k_0} - p_{e|k_0})/(1 - p_{e|k_0})$, where x could be a character such as "C" if it is the conditional Cohen kappa, or G if it is the conditional Gwet's AC_1. Moreover, $p_{a|k_0}$ and $p_{e|k_0}$ are the conditional percent agreement and conditional percent chance agreement given the true category k_0. In these notations, k_0 after the "|" sign is the symbolic representation of the conditioning event (i.e. "the true category of the subject to be rated is k_0"). The subscripts $a|k_0$ or $e|k_0$ on the other hand are the symbolic representations of the following 2 conditioned events:

- Both raters agree given the rating of k_0-subjects.
- Both raters agree by pure chance given the rating of k_0-subjects.

The basic probabilities discussed in this section are the building blocs for calculating the two core conditional probabilities $p_{a|k_0}$ and $p_{e|k_0}$. To compute the conditional percent agreement, one needs to compute 2 types of probabilities:

- The *joint classification probabilities* $p_{kl}^{(k_0)}$ *associated with categories* k *(for rater 1) and* l *(for rater 2) as a function of the "true" category* k_0. It represents the likelihood that a randomly selected subject turns out to be a k_0-subject classified into categories k and l by raters 1 and 2 respectively.
- The *conditional classification probabilities* $p_{kl|k_0}$ *associated with categories* k *(for rater 1) and* l *(for rater 2) given the "true" category* k_0. This is the conditional probability that raters 1 and 2 use categories k and l given that it is a k_0-subject that is being rated.

To illustrate how these probabilities are calculated, I will occasionally use Table 8.1 data. Throughout this section the true category is always assumed to be k_0.

► *Conditional classification probabilities* $p_{k|k_0}$

For any given category k_0, I want to compute the likelihood that two raters classify a subject into categories k and l, conditionally upon the subject's "true"

[4]One may wonder why these basic conditional probabilities are even needed, when the relevant pool of subjects can be identified first, before applying conventional methods to it. This basic approach may work with one major limitation. It will fail to adequately compute the standard errors. It is because the size of the relevant pool of subjects is random and cannot be determined prior to the experiment being conducted. The use of conditional probabilities as defined in this section resolves that problem.

category being k_0. This probability will be denoted by $p_{kl|k_0}$ (read p, k, l given k_0). For all practical purposes, this is the relative number of subjects that the two raters classified into categories k and l calculated based solely on k_0-subjects.

Using the data from Table 8.1, I computed the classification probabilities conditionally upon the "true" category being "Ectopic Pregnancy" ($k_0 = \mathsf{EP}$) as shown in Table 8.4. Only subjects whose true category matches the conditioning category are considered in this analysis. It follows from Table 8.4 that based on the 20 true EP patients, the two abstractors will agree on the EP category 65% of the times. Note that similar classification probabilities can be calculated conditionally upon the "true" category being "Intrauterine Pregnancy" (IP) as shown in Table 8.5.

Table 8.4: Computing conditional probabilities given the true category $k_0 = \mathsf{EP}$

Distribution of EP-Patients

	Abstractor 2		
Abstractor 1	EP	IP	Total
EP	13	4	17
IP	1	2	3
Total	14	6	20

Conditional probabilities[a] $p_{kl|k_0}$ given the k_0-patients where $k_0 = \mathsf{EP}$

Abstractor 1	Abstractor 2 (l)		
(k)	$l = \mathsf{EP}$	$l = \mathsf{IP}$	$p_{k+\|k_0}$
EP	0.65	0.20	0.85
IP	0.05	0.10	0.15
$p_{+l\|k_0}$	0.70	0.30	1.00

[a]Calculated by dividing the numbers on the left by 20 (the number of EP-subjects)

► *Joint classification probabilities $p_{kl}^{(k_0)}$ as a function of the "true" category k_0*

For any given category k_0, I want to compute the likelihood that two raters classify a k_0-subject into categories k and l. To compute these joint probabilities, only the ratings associated with k_0-subjects are considered. However, their frequencies are evaluated with respect to the entire pool of subjects that participated in the experiment. It is the case when probabilities are evaluated without condition. A typical joint classification probability will be denoted by $p_{kl}^{(k_0)}$ (read p, k, l of k_0). For all practical purposes, this is the number of k_0-subjects that were classified into categories k and l by the two raters, relative to the total number of subjects that participated in the experiment.

Table 8.5: Computing conditional probabilities given the true category $k_0 = \mathsf{IP}$

Distribution of IP-Patients

	Abstractor 2		
Abstractor 1	EP	IP	Total
EP	2	3	5
IP	2	73	75
Total	4	76	80

Conditional probabilities[a] $p_{kl|k_0}$ of k_0-Patients where $k_0 = \mathsf{IP}$

Abstractor 1	Abstractor 2 (l)		
(k)	$l = \mathsf{EP}$	$l = \mathsf{IP}$	$p_{k+\|k_0}$
EP	0.025	0.0375	0.0625
IP	0.025	0.9125	0.9375
$p_{+l\|k_0}$	0.05	0.95	1.00

[a]Calculated by dividing the numbers on the left by 20 (the number of EP-subjects)

Using the data from Table 8.1, I computed the joint classification probabilities for the "true" category "Ectopic Pregnancy" ($k_0 = \mathsf{EP}$) as shown in Table 8.6. Only subjects whose true category is EP are considered in the analysis, except that the denominator this time is 100 the total number of patients in the experiment. It follows from Table 8.6 that the patients whose pregnancy was rightfully classified by both abstractors as an EP represented 13% of all patients. Similar joint scoring probabilities can be calculated for the "true" Intrauterine Pregnancy (IP) category as shown in Table 8.7.

Table 8.6: Computing joint probabilities for the true category $k_0 = \mathsf{EP}$

Distribution of EP_{T} Patients

	Abstractor 2		
Abstractor 1	EP	IP	Total
EP	13	4	17
IP	1	2	3
Total	14	6	20

Joint probabilities[a] $p_{kl}^{(k_0)}$ for the k_0-patients where $k_0 = \mathsf{EP}_{\mathrm{T}}$

Abstractor 1	Abstractor 2 (l)		
(k)	$l = \mathsf{EP}$	$l = \mathsf{IP}$	$p_{k+}^{(k_0)}$
EP	0.13	0.04	0.17
IP	0.01	0.02	0.03
$p_{+l}^{(k_0)}$	0.14	0.06	0.20

[a]Obtained by dividing the numbers on the left by 100 (the total number of subjects)

Table 8.7: Computing joint probabilities for the true category $k_0 = \mathsf{IP}$

Distribution of IP_T Patients

	Abstractor 2		
Abstractor 1	EP	IP	Total
EP	2	3	5
IP	2	73	75
Total	4	76	80

Joint probabilities[a] $p_{kl}^{(k_0)}$ for the k_0-patients where $k_0 = \mathsf{IP}_\mathrm{T}$

Abstractor 1	Abstractor 2 (l)		
(k)	$l = \mathsf{EP}$	$l = \mathsf{IP}$	$p_{k+}^{(k_0)}$
EP	0.02	0.03	0.05
IP	0.02	0.73	0.75
$p_{+l\|k_0}$	0.04	0.76	0.80

[a]Calculated by dividing the numbers on the left by 100 (the total number of subjects)

► *Category Use Frequency $\pi_{k|k_0}$, Conditionally upon the True Category*

I want to be able to compute the relative number of times a specific category is used by raters when only a restricted group of subjects is selected based on their true category membership. For example, given that only *intrauterine pregnancies* (i.e. condition) are rated, the relative number of pregnancies categorized as Ectopic is denoted by $\pi_{\mathrm{E}|\mathrm{I}}$ (read pi, E given I). This represents the relative frequency of use of the ectopic pregnancy type when rating only true intrauterine pregnancies. Because of the specific condition imposed on the restricted pool of subjects to be rated, $\pi_{\mathrm{E}|\mathrm{I}}$ is referred to as the conditional relative frequency of ectopic given intrauterine. If there is no restriction on the pool of subjects to be rated, the relative frequency is said to be unconditional. More generally, let $\pi_{k|k_0}$ be the relative frequency with which category k is used *when rating k_0-subjects only*. This quantity is calculated as follows:

$$\pi_{k|k_0} = (p_{k+|k_0} + p_{+k|k_0})/2, \tag{8.2.1}$$

where $p_{k+|k_0}$ and $p_{+k|k_0}$ are the marginal conditional probabilities calculated in Tables 8.4 and 8.5.

8.2.2 *Conditional Reliability for 2 Raters in ACM Reliability Studies*

Unconditional reliability coefficients for ACM studies are identical to the regular inter-rater reliability coefficients covered in Part II of this book, and will not be discussed any further in this section. This section is devoted entirely to the

study of conditional reliability coefficients given the subject's "true" membership category. For example, a conditional reliability coefficient may quantify the extent of agreement among two abstractors under the condition that the woman's "true" pregnancy type is Ectopic. In this case, "Ectopic" is the conditioning category or the conditioning group.

> *As previously indicated, the primary purpose of conditioning is to have a reliability measure that is not affected by the distribution of women across the different types of pregnancy. Such a conditional reliability coefficient will facilitate comparison between reliability studies based on populations with different prevalence rates of Ectopic pregnancies. It also helps identify the specific categories where agreement is hard to reach.*

Although the conditional reliability coefficient can always be defined as seen in the previous paragraph, its actual computation requires the presence of a few subjects in the conditioning category. That is, if the gold standard does not include any subject into a category, then that category cannot be used for conditional analysis.

In this section, I will introduce 6 conditional agreement coefficients given a particular true category k_0 assuming it has been used at least once by the gold standard[5]. The 6 conditional agreement coefficients are the *conditional percent agreement*, *Gwet's* AC_2, *Cohen's Kappa*, *Scott's Pi*, *Krippendorff's alpha*, and *Brennan-Prediger*.

▶ Conditional Percent Agreement Coefficient

The conditional "raw" percent agreement given the "true" membership category k_0 and denoted by $p_{a|k_0}$ represents the conditional likelihood that two raters agree based on the pool of a k_0-subjects alone, and is evaluated as the weighted sum of the conditional probabilities that both raters classify a subject into two categories k and l. More formally, it is expressed as follows:

$$p_{a|k_0} = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl}\pi_{kl|k_0}, \tag{8.2.2}$$

where $\pi_{kl|k_0} = (p_{kl|k_0} + p_{lk|k_0})/2$ and $p_{kl|k_0}$ is the conditional probability that raters 1 and 2 classify a subject into categories k and l given the pool of k_0-subjects. Note that if the weight matrix (w_{kl}) is symmetrical - which is often the case - then equation 8.2.2 can be rewritten in a more simplified form as

[5] The category k_0 could designate EP or IP pregnancy for example.

follows:

$$p_{a|k_0} = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl}p_{kl|k_0}. \tag{8.2.3}$$

▶ Conditional AC_2 Coefficient

Let k_0 be an arbitrary category (e.g. k_0 could designate EP or IP pregnancy). The conditional AC_2 reliability coefficient given a "true" membership category k_0 is denoted by $\kappa_{G|k_0}$ (where G stands for "Gwet"). It is defined as the ratio of the percent of agreement for cause to the likelihood of no chance agreement[6]; both components being evaluated under the condition that k_0 is the true category of the subject being rated. More formally,

$$\begin{aligned} \widehat{\kappa}_{G|k_0} &= (p_{a|k_0} - p_{e|k_0})/(1 - p_{e|k_0}), \\ \textit{where } p_{e|k_0} &= \frac{T_w}{q(q-1)}\sum_{k=1}^{q}\pi_{k|k_0}(1-\pi_{k|k_0}). \end{aligned} \tag{8.2.4}$$

The conditional percent agreement $p_{a|k_0}$ is given by equation 8.2.2, and $p_{e|k_0}$ is the conditional percent chance agreement representing the conditional likelihood that two raters agree by pure chance based on the pool of k_0-subjects. The conditional probability $\pi_{k|k_0}$ used to compute the percent chance agreement is defined in equation 8.2.1.

Let us look more closely, step by step at the calculation of the conditional AC_2 coefficient using the rating data of Table 8.1. Since this reliability experiment involves two pregnancy types, I will need to compute two conditional AC_2 coefficients, which are $AC_{2|E}$ (for Ectopic), and $AC_{2|I}$ (for Intrauterine). Each of these conditional agreement coefficients require the calculation of a conditional percent agreement and a conditional percent chance agreement. I will confine myself to the calculation of unweighted agreement coefficients (i.e. Identity weights will be used in equation 8.2.4). Therefore, it is actually the conditional AC_1 coefficient that will be calculated as opposed to AC_2.

[6] *Note that the possibility of no chance agreement includes agreement for cause, and disagreements of all kinds as well.*

Calculating the conditional percent agreement values $p_{a|\text{E}}$ and $p_{a|\text{I}}$

Table 8.8 outlines the necessary steps for computing the two conditional percent agreement values $p_{a|\text{E}} = 0.75$, and $p_{a|\text{I}} = 0.9375$ (see the last row of step 2 in Table 8.8). The first input values going into these calculations are shown in step 1, and represent the agreement probabilities calculated separately for each true category and each reported category.

- The value $p_{a|\text{E}} = 0.75$ is obtained by summing the two conditional agreement probabilities $p_{\text{EE}|\text{E}} = 0.65$, and $p_{\text{II}|\text{E}} = 0.10$. Note that 0.65 represents the conditional relative number of times both abstractors would assign a pregnancy to the Ectopic category given that it is indeed an Ectopic pregnancy, whereas 0.10 is the conditional relative number of times both abstractors would assign a pregnancy to the intrauterine category given that it is a true Ectopic pregnancy.
- The conditional relative frequency $p_{\text{EE}|\text{E}} = 0.65$ is calculated as the ratio of the (unconditional) relative frequency $p_{\text{EE}}^{(\text{E})} = 0.13$ to the prevalence of the Ectopic pregnancy $p_{\text{E}} = 0.20$. The conditional relative frequency $p_{\text{II}|\text{E}} = 0.10$ on the other hand, is the ratio of the (unconditional) relative frequency $p_{\text{II}}^{(\text{E})} = 0.02$ to the prevalence of the Ectopic pregnancy $p_{\text{E}} = 0.20$.

The value $p_{a|\text{I}} = 0.9375$ can be calculated using the same procedure described above for $p_{a|\text{E}}$.

Calculating the conditional chance-agreement probabilities $p_{e|\text{E}}$ and $p_{e|\text{I}}$

Table 8.9 shows the steps for calculating the conditional percent chance agreement for each true category.

- The percent chance agreement conditionally upon the true ectopic type pregnancy is given by $p_{e|\text{E}} = 0.34875$, whereas the percent chance agreement conditionally upon the true intrauterine type pregnancy is given by, $p_{e|\text{I}} = 0.106172$. To get these two numbers, one needs to start with the four conditional marginal probabilities of Table 8.4 (i.e. $p_{k+|k_0}$ and $p_{+l|k_0}$), and the four conditional marginal probabilities of Table 8.5. These marginal probabilities are needed to compute the conditional probabilities $\pi_{k|k_0}$ that the reported category k is used (by either rater) given the k_0 subjects. I refer to $\pi_{k|k_0}$ as the conditional use probability of k given k_0.
- The next step is to compute the products of the conditional use probabilities and their complements (i.e. $\pi_{k|k_0} \times (1 - \pi_{k|k_0})$) as shown in the rightmost column of Table 8.9.

- Afterwards, you need to compute the "Total" row by summing the numbers column-wise. At this stage, the conditional chance agreement probabilities are calculated as prescribed by equation 8.2.4. Since we are computing unweighted coefficients, Identity weights are used and their total is q. Consequently the ratio T_w to $q(q-1)$ equals $q/(q\times(q-1)) = 1/(q-1) = 1/(2-1) = 1$. Therefore the conditional chance agreement probabilities are obtained by dividing the "Total" row by 1, which does not change their values.

Table 8.8: Calculation of the conditional percent agreement based on the true category

Step 1: Calculating Agreement Probabilities by True Category $\left(p_{ll}^{(k)}\right)$

Pregnancy Type Agreed Upon (l)	"True" Pregnancy Type (k)	
	$k = \mathsf{EP}_{\mathrm{T}}$	$k = \mathsf{IP}_{\mathrm{T}}$
EP	$13/100 = \mathbf{0.13}$	$2/100 = \mathbf{0.02}$
IP	$2/100 = \mathbf{0.02}$	$73/100 = \mathbf{0.73}$
"True" Category Prevalence	$20/100 = \mathbf{0.20}$	$80/100 = \mathbf{0.80}$

Step 2: Calculating Conditional Agreement Probabilities $(p_{ll|k})$

Pregnancy Type Agreed Upon (l)	"True" Pregnancy Type (k)	
	$k = \mathsf{EP}_{\mathrm{T}}$	$k = \mathsf{IP}_{\mathrm{T}}$
EP	$0.13/0.20 = \mathbf{0.65}$	$0.02/0.80 = \mathbf{0.025}$
IP	$0.02/0.20 = \mathbf{0.10}$	$0.73/0.80 = \mathbf{0.9125}$
Conditional % Agreement ($p_{a|k}$)	$0.65 + 0.10 = \boxed{\mathbf{0.75}}$	$0.025 + 0.9125 = \boxed{\mathbf{0.9375}}$

The last rows of Tables 8.8 and 8.9 contain the two percent agreement estimates, and two percent chance-agreement estimates needed to compute the conditional AC_1 coefficients named $\widehat{\kappa}_{\mathrm{G|E}}$ (for ectopic), and $\widehat{\kappa}_{\mathrm{G|I}}$ (for intrauterine). Therefore,

$$\begin{aligned}\widehat{\kappa}_{\mathrm{G|E}} &= \big(p_{a|\mathrm{E}} - p_{e|\mathrm{E}}\big)/\big(1 - p_{e|\mathrm{E}}\big),\\ &= (0.75 - 0.34875)/(1 - 0.34875) = 0.616.\\ \widehat{\kappa}_{\mathrm{G|I}} &= \big(p_{a|\mathrm{I}} - p_{e|\mathrm{I}}\big)/\big(1 - p_{e|\mathrm{E}}\big),\\ &= (0.9375 - 0.106172)/(1 - 0.106172) = 0.93.\end{aligned}$$

> *It follows from these conditional agreement coefficients that based on the AC_1 coefficient, the two abstractors would agree more often if they rate true intrauterine pregnancies ($AC_{1|\mathrm{I}} = 0.93$) than if they rate true ectopic pregnancies ($AC_{1|\mathrm{E}} = 0.616$). If the abstractors must be given further training, the training session would be more productive if it focuses on improving the rating of true ectopic pregnancies. One should note that the conditional agreement coefficients discussed here quantify the extent to which abstractors classify subjects into the same category. These are agreement coefficients. They do not evaluate the extent to which abstractors classify subjects into their correct and true category. Only validity coefficients discussed in sections 8.2.3 and 8.3.1 will quantify the propensity for raters to agree on the subject's true category.*

Table 8.9: Steps for computing the AC_1's conditional chance-agreement percentages given the true category k_0

Reported Pregnancy Type (k)	$\pi_{k\|k_0}$ = conditional use probabilities given k_0-subjects		$\pi_{k\|k_0} \times (1 - \pi_{k\|k_0})$	
	$k_0 = \mathrm{EP_T}$	$k_0 = \mathrm{IP_T}$	$k_0 = \mathrm{EP_T}$	$k_0 = \mathrm{IP_T}$
$k = \mathrm{EP}$	0.775	0.05625	0.174375	0.053086
$k = \mathrm{IP}$	0.225	0.94375	0.174375	0.053086
Total	1	1	0.348750	0.106172
% chance agreement[a] ($p_{e\|k_0}$)	N/A		**0.34875**	**0.106172**

[a] *This quantity is calculated as the ratio of the Total value (c.f. previous row) to the number of categories (in this case 2) minus 1*

Here is how some of Table 8.9's cells were calculated:

- $\pi_{\mathrm{E|E}} = 0.775 = (0.85 + 0.70)/2$ (from Table 8.4 and equation 8.2.1)

- $\pi_{\mathrm{E|I}} = 0.05625 = (0.0625 + 0.05)/2$ (from Table 8.5 and equation 8.2.1
- $\pi_{\mathrm{E|E}} \times \pi_{\mathrm{E|E}} = 0.174375 = 0.775(1 - 0.775)$
- $\pi_{\mathrm{I|E}} = 0.225 = (0.15 + 0.30)/2$ (from Table 8.4 and equation 8.2.1)
- $\pi_{\mathrm{I|I}} = 0.94375 = (0.9375 + 0.95)/2$ (from Table 8.5 and equation 8.2.1)
- $\pi_{\mathrm{I|E}} \times \pi_{\mathrm{I|E}} = 0.174375 = 0.225(1 - 0.225)$

▶ Conditional Kappa Coefficient

The conditional Kappa reliability coefficient given a "true" membership category k_0 is denoted by $\widehat{\kappa}_{\mathrm{C}|k_0}$ (where C stands for Cohen), and defined as follows:

$$\widehat{\kappa}_{\mathrm{C}|k_0} = \frac{p_{a|k_0} - p_{e|k_0}}{1 - p_{e|k_0}}, \ \textit{where } p_{e|k_0} = \sum_{k=1}^{q} p_{k+|k_0} p^{\star}_{+k|k_0}, \tag{8.2.5}$$

where $p^{\star}_{+k|k_0}$ is the weighted conditional probability that rater 2 classify a subject into category k given that it is a k_0-subject. This probability is calculated as follows:

$$p^{\star}_{+k|k_0} = \sum_{l=1}^{q} w_{kl} p_{+l|k_0} \tag{8.2.6}$$

The conditional percent agreement $p_{a|k_0}$ used for Kappa is the same as that used for the AC_1. However, the conditional percent chance agreement is different as seen from equation 8.2.5. Using the rating data from Table 8.1, I illustrate the calculation of the conditional percent chance agreement associated with kappa in Table 8.10. This table indicates that kappa's percent chance agreement conditionally upon true Ectopic pregnancy patients is 0.64, and is higher at 0.894 when calculated conditionally upon Intrauterine-pregnancy patients. Consequently, the two conditional kappa coefficients are given by,

$$\widehat{\kappa}_{\mathrm{C|EP}} = \frac{0.75 - 0.64}{1 - 0.64} = 0.306 \tag{8.2.7}$$

$$\widehat{\kappa}_{\mathrm{C|IP}} = \frac{0.9375 - 0.89375}{1 - 0.89375} = 0.4118. \tag{8.2.8}$$

Table 8.10: Computing Kappa's conditional chance-agreement percentages given the true category k_0

Reported Pregnancy Type (k)	Conditional agreement probabilities given k_0-subjects $\left(p_{+k\vert k_0} \times p_{k+\vert k_0}\right)$	
	$k_0 = \text{EP}$	$k_0 = \text{IP}$
$k = \text{EP}$	0.595^{a}	0.003125^{b}
$k = \text{IP}$	0.045^{c}	0.890625^{d}
Conditional chance-agreement percentage[e] $(p_{e\vert k_0})$	**0.64**	**0.89375**

[a] $0.595 = 0.85 \times 0.70$ (from Table 8.4)
[b] $0.003125 = 0.0625 \times 0.05$ (from Table 8.5)
[c] $0.045 = 0.15 \times 0.30$ (from Table 8.4)
[d] $0.890625 = 0.9375 \times 0.95$ (from Table 8.5)
[e] *This quantity is calculated by taking column totals (see equation 8.2.5)*

▶ Conditional Scott's Pi and Krippendorff's Alpha Coefficients

The Conditional Pi reliability coefficient given a "true" membership category k_0 is denoted by $\widehat{\kappa}_{\text{S}\vert k_0}$, and defined as follows:

$$\widehat{\kappa}_{\text{S}\vert k_0} = \frac{p_{a\vert k_0} - p_{e\vert k_0}}{1 - p_{e\vert k_0}}, \text{ where } p_{e\vert k_0} = \sum_{k=1}^{q} \pi_{k\vert k_0} \pi^{\star}_{k\vert k_0}, \tag{8.2.9}$$

and $\pi^{\star}_{k\vert k_0}$ is given by:

$$\pi^{\star}_{k\vert k_0} = \sum_{l=1}^{q} w_{kl} \pi_{l\vert k_0}. \tag{8.2.10}$$

Scott's conditional Pi coefficient also shares the same conditional percent agreement $p_{a\vert k_0}$ as kappa and AC_1, but is based on a different conditional percent chance agreement.

Krippendorff's conditional alpha coefficient (for 2 raters and with no missing

rating) on the other hand, is based on a percent agreement given by,

$$p^{\star}_{a|k_0} = (1-\varepsilon_n)p_{a|k_0} + \varepsilon_n, \text{ where } \varepsilon_n = 1/(2n). \tag{8.2.11}$$

The conditional percent chance agreement however, is identical to that of Scott shown in equation 8.2.9.

Using the same rating data of Table 8.1, I illustrate the calculation of the conditional percent chance agreement associated with Scott's Pi (and Krippendorff's alpha) in Table 8.11. This table indicates that Scott Pi's percent chance agreement conditionally upon true Ectopic-pregnancy patients is 0.65125, and is higher at 0.893828 when calculated conditionally upon Intrauterine pregnancy patients. Consequently, the two conditional Scott's Pi coefficients are given by,

$$\widehat{\kappa}_{\text{S}|\text{EP}} = \frac{0.75 - 0.6513}{1 - 0.6513} = 0.2832$$

$$\widehat{\kappa}_{\text{S}|\text{IP}} = \frac{0.9375 - 0.8939}{1 - 0.8939} = 0.4113.$$

Table 8.11: Computing Scott Pi's conditional chance-agreement percentages given the true category k_0

Reported Pregnancy Type (k)	$\pi_{k\|k_0}$ = conditional use probabilities[a] given k_0-subjects		$\pi^2_{k\|k_0}$	
	$k_0 = \text{EP}$	$k_0 = \text{IP}$	$k_0 = \text{EP}$	$k_0 = \text{IP}$
$k = \text{EP}$	0.775	0.05625	0.600625[b]	0.003164
$k = \text{IP}$	0.225	0.94375	0.050625[c]	0.890664
Conditional chance-agreement percentages[d] ($p_{e\|k_0}$)			**0.6513**	**0.8939**

[a] The calculation of these probabilities is described in details in Table 8.9
[b] Note that $0.600625 = 0.775^2$
[c] Note that $0.050625 = 0.225^2$
[d] *This quantity is obtained by summing the squared conditional use probabilities columnwise*

The two conditional Krippendorff's alpha coefficients are given by,

$$\widehat{\alpha}_{\text{S}|\text{EP}} = \frac{\big[1 - 1/(2 \times 100)\big] \times 0.75 + 1/(2 \times 100) - 0.6513}{1 - 0.6513} = 0.2866$$

$$\widehat{\alpha}_{\text{S}|\text{IP}} = \frac{\big[1 - 1/(2 \times 100)\big] \times 0.9375 + 1/(2 \times 100) - 0.8939}{1 - 0.8939} = 0.4139.$$

▶ **Conditional Brennan-Prediger Coefficient**

The Conditional Brennan-Prediger (BP) Coefficient given a "true" membership category k is denoted by $\widehat{\kappa}_{\text{BP}|k_0}$, and defined as follows:

$$\boxed{\widehat{\kappa}_{\text{BP}|k_0} = \frac{p_{a|k_0} - p_{e|k_0}}{1 - p_{e|k_0}}, \text{ where } p_{e|k_0} = T_w/q^2,} \tag{8.2.12}$$

and T_w is the summation of all weights. As it appears, this supposedly conditional percent chance agreement is not conditional after all. Its value does not depend on the conditioning category k_0. In the case of Table 8.1 data for example, it means that whether the patient has a true Ectopic pregnancy or a true Intrauterine pregnancy does not affect the propensity for two abstractors to agree by pure chance as it is defined in the Brennan-Prediger context. Note that $T_w = q$ for unweighted BP coefficients, since they are based on identity weights. With Table 8.1 data, $T_w = 2$ since the number of categories is 2. This leads to a percent chance agreement $p_{e|k_0} = 2/2^2 = 1/2 = 0.5$. The 2 unweighted BP coefficients conditionally upon the pregnancy types are given by,

$$\widehat{\kappa}_{\text{BP}|\text{EP}} = \frac{0.75 - 0.5}{1 - 0.5} = 0.5 \tag{8.2.13}$$

$$\widehat{\kappa}_{\text{BP}|\text{IP}} = \frac{0.9375 - 0.5}{1 - 0.5} = 0.875. \tag{8.2.14}$$

The BP coefficient also confirms the propensity of abstractors to agree to be higher for intrauterine pregnancies than for ectopic pregnancies.

8.2.3 *Unconditional Validity Coefficient for 2 Raters in ACM Studies*

The objective in this section is to quantify the extent to which two raters agree on subjects' "True" categories. The resulting metric will not simply measure how often the raters agree (i.e. reliability), instead it will measure how often the raters

agree on the subject's correct membership category. It will be a measure of validity as opposed to a measure of reliability discussed in the previous section. When all subjects used in the experiment are included into the calculations, the *Unconditional Validity Coefficient* is obtained.

Table 8.12 shows the equations needed for computing the unconditional validity version of several known agreement coefficients, including the AC_1, Pi, Kappa, and the BP (i.e. Brennan-Prediger). The (unconditional[7]) probability p_a that an agreement is reached on the correct category (also called the "percent agreement on the correct category") is common to all validity coefficients. However, the probability p_e that an agreement is reached by chance on the correct category is specific to each version of the validity coefficient, and its expression shown in column 3 of Table 8.12.

Table 8.12: Equations Associated with Various Unconditional Validity Coefficients

Coefficient	Validity Coefficient ($\widehat{\kappa}$)	Percent Chance Agreement (p_e)
% agreement (p_a)	$p_a = \sum_{k,l}\sum w_{kl}\pi_{kl}^{(k)}$	N/A
AC_2 ($\widehat{\kappa}_{2G}$)	$\widehat{\kappa}_G = (p_a - p_e)/(1-p_e)$	$p_e = \frac{T_w}{q(q-1)}\sum_{k=1}^{q}\pi_k(1-\pi_k)$
Scott's Pi ($\widehat{\kappa}_S$)	$\widehat{\kappa}_S = (p_a - p_e)/(1-p_e)$	$p_e = \sum_{k=1}^{q}\pi_k\pi_k^{\star}p_k$
Kappa ($\widehat{\kappa}_C$)	$\widehat{\kappa}_C = (p_a - p_e)/(1-p_e)$	$p_e = \sum_{k=1}^{q}p_k\Big(p_{k+}p_{+k}^{\star} + p_{+k}p_{k+}^{\star}\Big)/2$
BP ($\widehat{\kappa}_{BP}$)	$\widehat{\kappa}_{BP} = (p_a - p_e)/(1-p_e)$	$p_e = T_w/q^2$

Let $\pi_{kl}^{(k)}$ be the relative use frequency of categories k and l by raters 1 and 2 while rating k-subjects. This quantity is formally defined as $\pi_{kl}^{(k)} = (p_{kl}^{(k)} + p_{lk}^{(k)})/2$ where $p_{kl}^{(k)}$ (resp. $p_{lk}^{(k)}$) represents the relative number of times raters 1 and 2 classify a k-subject into categories k and l (resp. l and k) respectively. Moreover, $\pi_k^{\star}$ and $p_{+k}^{\star}$ are weighted sums of the probabilities π_l and p_{+l}, defined as follows:

$$\pi_k^{\star} = \sum_{l=1}^{q} w_{kl}\pi_l, \text{ and } p_{+k}^{\star} = \sum_{l=1}^{q} w_{kl}p_{+l}. \tag{8.2.15}$$

[7]The word "probability" used with no other specification will always refer to the unconditional probability, and should therefore be evaluated using all subjects that participated in the experiment.

T_w is defined as the weighted sum of the p_k's (the prevalence of k-subjects) and given by,

$$T_w = \sum_{k=1}^{q} \overline{w}_k p_{k}, \ \ where\ \overline{w}_k = (w_{k+} + w_{+k})/2. \tag{8.2.16}$$

For all unweighted analyses (i.e. based on identity weights), we will always have $T_w = 1$.

Example 8.1

This example illustrates the assessment of unconditional validity using the rating data of Table 8.1. These ratings are produced by 2 raters who rated 100 subjects with known "true" category membership. The 2 categories used in this study are Ectopic Pregnancy (or "E"), and Intrauterine Pregnancy (or "I").

For simplicity, I will confine myself to the unweighted analysis based on Identity weights where $w_{\text{EE}} = w_{\text{II}} = 1$, and $w_{\text{EI}} = w_{\text{IE}} = 0$. Consequently, the validity percent agreement is given by $p_a = \pi_{\text{EE}}^{(\text{E})} + \pi_{\text{II}}^{(\text{I})} = 0.13 + 0.73 = 0.86$. In fact, $\pi_{\text{EE}}^{\text{E}} = (p_{\text{EE}}^{\text{E}} + p_{\text{EE}}^{\text{E}})/2 = p_{\text{EE}}^{\text{E}} = 0.13$ (see Table 8.6). Likewise, $\pi_{\text{II}}^{\text{I}} = p_{\text{II}}^{(\text{I})} = 0.73$ (see Table 8.7). Validity percent agreement is expected to be smaller than reliability percent agreement, which in this case equals $0.9 = (15 + 75)/100$. Validity requires agreement to occur on the "true" category of the subject, whereas reliability only requires agreement to occur on any category.

It follows from the last 3 columns of Table 8.1 that the relative number of times category E is used is $\pi_{\text{E}} = (22/100 + 18/100)/2 = (0.22 + 0.18)/2 = 0.20$. Likewise the relative number of times category I is used is $\pi_{\text{I}} = (78/100 + 82/100)/2 = (0.78 + 0.82)/2 = 0.80$. Moreover, the prevalence of E-subjects and I-subjects is respectively given by $p_{\text{E}} = 0.20$, and $p_{\text{I}} = 0.80$.

- The percent chance agreement associated with the AC_1 coefficient is $p_e = (\pi_{\text{E}}(1 - \pi_{\text{E}}) + \pi_{\text{I}}(1 - \pi_{\text{I}}))/(2 \times (2 - 1)) = 0.16$. This leads to an AC_1 validity coefficient of $\widehat{\kappa}_{1\text{G}} = (0.86 - 0.16)/(1 - 0.16) = 0.833$. The AC_1 reliability coefficient on the other hand is 0.853.
- The percent chance agreement associated with Scott's Pi is $p_e = \pi_{\text{E}}^2 p_{\text{E}} + \pi_{\text{I}}^2 p_{\text{I}} = 0.2^2 \times 0.2 + 0.8^2 \times 0.8 = 0.52$. This leads to a Pi validity coefficient of $\widehat{\kappa}_{\text{S}} = (0.86 - 0.52)/(1 - 0.52) = 0.708$. For comparison, the Pi reliability coefficient is estimated at 0.687.
- The percent chance agreement associated with Cohen's Kappa is $p_e = p_{\text{E}+}p_{+\text{E}}p_{\text{E}} + p_{\text{I}+}p_{+\text{I}}p_{\text{I}} = 0.22 \times 0.18 \times 0.2 + 0.78 \times 0.82 \times 0.8 = 0.5196$. This leads to a kappa validity coefficient of $\widehat{\kappa}_{\text{S}} = (0.86 - 0.5196)/(1 - 0.5196) = 0.709$, which is slightly higher than the kappa reliability coefficient of 0.688.
- The percent chance agreement associated with Brennan-Prediger (BP) coefficient is $p_e = 1/2^2 = 0.25$. This leads to a BP validity coefficient of $\widehat{\kappa}_{\text{BP}} = (0.86 - 0.25)/(1 - 0.25) = 0.813$. Note that the BP reliability coefficient is slightly lower and estimated to be 0.80.

8.2.4 *Concluding Remarks on Section 8.2*

In section 8.2, I presented various ways for computing the extent to which two raters agree conditionally upon specific categories. These conditional agreement coefficients are based on the restricted pool of subjects that the gold-standard rater assigned to the conditioning categories, and were presented in section 8.2.2. In section 8.2.3, I discussed the unconditional[8] validity coefficient for two raters with respect to gold-standard ratings. I will often omit any reference to the term "unconditional" and only use the term "validity coefficient" when no confusion is possible.

We saw in Example 8.1 that validity coefficients are not always smaller than their reliability counterparts, in spite of validity requiring more stringent conditions than reliability. This nonintuitive fact stems from the notion of chance correction. While the percent agreement and percent chance agreement are smaller when used with validity coefficients than when used for reliability coefficients, the difference between these 2 percentages may turn out to be higher for validity coefficients. In other words, the propensity for agreeing by pure chance on the correct category will generally be smaller than the propensity for agreeing by pure chance on any category. It is why experiments where chance agreement is high will generally yield higher validity coefficients and lower reliability coefficients.

You may have noticed that I did not mention conditional validity coefficients. Although validity coefficients may be computed conditionally upon specific categories, this is unnecessary if you have the conditional reliability and the unconditional validity coefficients. Conditional validity is good only if both conditional reliability and unconditional validity are. If either conditional reliability or unconditional validity is low, then conditional validity will be low as well.

8.3 Validity and Conditional Agreement Coefficients for 3 Raters or More in ACM Studies

This section aims at extending the validity and conditional agreement coefficients of the past few sections to studies involving 3 raters or more. When the number of raters is 3 or more, validity coefficients quantify the extent of agreement between a roster of raters and the "gold standard" generally accepted as the reference to match. Validity coefficients are discussed in section 8.3.1, while section 8.3.2 is devoted to the conditional agreement analysis for multiple raters in ACM studies.

[8] This validity coefficient is referred to as "unconditional" because conditioning categories are not used, and all subjects are are used in the analysis.