CEMETR-2022-02
February 2022

CEME Technical Report

The Center for Educational Measurement and Evaluation

Examining Inter-Rater Reliability of Evaluators Judging Teacher Performance: Proposing an Alternative to Cohen's Kappa

Richard Lambert
T. Scott Holcomb
Bryndle Bottoms

A PUBLICATION OF THE CENTER FOR EDUCATIONAL MEASUREMENT AND EVALUATION

Examining Inter-Rater Reliability of Evaluators Judging Teacher Performance: Proposing an Alternative to Cohen's Kappa

Richard G. Lambert
Center for Educational Measurement and Evaluation
UNC Charlotte

T. Scott Holcomb
Duke University

Bryndle Bottoms
University of South Carolina

An earlier version of this paper was presented to the virtual Annual Meeting of the National Council on Measurement in Education, June, 2021.

Abstract

The validity of the Kappa coefficient of chance-corrected agreement has been questioned when the prevalence of specific rating scale categories is low and agreement between raters is high. The researchers proposed the Lambda Coefficient of Rater-Mediated Agreement as an alternative to Kappa to address these concerns. Lambda corrects for chance agreement based on specific assumptions about raters and the rater-mediated assessment process including rater-specific tendencies for strict or lenient ratings. Actual ratings of teacher profiles from an inter-rater reliability exercise confirmed the shortcomings of Kappa. The rater data also demonstrated the robustness of Lambda-1, Lambda-2, Gwet's AC1, and Gwet's AC2 to the data conditions that are problematic for Kappa. All four alternative chance-corrected agreement coefficients showed less variability across the 65 raters than Kappa. However, AC-2 was undetermined for 57 of the 65 raters. Simulation data demonstrated the robustness of the Lambda Coefficient of Rater-Mediated Agreement to the data conditions that are problematic for Kappa.

Examining Inter-Rater Reliability of Evaluators Judging Teacher Performance: An Alternative to Cohen's Kappa

Cohen's Kappa (Cohen, 1960) and Weighted Kappa (Cohen, 1968) are widely used measures of chance-corrected agreement between raters. Researchers have raised questions about whether Kappa is actually correcting for chance agreement and whether it is useful for identifying and separating various sources of disagreement. The validity of Kappa coefficients has been questioned when prevalence of specific categories on a rating scale is low and agreement is high. In addition, researchers have raised questions about the generalizability of Kappa coefficients across populations and study conditions (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti; 1990; Thompson & Walter, 1988). Gwet (2008) introduced AC1 and AC2 as alternatives to Kappa and demonstrated they are robust indexes not susceptible to the identified shortcomings of Kappa.

This study proposed an alternative to Kappa, Weighted Kappa, AC1, and AC2 that is rooted in theory regarding rated-mediated assessment (Engelhard & Wind, 2018). We designed the Lambda Coefficient of Rater-Mediated Agreement for use with the ordinal scales often used to evaluate teacher performance. It examines inter-rater agreement corrected for the probability that raters may agree with expert raters by chance due to the response process they employ when they are uncertain about how to place a teacher on a rubric.

Rater-Mediated Assessment Theory

Most research on rater cognition focuses on the mental processes used by raters of student or examinee performance. The rating process followed by raters judging constructed responses is intricate; however, the judgment of teacher performance can pose even greater complexities. Similar to other examples of rater-mediated assessments, an observers' level of

expertise, and the overall scoring task demands, can influence ratings of teacher performance (Bell et al., 2018; Suto, 2012). Understanding rater cognition is crucial to making a validity argument to support the use of any rater-mediated assessment measure. According to Standard 1.12, which addresses "evidence regarding cognitive processes", in the *Standards for Educational and Psychological Testing* (2014):

> "If the rationale for score interpretation for a given use depends on premises about the psychological processes or cognitive operations of test takes, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by observers or scorers are part of the argument for validity, similar information should be provided." (p. 26)

It is common to provide raters of any construct a rubric and a set of exemplar and counter example responses to serve as anchors or benchmarks in scoring decisions. This standard makes it important to go beyond rater training to analyze and interpret raters' cognitive processes, perceptions of exemplar responses, and to make sure those processes are congruent with the measured construct (Bejar, 2012).

Engelhard et al. (2018) asserted the validity, reliability, and fairness of rater-mediated assessments relies on both the quality of the rater's cognitive process and psychometric properties of the measure. In rater-mediated assessment, understanding raters' scoring processes is an important component of understanding what an assessment is actually measuring (Crisp, 2012). Interpreting and anticipating a rater's cognitive process "can provide practical information to assist those who are designing performance tasks and rubrics, selecting raters, training raters, and developing quality control procedures to monitor rater performance, particularly in 'real time' as a scoring session is proceeding" (Myford, 2012, p. 49).

Brunswik (1952) proposed the Lens Model as a human judgment and decision-making conceptual framework. This model was adapted by Engelhard (2013) as a conceptual framework for rater judgment and decision-making. The goal of this adapted version of the lens model "is to have a close correspondence between the latent variable and the observed ratings" (Engelhard & Wind, 2018, p. 81). The connection takes place between the items, raters, and the rating scale. Teacher evaluation is multidimensional and requires raters to have a high level of expertise and skill in interpreting information through this model.

Interrater Reliability Coefficients

Bennett, Alpert, and Goldstein's *S* (1954; referred to as *S* throughout remainder of text) coefficient is not a chance-corrected agreement coefficient; however, it was an initial attempt at providing information more meaningful than exact percentage agreement figures in IRR studies. They based *S* on the proportion of agreement according to the number of categories on a rating scale. It produces a constant value for all ratings with the same number of categories and level of exact agreement. The original chance-corrected agreement coefficient, Scott's π, was introduced by Scott (1955) as an improvement over the use of simple observed agreement percentages and was designed for nominal data in communication studies. Cohen's Kappa (1960) was developed as an improved chance-corrected agreement coefficient. Kappa differed slightly in the calculation of expected agreement as compared to Scott's π based on how the marginal distribution of ratings were involved in this calculation (Banerjee et al., 1999).

The use of IRR coefficients has expanded across various rating contexts involving ratings made using varying scales of measurement. The expanding number of applications have exposed researchers across many disciplines to the flaws of the Kappa coefficient. Consequently, researchers across many fields have introduced measures of interrater agreement that provide

information beyond simple percentage agreement (Zwick, 1988). Most of these coefficients differ in how each defines agreement due to chance.

The main issues with the Kappa coefficient involve situations where actual agreement is high, and the number of rating categories used is low. These situations result in unrealistically low Kappa values. Gwet's first-order agreement coefficient (AC1) was proposed as being resistant to Kappa's known paradoxes by setting a maximum limit of 0.5 on the proportion of chance-agreement. Kappa allows this value to range from 0.0 to 1.0 depending on the marginal distribution of ratings (Gwet, 2001). AC1 produces coefficients close to the percentage of agreement while still accounting for the random chance of agreement and was designed to use with any number of raters using categorical rating systems. In clinical psychology applications, AC1 is recommended in place of, or at least along with, Kappa (Wongpakaran et al., 2013) because it offers a more stable measure than Kappa and is simpler to calculate as compared to other IRR coefficients. Another distinguishing trait of AC1 is that it does not require the assumption of independence between raters (Gwet, 2008). Gwet's second-order agreement coefficient (AC2) is a weighted version of AC1 and was intended to be used for ordinal, interval, and ratio data (Gwet, 2014; 2016).

Rater Cognition

Rater cognition is the term used to describe the process raters go through when assigning ratings to performances or products and related mental activities. According to Eckes (2011), "rater cognition refers to the mental structures and processes involved in assigning ratings to examine performances of products" (p. 189). Developing and acknowledging a deeper understanding of the process raters go through when making ratings adds value to the

interpretation of scores. The consistency of raters' response processes is one source of validity evidence explained in the *Standards* (AERA et al., 2014).

Whenever human raters are used to evaluate or score assessment performance there is a potential for subjectivity and inaccuracies. According to Bejar (2012), in instances where raters are used to make rating judgments, rater cognitive processes should be considered in the assessment design phase as well as the assessment scoring phase. During the design phase of an assessment, raters should be recruited and trained. This training process allows "raters to form a mental scoring rubric", which is built into a rater's cognitive process during appropriate training (Bejar, 2012, p. 5). While training will not alleviate all rater effects, it can assist raters in forming and developing proper mental processes that align with the rubric or rating scale. Understanding and researching rater cognition are beneficial practices that can help minimize threats to the validity of ratings related to rater judgment. Therefore, consistency of rater cognition is required for valid, reliable, and fair assessment practices to occur in rater-mediated assessment.

Rater Assumptions

The following set of assumptions about raters, and the complex response process they use to arrive at ratings, serve as a theoretical foundation for the Lambda Coefficient of Rater-Mediated Agreement. We posit the following principles regarding the internal cognitive process raters employ when they are confident about which rating to assign:

- Raters are trained evaluators and function as expert professionals.
- Rather than acting as scoring machines, raters bring their own experiences and expertise to the rating process.
- Raters use a complex, three-stage internal response process to make ratings.
- First, raters acquire an overall impression, based on global evidence, to arrive at a starting

point on a rubric or rating scale.

- Second, raters synthesize information from previous ratings, analyze observational data, and interpret evidence and artifacts.
- Third, raters combine their overall impressions with their analysis of evidence to settle on a final placement on a rubric or rating scale.
- A rater's individual tendencies toward strictness and leniency influence this complex internal response process.

This process functions at several levels. When raters consider how to make a rating on an item that addresses a specific area of practice, such as ratings focused on particular teaching competencies, they start with their overall impression, analyze a variety of item-specific pieces of evidence, synthesize the ratings they have made across items that address similar content, and then settle on a final rating. Similarly, when raters make global ratings, such as ratings of overall teaching effectiveness or quality, they may start with their overall impression, analyze a variety of pieces of evidence, synthesize the ratings they have made across items that address various content areas, and then settle on a final rating. However, raters do resort to guessing, or at least a random process similar to guessing, when they are uncertain about a particular rating. We posit the following principles regarding the internal cognitive process raters employ when they are uncertain about which rating to assign:

- Professional raters can be, on occasion, uncertain about their selection of ratings.
- A professional rater may, on occasion, lack the experience, expertise, or evidence to have confidence in a particular rating.
- When uncertain, raters make ratings by a random process that mimics the three-stage internal cognitive response process they use when confident in their ratings.

These ratings may, by chance, agree with the ratings of another rater or those from an expert panel.

- When uncertain, raters select a random starting point for deliberations.
- When uncertain, raters may synthesize previous ratings and analyze evidence, but this process does not resolve their uncertainty. When uncertain, raters may have little confidence in their previous ratings and may lack sufficient evidence to support a particular rating.
- When uncertain, raters combine their initial random starting point with their inconclusive analysis of evidence to settle on a final rating.
- A rater's individual tendencies toward strictness and leniency influence this random response process.

We developed the Lambda Coefficient of Rater-Mediated Agreement based on these assumptions concerning the response process raters use when applying ordinal ratings scales to tasks such as teacher evaluation.

The Lambda Coefficient of Rater-Mediated Agreement

Kappa (Cohen, 1960) is equivalent to the proportion of the ratings that are in agreement with another rater, after removing the proportion of the agreement ratings that may have occurred by chance. The formulae take the following forms:

$$\kappa = \frac{p_a - p_e}{1 - p_e} \tag{1}$$

$$\sigma_x = \sqrt{\frac{p_a(1-p_a)}{n(1-p_e)^2}} \tag{2}$$

Where:

p_a = Proportion of exact agreement.

p_e = Expected proportion chance agreement. For Kappa, this quantity is equal to the sum of the products of the marginal proportions associated with each cell.

n = number of ratings.

There have been various alternatives to Kappa proposed in the years since (Holley & Guilford, 1964; Maxwell, 1970; Krippendorff, 1970; Jason & Vegelius, 1979; Brennan & Prediger, 1981; Perreault & Leigh, 1989; Bryt, Bishop, & Carlin, 1993; Gwet, 2008). In addition, Bennett, Albert, and Goldstein (1954) proposed a method equivalent to the method reintroduced by Bryt, Bishop, & Carlin (1993) prior to the introduction of Kappa. These alternatives to Kappa were developed based on the assumptions of generalizability theory, applications for nominal scales, or both. Their focus was primarily on agreement among raters, judges, or observers with respect to the presence or absence of specific characteristics, symptoms, or diagnoses. Such applications do not involve rater strictness or leniency as is often present in rater use of ordinal rating scales.

We are proposing the Lambda Coefficient of Rater-Mediated Agreement for a different set of applications. The researchers are proposing Lambda based on the theoretical propositions of rater-mediated assessment (Engelhard & Wind, 2018) and our applied work with teacher performance evaluations. Teacher performance evaluations are typically conducted using ordinal scales. The agreement of interest is between individual raters and expert raters. Furthermore, teacher performance evaluations are high stakes endeavors, and strict or lenient ratings can have significant consequences for teachers. Evaluators make placements on such scales based on observational data, classroom artifacts, student work samples, and general overall impressions. Raters can and will have their own personal tendencies toward strictness or leniency, or even biases. In addition, teacher evaluators can be uncertain about a particular rating and can use a

random process to settle on their final placements on ordinal rating scales. Our goal is to correct for chance agreement that may occur due to this complex cognitive process.

We set out to develop a coefficient for ordinal scales that meets several important criteria. First, Lambda had to agree with Kappa when all ratings fall on the main diagonal of the ratings matrix. When all rater placements are in agreement with the expert ratings, Lambda and Kappa both = 1.0. Furthermore, when rater agreement, strictness, and leniency are all equal this is equivalent to a rater cognitive process that involves simple guessing. Therefore, Kappa and Lambda should agree in these circumstances and they do. Next, we sought a coefficient that would equal zero when all ratings in the ratings matrix have equal frequency, and both Kappa and Lambda equal zero under these circumstances. Finally, we sought to develop a coefficient that had a reasonable upper bound on the magnitude of the correction for chance, similar to Gwet's approach (2008). Lambda-1, described below, met all of these criteria. For example, Lambda-1 has an upper bound on chance agreement of .5 for a 4x4 ratings matrix as defined by these quantities.

$$p_e \leq 2L / q \tag{3}$$

$$p_e \leq 2S / q \tag{4}$$

Where:

p_e = Expected proportion chance agreement.

L = proportion of ratings that are lenient, or above the "correct" or "expert" rating.

S = proportion of ratings that are strict, or below the "correct" or "expert" rating.

q = number of steps on the ordinal rating scale.

The general form for λ, applicable to both λ_1 and λ_2, and to rating scales with any number of steps, can be expressed as:

$$\lambda = \frac{p_a - p_e}{1 - p_e} \quad (5)$$

$$p_e = \sum p_s \, p_c p_f \quad (6)$$

$$\sigma_\lambda = \sqrt{\frac{p_a(1 - p_a)}{n\,(1 - p_e)^2}} \quad (7)$$

Where:

p_a = Proportion of exact agreement.

p_e = Proportion expected chance agreement.

Σ = Sum across all cells from r=1, c=1 to r=q, c=q.

r = row.

c = column.

n = number of ratings.

q = The number of steps on the ordinal rating scale.

p_s = Probability of picking the given cell as a starting point (s) for deliberation.

p_c = Proportion of ratings for which the given column (c) is used as a correct answer.

p_f = Expected probability of exact agreement when the given cell was used as a starting point, and the rater makes a final (f) rating informed by their tendency for agreement, strictness, and leniency.

The only difference between λ_1 and λ_2 is the formula for p_s. For λ_1, $p_s = 1/q$. This value assumes when a rater is uncertain about which rating to give, and arrives at a random starting point for their deliberations, they are equally likely to select any of the points on the rating scale as a starting point. For λ_2, p_s is set to the population proportion of the total ratings for which raters used the rating scale level associated with the cell in question. This value is the marginal proportion for the given row in the agreement matrix formed by using the population data. This

value also assumes when a rater is uncertain about which rating to give, and uses guessing as a means to arrive at a starting point for their deliberations, their internal guessing process weights the points on the rating scale according to how frequently they are used. So for example, if the population of raters very rarely uses a particular point on the rating scale, λ_2 assumes an individual rater would be much less likely to select that point as a starting point for deliberations. To illustrate how Lambda-1 and Lambda-2 work in practice, see Figure 1 for a four-point ordinal rating scale. Figure 1 illustrates how the chance agreement values would be calculated for each cell in the agreement matrix. These values would then be summed across all cells to compute chance agreement. Just for illustration purposes, we have included category labels that might apply to a teacher performance evaluation rubric. Note that for the Lambda-2 demonstration in Figure 1, we assume the population proportions (P_1, P_2, P_3, and P_4) for each step on the rating scale are known. Figures 2 and 3 illustrate this process for two- and three-point ordinal rating scales.

The Current Study

The purpose of this study was twofold. First, the researchers sought to test the performance of Lambda relative to Kappa, AC-1, and AC-2 using field data. Second, the researchers sought to evaluate Lambda relative to Kappa, AC-1, and AC-2 with simulated data that represents the high agreement / low frequency of specific categories data conditions under which Kappa is known to yield paradoxical results. Specifically, this study examined the following research questions:

1. How does the Lambda Coefficient of Rater-Mediated Agreement perform relative to Kappa, AC-1, and AC-2 given real world teacher performance evaluation data?

2. Does the Lambda Coefficient of Rater-Mediated Agreement yield chance-corrected coefficients of agreement that are robust to data conditions that have been shown to be problematic for Kappa (high agreement and some rating scale categories with low prevalence)?

Methods

Evaluators charged with conducting state-mandated performance evaluations of all licensed pre-kindergarten teachers working in non-public school settings within one state participated in an inter-rater reliability certification exercise. Evaluators (n=65) made placements on five progressions across each of 10 online teacher profiles for a total of 3,250 ratings. The evaluators rated the teacher profiles using the North Carolina Teacher Evaluation Process rubric, which is the same teacher performance evaluation measure used to conduct evaluations across the entire state. The measure includes five progressions, called standards, each of which contains specific behavioral anchors and is supported by a series of rubrics called "elements." The evaluators used the same four-point rating scale for all standards and elements. The ordinal scale points were labeled (1) Developing, (2) Proficient, (3) Accomplished, and (4) Distinguished. Agreement was evaluated against the "correct answer" ratings from a panel of five experts who had previously achieved consensus. The following statistics were calculated for each evaluator using the criteria for exact agreement with the expert panel: a.) agreement, leniency, and strictness percentages, b.) Cohen's Kappa, c.) Lambda-1 Coefficient of Rater-Mediated Agreement, d.) Lambda-2 Coefficient of Rater-Mediated Agreement, e.) Gwet's AC-1, and f.) Gwet's AC-2.

An alternative scoring strategy that allowed for agreement between some adjacent ratings was also developed. Adjacent agreement was defined as an expert panel rating of "Proficient"

and an evaluator rating of either “Proficient” or “Accomplished”, or an expert panel rating of “Accomplished” and an evaluator rating of either “Proficient” or “Accomplished.” Exact agreement was still required for expert panel ratings of either “Developing” or “Distinguished.” The rationale was there is no difference in how teachers who are rated as “Proficient” or “Accomplished” are treated within either a mentoring or a performance evaluation context in the particular state under investigation. Teachers must obtain ratings of at least “Proficient” across all standards by the end of their third year of teaching. Therefore, teachers rated as “Developing” receive additional mentoring. Teachers rated “Distinguished” are rare and may be asked to serve as model teachers, mentors, or evaluators. The same statistics were calculated for each evaluator using the criteria for adjacent agreement with the expert panel: a.) agreement, leniency, and strictness percentages, b.) Cohen’s Kappa, c.) Lambda-1 Coefficient of Rater-Mediated Agreement, d.) Lambda-2 Coefficient of Rater-Mediated Agreement, e.) Gwet’s AC-1, and f.) Gwet’s AC-2. It should be noted that the Kappa coefficient for the adjacent agreement condition is equivalent to a special case of Weighted Kappa (Cohen, 1968) with weights assigned according to this particular adjacent scoring scheme.

To address research question 2, we used a simulation design that extended the approach of Xie (2013) to include Lambda-1 and Lambda-2. For the purpose of this simulation study, we defined the Bias Index as Strictness minus Leniency (expressed as proportions). We defined the Prevalence Index as the proportion of rater selections using the lowest point on the ratings scale minus the proportion of rater selections using the highest point on the rating scale. We varied the Prevalence Index across all possible values for each condition and calculated the values of Kappa, Lambda-1, and Lambda-2 for each condition. This study was not a Monte Carlo simulation, sampling error as not involved, and similar to Xie (2013) we simply included the

values each coefficient would take on under the specific conditions. For Lambda-2, we used the following population values from our knowledge of the actual population distribution of rating made in the field: Developing – 5%, Proficient – 65%, Accomplished – 25%, Distinguished – 5%. Four simulated conditions used a four point rating scale similar to the real world data conditions reported for research question one. These four conditions included high agreement and low category frequency conditions known to be problematic for Kappa. The four conditions were: 1.) Agreement = 95%, Bias Index = .05, Prevalence ranged from .95 to -.95, 2.) Agreement = 90%, Bias Index = .10, Prevalence ranged from .90 to -.90, 3.) Agreement = 85%, Bias Index = .15, Prevalence ranged from .85 to -.85, and 4.) Agreement = 80%, Bias Index = .20, Prevalence ranged from .80 to -.80. We calculated Kappa, Lambda-1, and Lambda-2 for each of the four conditions across the applicable range of the Prevalence Index.

Results

First, we examined the distribution of the agreement, strictness, and leniency percentages for all 65 raters across both the exact and adjacent agreement conditions. Table 1 contains the mean, standard deviation, and five number summary for each of these percentages. The mean percent exact agreement across the 65 evaluators was 69.3% (*SD* = 9.3) and values ranged from 42.0% to 88.0%. The mean percent lenient for exact agreement was 9.8% (*SD* = 8.4) and values ranged from 0.0% to 44.0%. The mean percent strict for exact agreement was 21.0% (*SD* = 11.8) and values ranged from 2.0% to 58.0%. Therefore, the raters as a group displayed moderate levels of agreement. However, a substantial minority of raters (*n* = 7) agreed with the expert panel for less than 60% of their ratings. The raters as a group also displayed more strictness in their ratings than they did leniency.

As expected, agreement percentages increased, and strictness and leniency percentages decreased, for the adjacent agreement condition. Only four of the 65 raters displayed adjacent agreement percentages less than 80%. The mean percent adjacent agreement across the 65 evaluators was 88.5% (*SD* = 6.6) and values ranged from 64.0% to 100.0%. The mean percent lenient for adjacent agreement was 3.4% (*SD* = 4.2) and values ranged from 0.0% to 20.0%. The mean percent strict for adjacent agreement was 8.2% (*SD* = 7.1) and values ranged from 0.0% to 36.0%. For the adjacent agreement condition, strictness was again greater than leniency; however, both values were much lower than they were in the exact agreement condition.

The distributions of each of five coefficients of chance-corrected agreement were compared to address research question one. Table 2 contains the mean, standard deviation, and five number summary for each of the coefficients were calculated. The mean Kappa for exact agreement was .489 (*SD* = 0.152) and values ranged from .121 to .790. The mean Lambda-1 Coefficient of Rater-Mediated Agreement for exact agreement was .587 (*SD* = .127) and values ranged from .199 to .840. The mean Lambda-2 Coefficient of Rater-Mediated Agreement for exact agreement was .511 (*SD* = .136) and values ranged from .121 to .798. The mean Gwet's AC-1 for exact agreement was .615 (*SD* = .117) and values ranged from .273 to .852. The mean Gwet's AC-2 for exact agreement was .507 (*SD* = .113) and values ranged from .356 to .706. However, Gwet's AC-2 was undetermined for 57 of the 65 raters. Figure 4 displays these distributions as boxplots. AC-2 was not included due to small sample size (*n* = 8). The boxplots show that Lambda-1, Lambda-2, and Gwet's AC-1 yielded less severe and less variable corrections for chance agreement than Kappa. For the exact agreement condition, Kappa did not show sensitivity to one outlier rater detected by the other coefficients.

A very consistent rank order of correction among the alternatives to Kappa emerged for the exact agreement condition (see Figure 6). Lambda-2 was closest to Kappa for all 65 raters. Lambda-1 yielded coefficients that were consistently higher than Kappa and Lambda-2 and lower than AC-1 and this pattern held for 63 of the 65 raters. AC-1 emerged as yielding the consistently highest coefficients and this pattern held for 63 of the 65 raters. AC-2 was not included in these comparisons due to the small sample size. As seen in Figure 6, the most dramatic changes to the relatively consistent rank order of the coefficients appear between Lambda-1 and AC-1 as indicated by several lines that break from the overall pattern.

The mean Kappa for adjacent agreement was .661 (*SD* = .176) and values ranged from as low as -.018 to 1.000 (see Figure 5). The mean Lambda-1 Coefficient of Rater-Mediated Agreement for adjacent agreement was .826 (*SD* = .100) and values ranged from .442 to 1.000. The mean Lambda-2 Coefficient of Rater-Mediated Agreement for adjacent agreement was .657 (*SD* = .159) and values ranged from .192 to 1.000. The mean Gwet's AC-1 for adjacent agreement was .860 (*SD* = .083) and ranged from .532 to 1.000. The mean Gwet's AC-2 for adjacent agreement was .800 (*SD* = .063) and values ranged from .687 to .911. However, again Gwet's AC-2 was undetermined for 57 of the 65 raters. AC-2 was not included due to small sample size (*n* = 8). Lambda-1, Lambda-2, and Gwet's AC-1 yielded less severe and less variable corrections for chance agreement than Kappa. Lambda-1 and AC-1 yielded similar distributions, and displayed less variability than Kappa or Lambda-2. For the adjacent agreement condition, Kappa was sensitive to the three outlier raters and the other three coefficients detected the same single outlier rater.

A very consistent rank order of correction among the alternatives to Kappa emerged for the adjacent agreement condition as well. Lambda-2 was closest to Kappa for 63 of the 65 raters.

Lambda-1 coefficients were higher than Kappa and Lambda-2 and lower than AC-1 for 63 of the 65 raters. AC-1 yielded the highest coefficients for 64 of the 65 raters. AC-2 was not included in these comparisons due to the small sample size (n = 8). There was only one exception to this pattern. One rater had 100% adjacent agreement with the expert panel and all coefficients equaled 1.00.

The results of the simulation study addressed research question two across the four proposed conditions. For condition 1 (95% agreement), the mean Kappa, across the full range of the Prevalence Index, was .820 (*SD* = .147) and values range from -.053 to .904. The mean Lambda-1 was .933 (*SD* = .001) and values range from .932 to .934 (see Figure 8). The mean Lambda-2 was .947 (*SD* = .001) and values range from .946 to .948. Therefore, Lambda-1 and Lambda-2 yielded very consistent values and Kappa yielded very inconsistent and over-corrected values. The remaining three conditions yielded similar patterns. For condition 2 (90% agreement), the mean Kappa was .739 (*SD* = .111) and values range from .298 to .825. The mean Lambda-1 was .867 (*SD* = .002) and values range from .863 to .871. The mean Lambda-2 was .892 (*SD* = .002) and values range from .888 to .895. For condition 3 (85% agreement), the mean Kappa was .674 (*SD* = .095) and values range from .355 to .756. The mean Lambda-1 was .801 (*SD* = .005) and values range from .793 to .810. The mean Lambda-2 was .836 (*SD* = .005) and values range from .827 to .843. For condition 4 (80% agreement), the mean Kappa was .621 (*SD* = .082) and values range from .370 to .696. The mean Lambda-1 was .737 (*SD* = .008) and values range from .722 to .750. The mean Lambda-2 was .779 (*SD* = .008) and values range from .764 to .792. For all four conditions, Lambda-1 and Lambda-2 tended to remain very consistent across the values of the Prevalence Index. Kappa tended to get much smaller and over-correct as values of the Prevalence Index moved away from zero.

Discussion

The results of this study confirmed and extended previous research (Gwet, 2008) by illustrating the shortcomings of Kappa as a measure of chance-corrected agreement and the robustness of AC-1 to the data conditions associated with these shortcomings. These results also illustrated how the proposed Lambda-1 Coefficient of Rater-Mediated Agreement is resistant to the data conditions that are problematic for Kappa, and offers a slightly more conservative, less variable measure of chance-corrected agreement than AC-1 while also demonstrating greater sensitivity to outlier raters.

The data from this study contained various examples, both real and simulated, of the high agreement / low frequency of specific rating scale categories problem. For example, the real world data included very infrequent use of the "Distinguished" category by the evaluators or experts. In practice, raters use "Distinguished" very rarely and reserve its use for truly exceptional teachers. The paradoxical performance of Kappa found in previous studies under these data conditions was confirmed (Cicchetti & Feinstein, 1990; Gwet, 2008). Consistent with previous research, Cohen's Kappa was overly sensitive and over-corrected when agreement was high and there was low frequency of specific categories on the rating scale. Cohen's Kappa yielded results consistent with the paradox problem in which percent agreement is high and Cohen's Kappa is low or even .000. However, both Lambda-1 Coefficient of Rater-Mediated Agreement and Gwet's AC-1 yield results that were robust to both of these data conditions.

Separate examinations of the fictitious teacher profiles revealed several stark examples of this pattern. For example, neither the expert panel nor the evaluators selected ratings of "Accomplished" or "Distinguished" for Profile 4. Across all 65 raters, agreement was high (93.33%) for both the exact and adjacent methods. However, Cohen's Kappa was .000 for the

exact method while Lambda-1 equaled .913 and AC-1 equaled .932. Cohen's Kappa was also .000 for the adjacent method while Lambda-1 equaled .903 and AC-1 equaled .933. Similarly, for Profile 5 no ratings of "Developing" or "Distinguished" were selected by either the expert panel or the evaluators, and agreement was moderate (63.11%) for exact method and high for the adjacent method (96.44%). However, Cohen's Kappa was .000 for the exact method while Lambda-1 equaled .508 and AC-1 equaled .589. Cohen's Kappa was also .000 for the adjacent method while Lambda-1 equaled .947 and AC-1 equaled .964.

The results of the simulation portion of this study also demonstrated that Lambda is less susceptible to the shortcomings of Kappa under specific data conditions. However, this study contained only four specific simulated data conditions, all of which included high agreement, low category frequency, and low Bias Index values. A more extensive simulation study extended this work across many more data conditions with similar results (Holcomb et al., 2022b). Future research is still needed to examine the performance of Lambda relative to both Kappa and its alternatives across additional simulated and real world data conditions.

Collectively, the results of this study demonstrate that both the magnitude of individual chance-corrected agreement coefficients, and the variability of those coefficients across a team of raters, are contingent upon the researcher's choice of index of agreement. Tables 3 and 4 demonstrate that the choice of coefficient can determine whether an individual rater's level of chance-corrected agreement fall in the acceptable range. Specifically, Kappa classified fewer raters in the "good" to "excellent" range than all other coefficients tested. A separate study, using data from some of the real world evaluators involved in the current study, confirmed these findings, and demonstrated that rater behavior can vary quite extensively even when rating the same teachers (Holcomb et al., 2022a). Holcomb et al. (2022a) also demonstrated, using multiple

criteria for acceptable levels of agreement (Landis and Koch, 1977; Fleiss, 1981; Altman, 1991), that raters can be classified in the acceptable range using one coefficient and in the unacceptable range using another. Therefore, choice of agreement statistic can matter to both researchers and practitioners, and can impact decisions to retain, retrain, or support raters.

It is important to point out that agreement and or chance-corrected agreement statistics are not sufficient to identify problematic raters and the specific supports they may need. Lambda is not meant to provide the rich information that a more complex measurement model can provide about individual raters and their tendencies. For example, the Many-Facets Rasch Model (Linacre, 1989) can provide a detailed calibration of individual rater strictness and leniency and potential biases. Furthermore, Lambda is a single coefficient that is agnostic to where in the rating space strictness or leniency occurs. Lambda is not influenced by the specific steps on a rating scale that are associated with a rater's tendencies for strictness or leniency. It is, however, useful as a red flag, as one indicator among many, of the need to retain, certify, support, or retrain individual raters.

In conclusion, this study confirmed the advantages of AC-1 over Kappa demonstrated in previous research (Gwet, 2008). In addition, this study introduced the Lambda Coefficient of Rater-Mediated Agreement, a coefficient rooted in the theoretical underpinnings of rater-mediated assessment (Engelhard & Wind, 2018). It operationalizes a series of proposed principles regarding the complex process by which raters make placements on ordinal progressions. Future research is needed to test these theoretical propositions further and to investigate the cognitive processes raters use when they feel confident in their ratings and when they are uncertain. The current study, with both field data and simulated data, highlighted the robustness of the Lambda Coefficient of Rater-Mediated Agreement to the data conditions that

are problematic for Kappa, while offering a less variable metric that is more sensitive to outlier raters.

References

Altman, D. (1991). *Practical statistics for medical research*. CRC Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond Kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics, 27*(1), 3-23.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2-9.

Bell, C. A., Jones, N. D., Qi, Y., & Lewis, J. M. (2018). Strategies for assessing classroom teaching: Examining administrator thinking as validity evidence. *Educational Assessment, 23*(4), 229-249. https:doi.org/10.1080/10627197.2018.1513788

Bennett E. M., Albert R., & Goldstein A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly*, *18* (3), 303-308. https://doi.org/10.1086/266520

Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41* (3), 687-699.

Brunswik, E. (1952). *The conceptual framework of psychology.* University of Chicago Press.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence, and kappa. *Journal of Clinical Epidemiology, 46* (5), 423-429.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology, 43(6), 551–558*.

Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, *20(1), 37–46*. https://doi.org/10.117/001316446002000104

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

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement: Issues and Practice, 31*(3), 10-20.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (1st ed.). Peter Lang.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences.* Routledge.

Engelhard, G., Wang, J., & Wing, S. A. (2018). A tale of two models: Psychometric and cognitive perspectives on rater-mediated assessments using accuracy ratings. *Psychological Test and Assessment Modeling, 60*(1), 33-52.

Engelhard, G. & Wind, S. (2018). Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments. Routledge.

Feinstein, A.R. & Cicchetti, D.V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43(6), 543-549*.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31*, 651-659.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. Wiley.

Gwet. K. (2001). *Handbook of inter-rater reliability*. STATAXIS Publishing Company.

Gwet, K. (2008). Computing inter-rater reliability and its variance in the presence of high Agreement. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1348/000711006X126600

Gwet. K. (2014). *Handbook of inter-rater reliability* (4th ed.). Advanced Analytics Press.

Gwet, K. L. (2016). Testing the difference of correlated agreement coefficients for statistical significance. *Educational and Psychological Measurement, 76*(4), 609-637. DOI: https://doi.org/10.1177/0013164415596420

Holcomb, S., Lambert, R., and Bottoms, B. (2022a). Reliability evidence for the NC teacher evaluation process using a variety of indicators of inter-rater agreement. *Journal of Educational Supervision, 5(1),* 27-43. DOI: https://doi.org/10.31045/jes.5.1.2

Holcomb, S., Lambert, R., & Bottoms, B. (April, 2022b). *Lambda Coefficient of Rater-Mediated Agreement: Demonstration of a Chance-Corrected Agreement Coefficient.* Paper presented to the Annual Meeting of the American Educational Research Association, San Diego, CA.

Holley, W., & Guilford, J. P. (1964). A note on the G-index of agreement. *Educational and Psychological Measurement, 24*, 749-754.

Janson, S., & Vegelius, J. (1979). On generalizations of the G index and the phi coefficient to nominal scales. *Multivariate Behavioral Research, 14*, 255-269.

Krippendorff, K. (1970). Bivariate agreement coefficients for reliability data. In E. R. Borgatta & G. W. Bohrnstedt (Eds.), *Sociological Methodology 1970* (pp. 139-150). Jossey Bass.

Lambert, R., Holcomb, S., & Bottoms, B. (June, 2021). *Examining Inter-Rater Reliability of Evaluators Judging Teacher Performance: Proposing an Alternative to Cohen's Kappa.* Paper presented to the virtual Annual Meeting of the National Council on Measurement

in Education.

Lambert, R., Holcomb, S., & Bottoms, B. (February, 2022). *The Examining the interrater reliability of evaluators judging teacher performance: Proposing an alternative to Cohen's Kappa.* Retrieved from the Center for Educational Measurement and Evaluation website: http://ceme.uncc.edu/ceme-technical-reports

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics,* 33, 159-174.

Linacre, J.M. (1989). *Many-Facets Rasch Measurement*. MESA Press.

Maxwell, A.E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry*, *130*, 79–83.

Myford, C. M. (2012). Rater cognition research: Some possible directions for the future. *Educational Measurement: Issues and Practice, 31*(3), 48-49.

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

Porter, J. M. & Jelinek, D. (2011). Evaluating inter-rater reliability of a national assessment model for teacher performance, *International Journal of Educational Policies*, *5*(2), *74-87*.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*, 321-325.

Suto, I. (2012). A critical review of some qualitative research methods used to explore rater cognition. *Educational Measurement: Issues and Practice, 31*(3), 21-30. https://doi.org/10.1111/j.1745-399.2012.00240.x

Thompson, W. D. & Walter, S. D. (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology, 41(10), 949-958*.

Uebersax, J. (2006). Tests of marginal homogeneity. Retrieved from https://www.john-uebersax.com/stat/margin.htm

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology Journal*, *13*, 1-7. https://doi.org/10.1186/1471-2288-13-61

Xie, Q. (2013). Agree or disagree? A demonstration of an alternative statistic to Cohen's Kappa for measuring the extent and reliability of agreement between observers. Federal Committee on Statistical Methodology, U.S. Office of Management and Budget. Retrieved from https://nces.ed.gov/FCSM/pdf/J4_Xie_2013FCSM.pdf.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103*(3), 374-378. http://dx.doi.org/10.1037/0033-2909.103.3.374

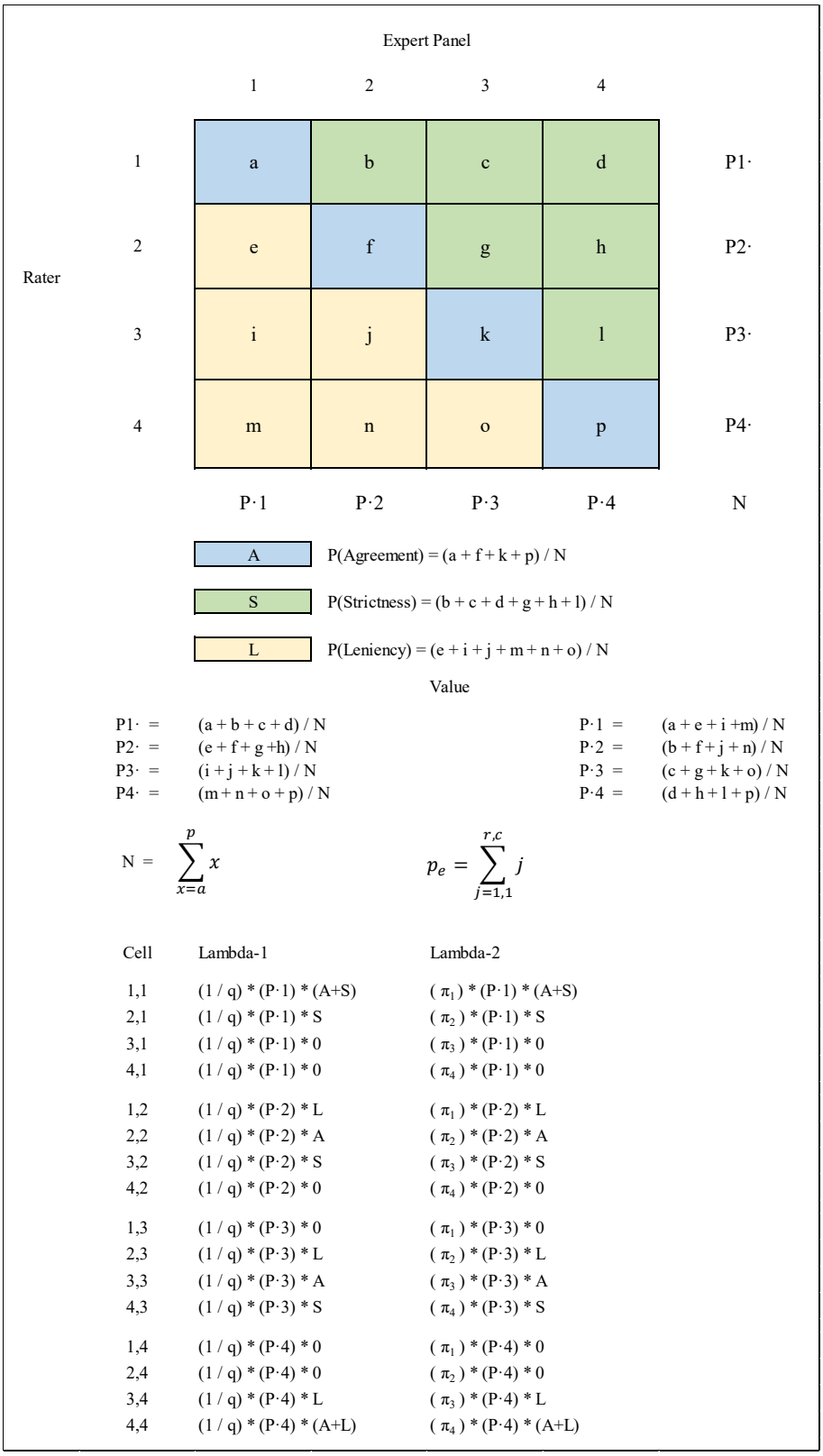

Figure 1. Calculation of Lambda for a 4x4 agreement matrix.

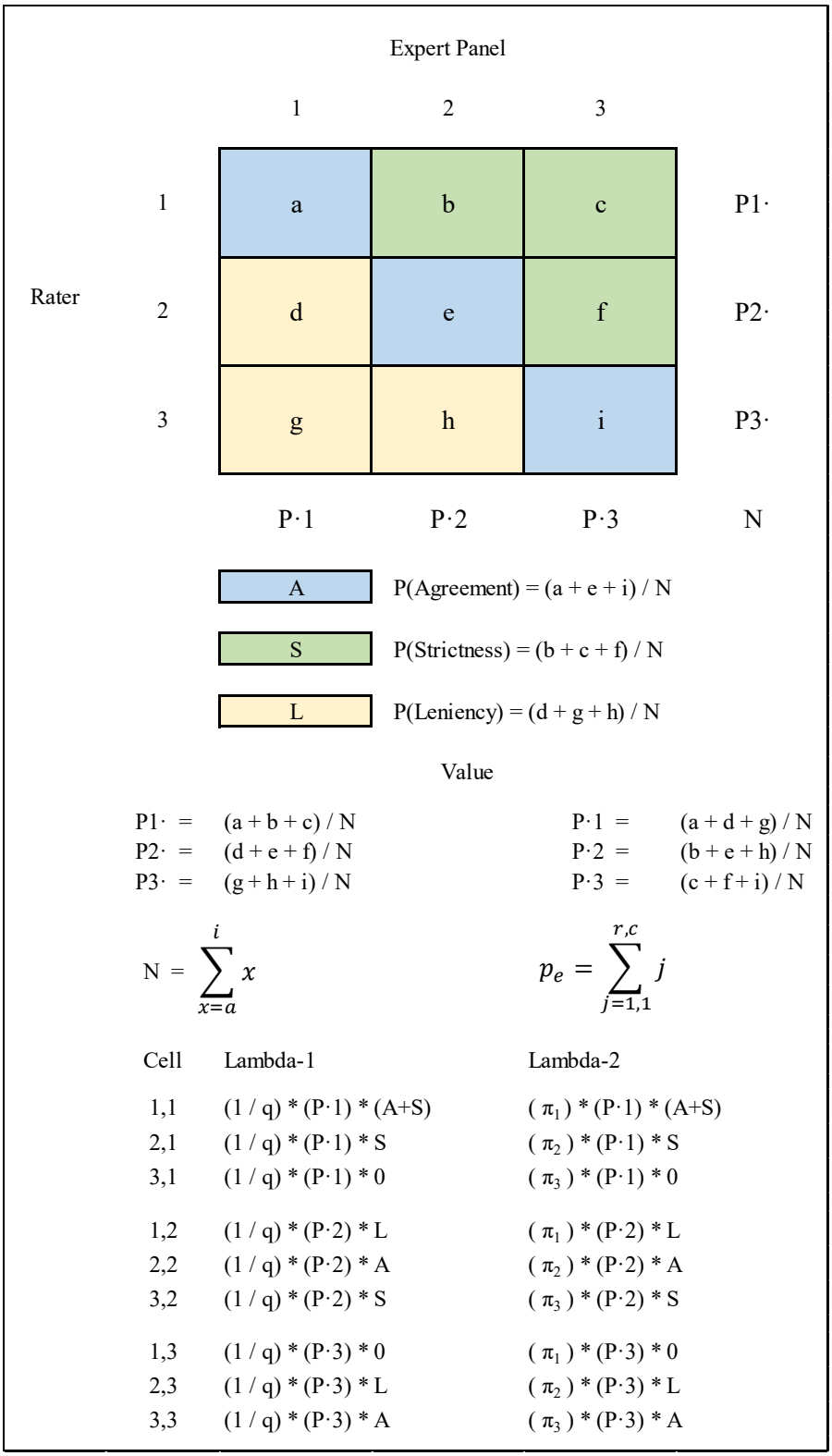

Figure 2. Calculation of Lambda for a 3x3 agreement matrix.

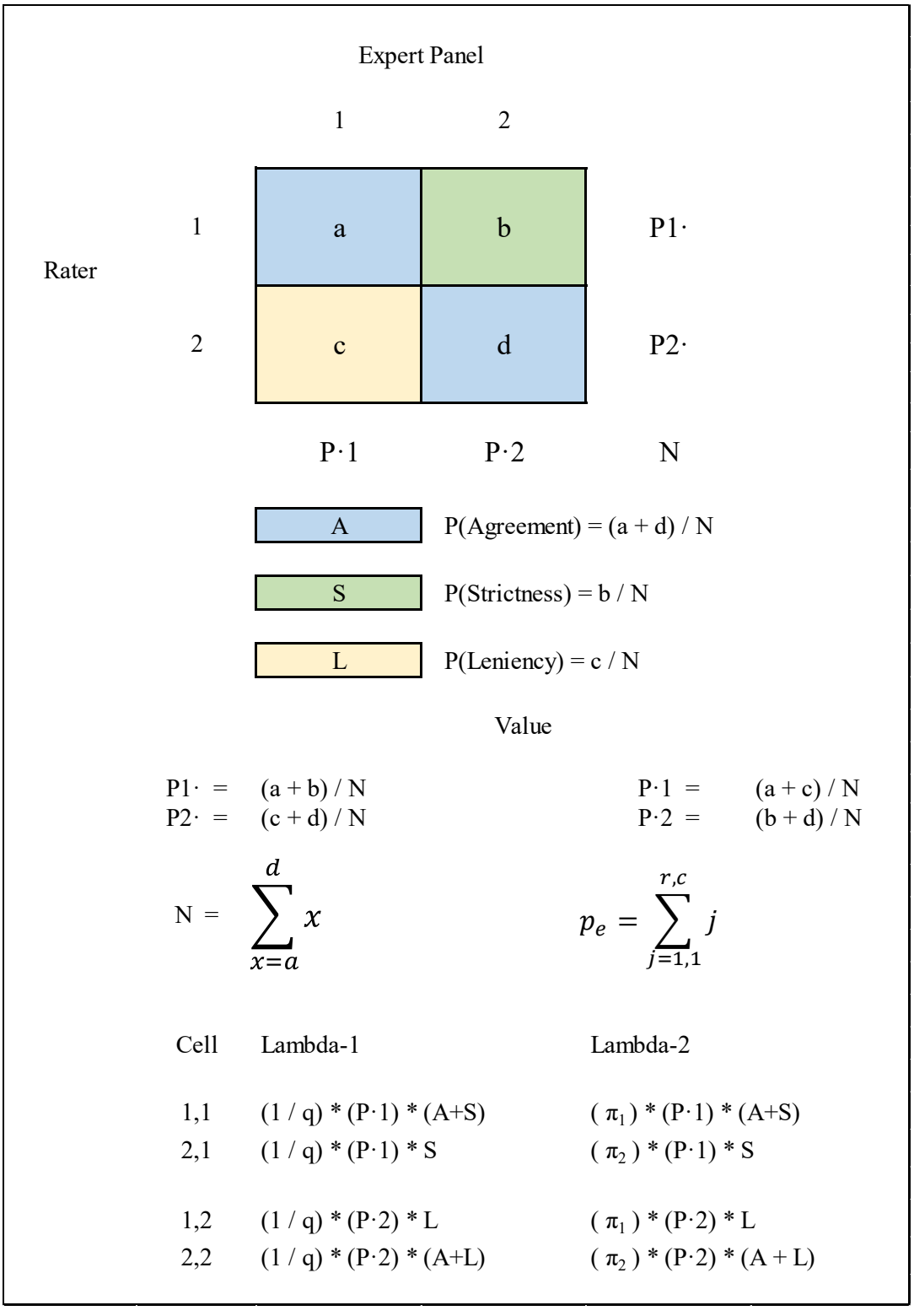

Figure 3. Calculation of Lambda for a 2x2 agreement matrix.

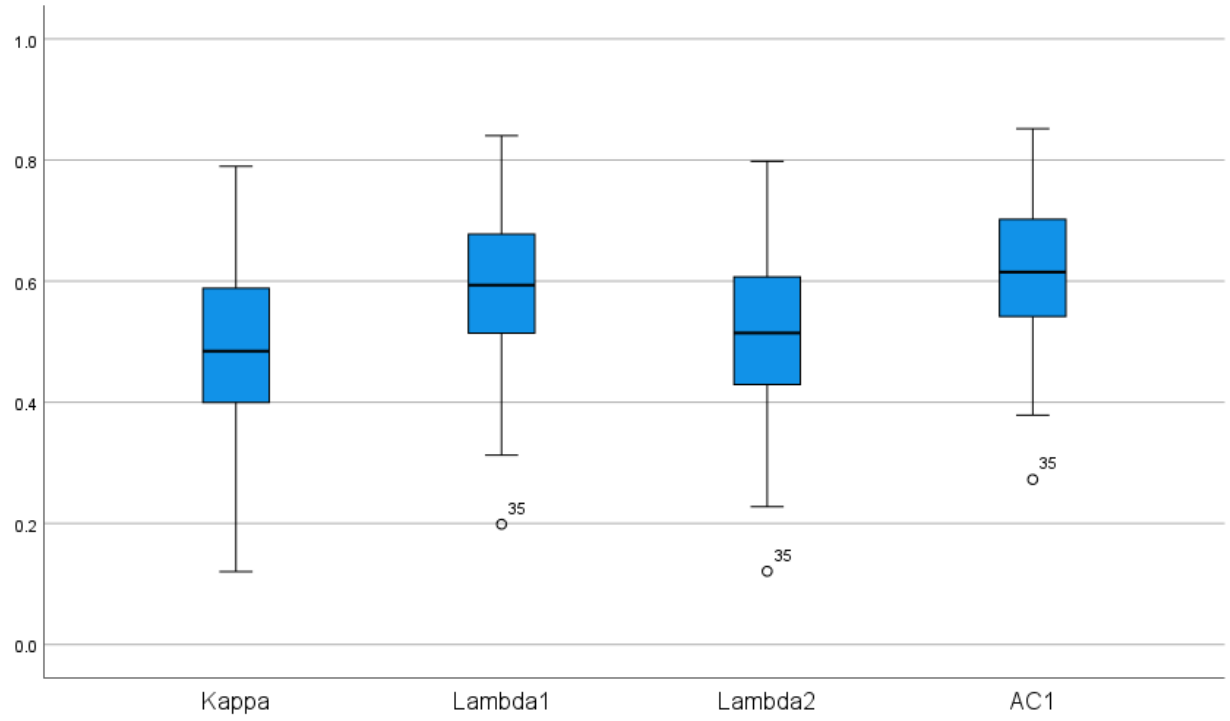

Figure 4. Boxplots of Kappa, Lambda-1, Lambda-2, and AC-1 across all raters in the sample for the exact agreement condition.

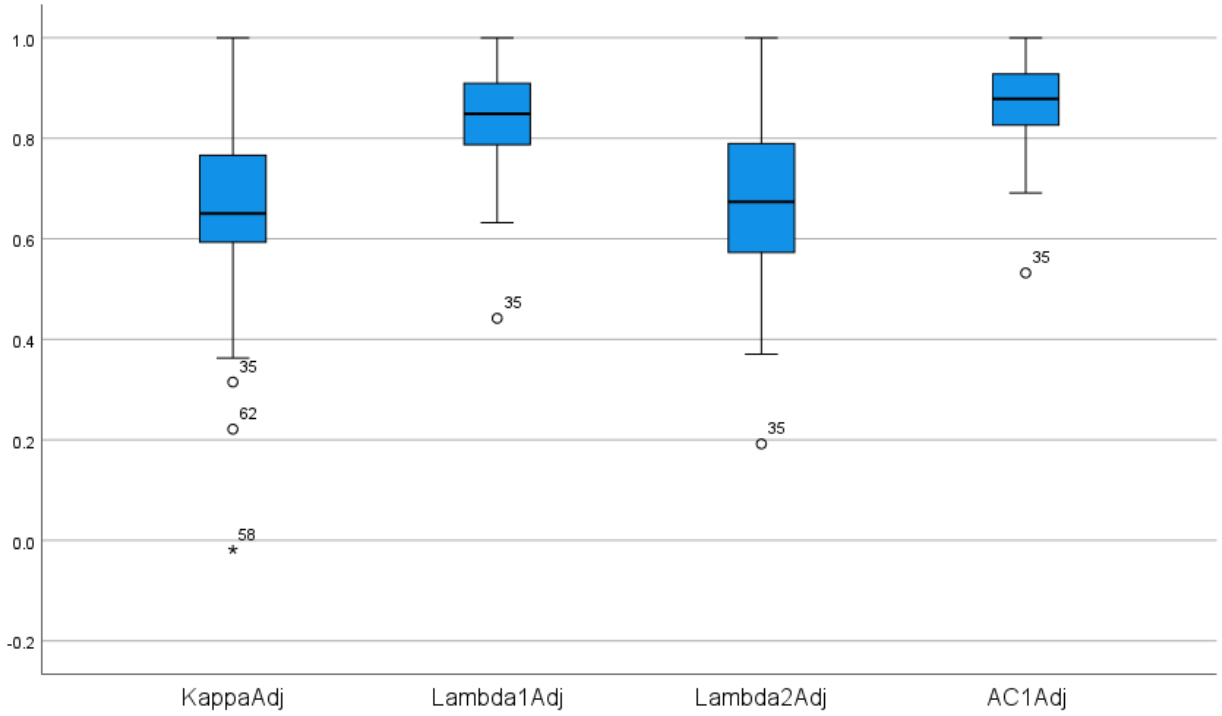

Figure 5. Boxplots of Kappa, Lambda-1, Lambda-2, and AC-1 across all raters in the sample for the adjacent agreement condition.

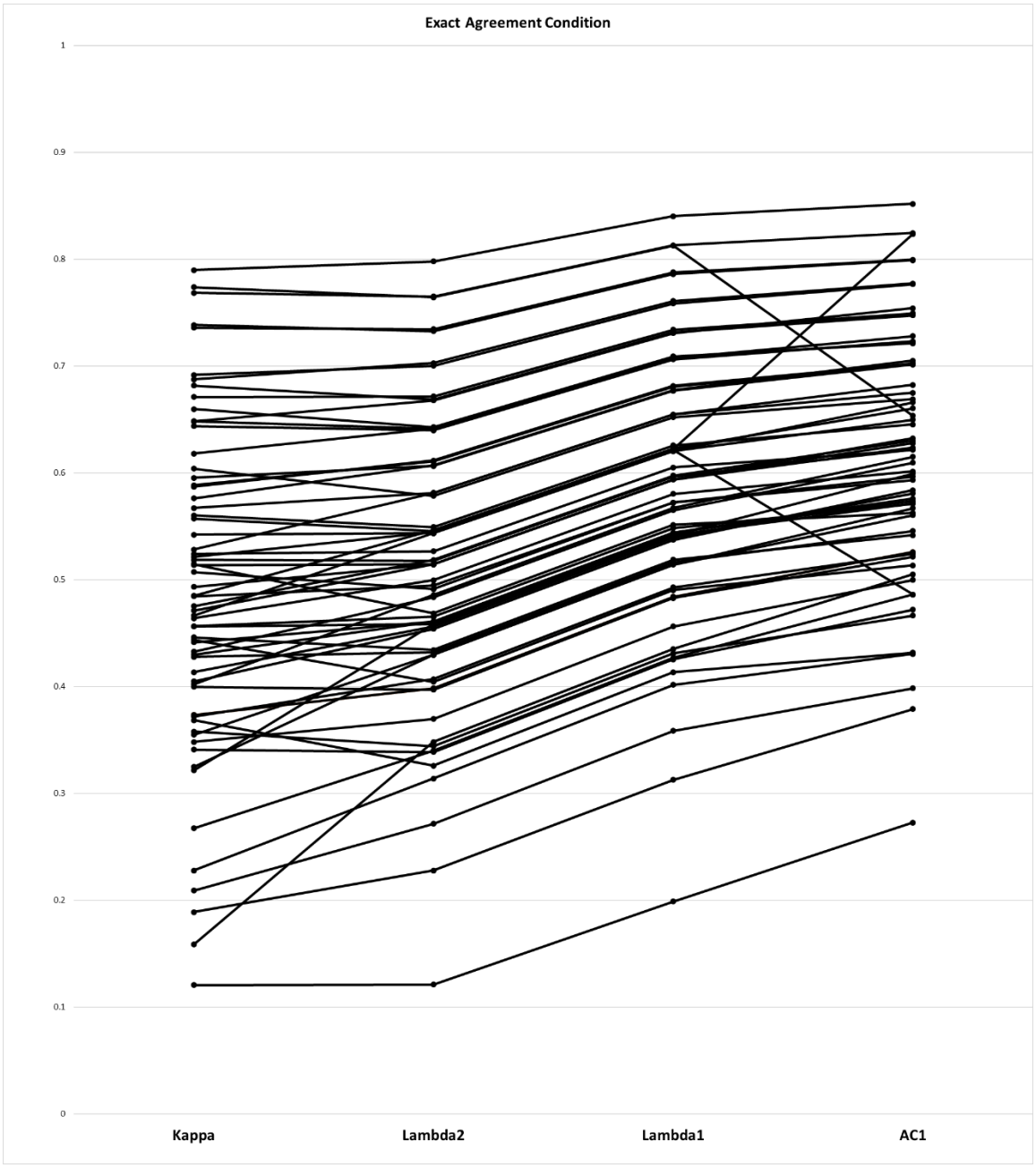

Figure 6. Kappa, Lambda-1, Lambda-2, and AC-1 for each rater under the exact agreement condition.

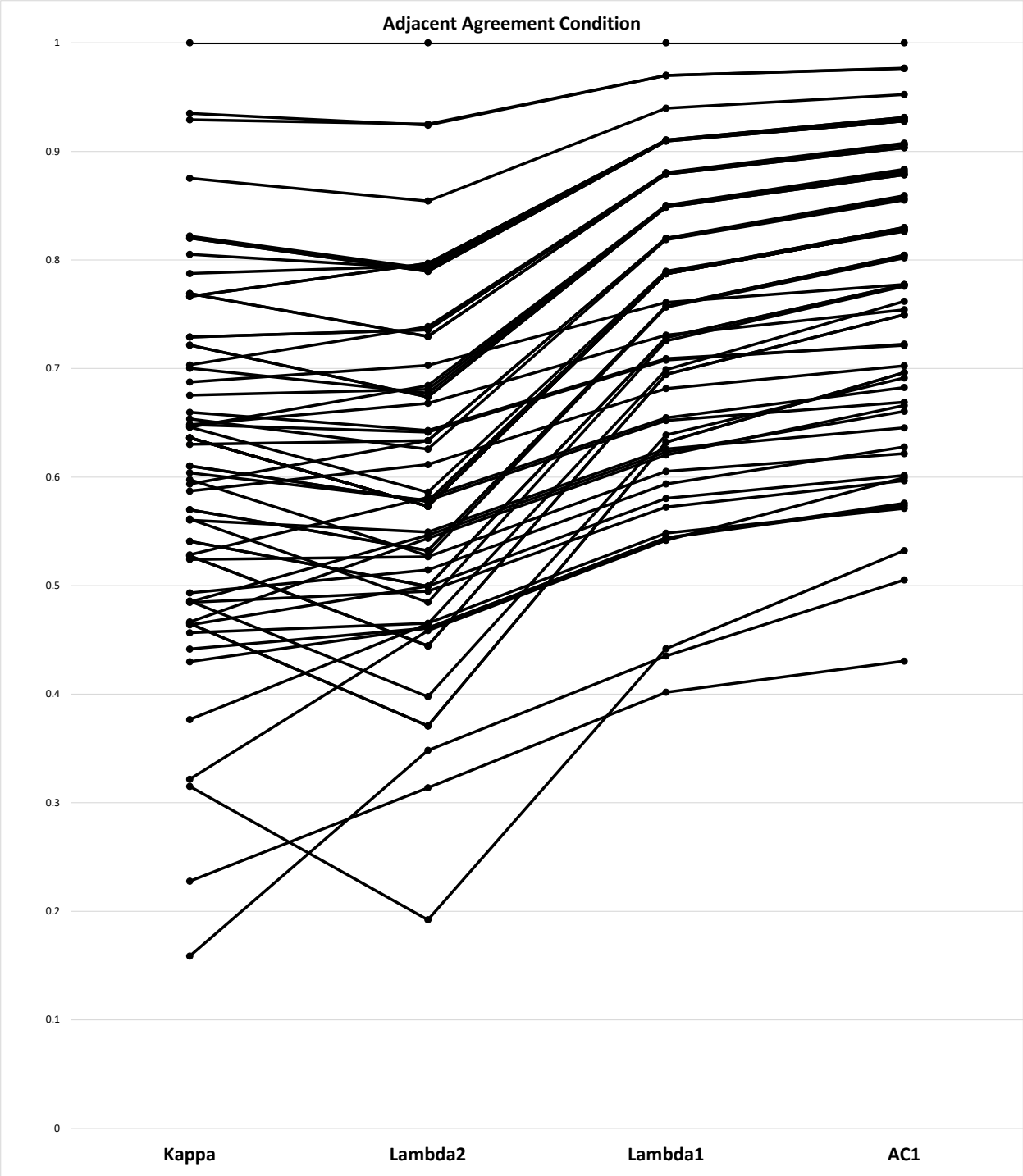

Figure 7. Kappa, Lambda-1, Lambda-2, and AC-1 for each rater under the adjacent agreement condition.

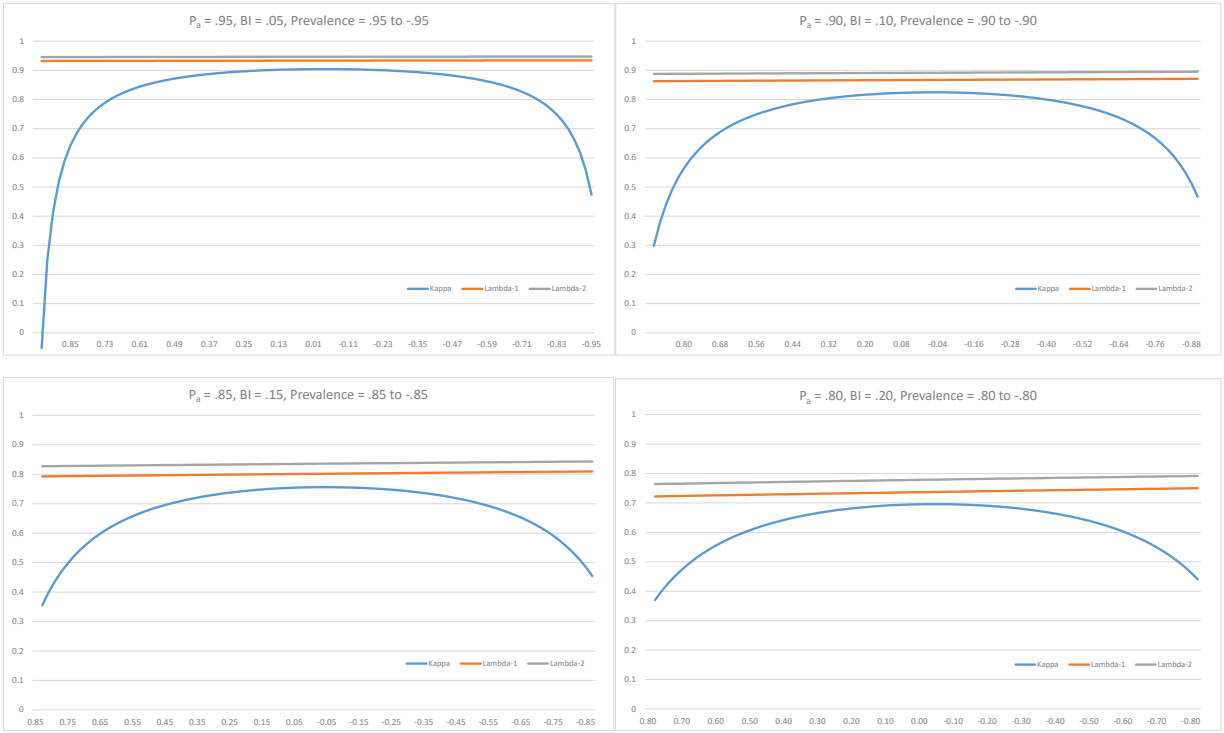

Figure 8. Simulation results. X-axis = Prevalence Index, Y-axis = Chance-corrected agreement.

Table 1

Agreement, leneincy, and strictness percentages across both the exact and adjacent agreement conditions

	Exact Agreement			Adjacent Agreement		
	Agreement	Leniency	Strictness	Agreement	Leniency	Strictness
Mean	69.3%	9.8%	21.0%	88.5%	3.4%	8.2%
SD	9.3%	8.4%	11.8%	6.6%	4.2%	7.1%
Minimum	42.0%	0.0%	2.0%	64.0%	0.0%	0.0%
25th percentile	64.0%	4.0%	10.0%	85.0%	0.0%	2.0%
Median	70.0%	8.0%	22.0%	90.0%	2.0%	8.0%
75th percentile	76.0%	14.0%	28.0%	94.0%	6.0%	12.0%
Maximum	88.0%	44.0%	58.0%	100.0%	20.0%	36.0%

Table 2
Chance corrected agreement for the exact agreement condition

	Kappa	Lambda-1	Lambda-2	AC-1	AC-2
Mean	0.489	0.587	0.511	0.615	0.507
SD	0.152	0.127	0.136	0.117	0.113
Minimum	0.121	0.199	0.121	0.273	0.356
25th percentile	0.387	0.514	0.430	0.534	0.391
Median	0.485	0.594	0.514	0.615	0.527
75th percentile	0.592	0.679	0.609	0.702	0.548
Maximum	0.790	0.840	0.798	0.852	0.706

Table 3
Classification of the evaluator performance using the Fleiss (1981) criteria

		Kappa		Lambda 1		Lambda 2		AC 1	
		n	%	n	%	n	%	n	%
Exact Agreement	Poor	15	26.3	3	5.3	11	19.3	3	5.3
	Fair	27	47.4	29	50.9	29	50.9	23	40.4
	Good	15	26.3	22	38.6	17	29.8	28	49.1
	Excellent	0	0.0	3	5.3	0	0.0	3	5.3
Adjacent Agreement	Poor	2	3.5	0	0.0	4	7.0	0	0.0
	Fair	12	21.1	1	1.8	18	31.6	1	1.8
	Good	31	54.4	21	36.8	28	49.1	9	15.8
	Excellent	12	21.1	35	61.4	7	12.3	47	82.5

Table 4
Classification of the evaluator performance using the Altman (1991) criteria

		Kappa		Lambda 1		Lambda 2		AC 1	
		n	%	n	%	n	%	n	%
Exact Agreement	Poor	2	3.5	1	1.8	1	1.8	0	0.0
	Fair	13	22.8	2	3.5	10	17.5	3	5.3
	Moderate	27	47.4	29	50.9	29	50.9	23	40.4
	Good	15	26.3	22	38.6	17	29.8	28	49.1
	Very Good	0	0.0	3	5.3	0	0.0	3	5.3
Adjacent Agreement	Poor	0	0.0	0	0.0	1	1.8	0	0.0
	Fair	2	3.5	0	0.0	3	5.3	0	0.0
	Moderate	12	21.1	1	1.8	18	31.6	1	1.8
	Good	31	54.4	21	36.8	28	49.1	9	15.8
	Very Good	12	21.1	35	61.4	7	12.3	47	82.5