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

Richard Lambert

Scott Holcomb

Bryndle Bottoms

Center for Educational Measurement and Evaluation

UNC Charlotte

An earlier edition 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 when used within the teacher performance evaluation process. Rater data also demonstrated the robustness of Lambda and Gwet's AC-1 to the data conditions known to be problematic for Kappa. All alternative chance-corrected agreement coefficients evaluated showed less variability across the 57 raters than Kappa. Simulation results demonstrated the robustness of the Lambda Coefficient of Rater-Mediated Agreement to the data conditions that are problematic for Kappa.

Examining the 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. Various questions have been raised about whether Kappa is actually correcting for chance agreement, whether it is useful for identifying and separating various sources of disagreement, the validity of Kappa coefficients when prevalence of specific categories on a rating scale is low, the validity of Kappa coefficients when agreement is high, and the generalizability of Kappa coefficients across populations and study conditions (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti; 1990; Thompson & Walter, 1988). Gwet (2008) introduced AC-1 and AC-2 as alternatives to Kappa and demonstrated they are robust indexes not susceptible to the identified shortcomings of Kappa.

This study proposes an alternative to Kappa, Weighted Kappa, AC-1, and AC-2 that is rooted in theory regarding rater-mediated assessment (Engelhard & Wind, 2018). The Lambda Coefficient of Rater-Mediated Agreement is designed for use with ordinal scales that are 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.

Most research on rater cognition focuses on the mental processes used by raters of student or examinee performance. The rating process employed by raters judging examinee constructed responses can be intricate, complex, and require subtle judgments between adjacent categories on a rubric. However the judgment of teacher performance can pose even greater complexities. In addition to the complexities of grading constructed response questions, teacher evaluators can be asked to make many ratings across multiple dimensions and base their ratings

on a complex set of indicators that can include classroom observations, interviews with students and teachers, analysis of student work, and review of artifacts from the instructional process. Similar to other examples of rater-mediated assessments, an observers' level of expertise and classroom experience, the availability of evidence and artifacts, and the overall scoring task demands, can all drive ratings of teachers (Bell et al., 2018; Suto, 2012).

Rater-Mediated Assessment Theory

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 takers, 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)

To meet this standard, validity evidence must extend beyond a review of rater training. Analysis and interpretation of raters' cognitive processes are required to ensure those processes are congruent with the measured construct (Bejar, 2012). In addition, simulated rating exercises and standards for agreement with expert raters are essential components of the validity argument. For example, it is common to provide raters a set of exemplar and non-exemplar responses to serve as anchors or benchmarks for scoring decisions.

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 in understanding what is actually being measured by an assessment (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).

The lens model proposed by Brunswik (1952) was initially designed as a human judgment and decision-making model. 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). This connection resides within the process through which raters interact with the items and the rating scale. In the case of teacher performance evaluation, observers interact with a multidimensional rubric which requires raters to have a high level of expertise, skill in interpreting a wide range of indicators, and an ability to analyze evidence to arrive at placements on a rating scale.

The following set of assumptions about raters of teacher performance, and the complex response process they use to arrive at ratings, serve as a theoretical foundation for the Lambda Coefficient of Rater-Mediated Agreement. These assumptions are rooted in our work as trainers, managers, and observers who are charged with ensuring the validity of one state's teacher performance evaluation process. 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 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 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.

Furthermore, we posit 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 performance evaluation.

The Lambda Coefficient of Rater-Mediated Agreement

Cohen (1960) introduced the Kappa coefficient of chance-corrected agreement. Kappa 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:

$$\kappa = (p_a - p_e) / (1 - p_e) \quad (1)$$

$$\sigma_\kappa = \sqrt{ \{ [p_a(1 - p_a)] / [n(1 - p_e)^2] \} } \quad (2)$$

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 (Brennan & Prediger, 1981; Bryt et al., 1993; Gwet, 2008; Holley & Guilford, 1964; Jason & Vegelius, 1979; Krippendorff, 1970; Maxwell, 1970; Perreault & Leigh, 1989). In addition, Bennett et al. (1954) proposed a method equivalent to the method reintroduced by Bryt et al. (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 that 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. We are proposing the 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, interviews with teachers, classroom conditions and artifacts, student work samples, and general overall impressions. Raters bring their own personal tendencies toward strictness or leniency, or even biases, to this rating process. 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 chance-corrected agreement coefficient for use in a specific, yet commonly occurring rating situation: raters using holistic scoring to arrive at placements on an ordinal scale. The goal was to evaluate individual agreement with an expert panel that has achieved consensus regarding "correct" answers. Therefore, we were not interested in creating a measure of the consistency of a group of ratings, but rather focused on the accuracy of individual raters relative to a standard. The target was a descriptive statistic about the behavior of individual raters that may provide useful information within a training or inter-rater reliability (IRR) certification process.

We established several desirable criteria for the proposed statistic. First, Lambda had to agree with Kappa when all ratings fall on the main diagonal of the ratings matrix formed by completely crossing an evaluators ratings with those of an expert panel. 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 has a reasonable upper bound on the magnitude of the correction for chance, similar to Gwet's approach (2008). This upper bound for both Lambda and Gwet's coefficients is set to .50. This value has an intuitive and practical appeal in addition to its mathematical advantages. When raters use an ordinal rating scale, "guessing" often resides within deliberations between adjacent steps on the rating scale. Raters may struggle to resolve uncertainty between adjacent steps on a rating scale more often than they randomly select a rating from across an entire rating scale. 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.

$$p_e \leq 2L / q \quad (3)$$

$$p_e \leq 2S / q \quad (4)$$

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 Lambda (λ), applicable to both Lambda-1 (λ_1) and (λ_2), and to rating scales with any number of steps, can be expressed as:

$$\lambda = (p_a - p_e) / (1 - p_e) \quad (5)$$

$$p_e = \Sigma \; p_s p_c p_f \quad (6)$$

$$\sigma_\lambda = \sqrt{ \{ \; [p_a(1 - p_a)] \; / \; [n(1 - p_e)^2] \; \} } \quad (7)$$

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 the rater is uncertain about which rating to give, arrives at a random starting point for their deliberations, and is 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 proportion of the total ratings given in the population case associated with the cell in question. This value is the marginal proportion for the given row in the agreement matrix. This value also assumes the rater is uncertain about which rating to give, uses guessing as a means to arrive at a starting point for their deliberations, and their internal guessing process weights the points on the rating scale according to how frequently they encounter each level on the rating scale in the population. This approach assumes the population proportions of ratings are known and expectations regarding the skill levels evaluators typically encounter in

the field influence an evaluator's selection of starting point for deliberation. So for example, if a rater rarely encounters a teacher with the skill level associated with a particular point on the rating scale, λ_2 assumes the rater would be much less likely to select that point as a starting point for deliberations.

It is important to note that when all steps on a ratings scale are equally likely in the population, $\lambda_1 = \lambda_2$. To illustrate how λ_1 and $\lambda_2$work in practice, see Figure 1 for a two-point ordinal rating scale. See Figure 2 for a three-point ordinal rating scale and Figure 3 for a four-point ordinal rating scale. Just for illustration purposes, we have included category labels that might apply to a teacher performance evaluation rubric.

The Current Study

The purpose of this study was twofold. First, the researchers sought to test the performance of λ relative to Kappa, AC-1, and AC-2 using field data. Second, the researchers sought to evaluate λ 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 IRR certification exercise. Evaluators (n=57) made placements on five progressions across each of 10 online teacher profiles for a total of 2,850 ratings. The evaluators rated 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 for teachers of all licensure levels. The measure includes five progressions, each of which measures performance relative to a specific teaching standard. Each progression 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 by comparing an evaluator's ratings to the "correct answer" which consisted of consensus ratings from a panel of five experts. 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, f.) Gwet's AC-2, and g.) the level of each rater's performance according to well established classification systems (Altman, 1991; Fleiss, 1981).

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 support and 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, f.) Gwet’s AC-2, and g.) the level of each rater’s performance according to well established classification systems (Altman, 1991; Fleiss, 1981). 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.

Results

First, we examined the distribution of the agreement, strictness, and leniency percentages for all 57 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 57 evaluators was 69.6% (SD=9.5) and values ranged from 42.0% to 88.0%. The mean percent lenient for exact agreement was 8.1% (SD=6.5) and values ranged from 0.0% to 26.0%. The mean percent strict for exact agreement was 22.2% (SD=11.6) and values ranged from 4.0% to 58.0%. Therefore, the raters as a group displayed moderate

levels of agreement and were more likely to be strict than lenient. However, a substantial minority of raters (n = 7, 12.3%) agreed with the expert panel for less than 60% of their ratings.

As expected, agreement percentages increased, and strictness and leniency percentages decreased, for the adjacent agreement condition. Only four of the 57 raters (7.0%) displayed adjacent agreement percentages less than 80%. The mean percent adjacent agreement across the 57 evaluators was 88.6% (SD=6.8) and values ranged from 64.0% to 100.0%. The mean percent lenient for adjacent agreement was 2.4% (SD=2.8) and values ranged from 0.0% to 10.0%. The mean percent strict for adjacent agreement was 9.0% (SD=7.2) 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 .497 (SD=.152) and values ranged from .121 to .790. The mean Lambda-1 Coefficient of Rater-Mediated Agreement for exact agreement was .591 (SD=.131) and values ranged from .199 to .840. The mean Lambda-2 Coefficient of Rater-Mediated Agreement for exact agreement was .515 (SD=.141) and values ranged from .121 to .798. The mean Gwet's AC-1 for exact agreement was .618 (SD=.121) and values ranged from .273 to .852. The mean Gwet's AC-2 for exact agreement was .472 (SD=.092) and values ranged from .356 to .548. However, Gwet's AC-2 was undetermined for 51 of the 57 raters. Figure 4 displays these distributions as boxplots. AC-2 was not included due to small sample size (n = 6). The boxplots show that λ_1, λ_2, and Gwet's AC-1 yielded less severe and less variable corrections for chance

agreement than Kappa. λ_1, λ_2, and AC-1 showed sensitivity to one outlier rater not detected by Kappa.

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 almost all raters. λ_1 yielded coefficients that were consistently higher than Kappa and λ_2 and lower than AC-1. AC-1 emerged as yielding the consistently highest coefficients. 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 λ_1 and AC-1 as indicated by several lines that break from the overall pattern.

The mean Kappa for adjacent agreement was .686 (SD=.145) and values ranged from .315 to 1.000 (see Table 3). The mean λ_1 for adjacent agreement was .828 (SD=.104) and values ranged from .442 to 1.000. The mean λ_2 for adjacent agreement was .661 (SD=.166) and values ranged from .192 to 1.000. The mean Gwet's AC-1 for adjacent agreement was .861 (SD=.087) and ranged from .532 to 1.000. The mean Gwet's AC-2 for adjacent agreement was .787 (SD=.052) and values ranged from .687 to .831. However, again Gwet's AC-2 was undetermined for 51 of the 57 raters. Figure 5 displays these distributions as boxplots. AC-2 was not included due to small sample size (n = 6). The boxplots show that λ_1, λ_2, and Gwet's AC-1 yielded less severe and less variable corrections for chance agreement than Kappa. λ_1 and AC-1 yielded similar distributions and displayed less variability than Kappa or λ_2. All four coefficients detected the same outlier rater.

A very consistent rank order of correction among the alternatives to Kappa emerged for the adjacent agreement condition as well (see Figure 7). λ_2 was closest to Kappa for almost all raters. λ_1 coefficients were higher than Kappa and λ_2 and lower than AC-1. AC-1 yielded the

highest coefficients for almost all raters. AC-2 was again not included in these comparisons due to the small sample size (n = 6). 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 research literature regarding chance-corrected agreement contains several systems for classifying coefficients as above or below acceptable levels. Altman's system (1991) includes the following criteria: poor (.00 - .20), fair (>.20 - .40), moderate (>.40 - .60), good (>.60 - .80, and very good (>.80). Similarly, Fleiss (1981) offered the following criteria: poor (≤.40), fair (.40 - .60), good (>.60 - .80), and excellent (>.80). As a final examination of evidence to address research question one, the coefficients from the current study were classified according to these two sets of standards. Table 4 contains the results of applying Altman's criteria (1991) to the Kappa, λ_1, λ_2, and AC-1 coefficients for all 57 evaluators and Table 5 contains the results of applying Fleiss' criteria (1981). Good and Excellent according to Fleiss (1981) and Good and Very Good according to Altman (1991) use the same criteria and therefore produce the same classifications. If these levels were used to indicate acceptable evaluator performance, only 26.3% of evaluators would have passed the exercise using Kappa. Using λ_2, 29.8% would have passed, followed by 43.9% for λ_1 and 54.4% for AC-1. As expected, the passing rates would have been quite higher using the adjacent agreement scoring procedure: Kappa (75.6%), λ_2 (61.4%), λ_1 (98.2%), and AC-1 (98.2%). Therefore, substantially more evaluators would have passed the IRR exercise using either λ_1 or AC-1.

The two systems differ in two ways. Altman (1991) includes a moderate agreement level while Fleiss (1981) does not. Fleiss (1981) suggests a broader range for poor agreement (≤.40) compared to Altman's (1991) (≤.20). Very few evaluators, two or fewer, met Altman's criteria for "poor" across all four coefficients using either the exact or adjacent agreement methods.

However, using Fleiss' criteria, a substantial minority of evaluators would have been classified as exhibiting "poor" agreement using exact agreement and Kappa (26.3%). The exact agreement method and λ_2 also identified a substantial minority of evaluators as exhibiting poor agreement (19.3%). Exact agreement using λ_1 or AC-1 classified only three evaluators in the poor agreement category (5.3%)

A simulation study addressed research question two. The simulation design extended the approach of Xie (2013) to include λ_1 and λ_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. 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, λ_1, and λ_2 for each of the four conditions across the applicable range of the Prevalence Index.

For condition 1 (95% agreement), the mean Kappa was .820 (SD = .147) and values range from -.053 to .904. The mean λ_1 was .933 (SD = .001) and values range from .932 to .934. The mean λ_2 was .842 (SD = .094) and values range from .487 to .907. Therefore, λ_1 yielded very consistent values (see Figure 8), Kappa yielded very inconsistent and over-corrected values, and

λ_2 followed a similar pattern as Kappa but did not over-correct as much when the Prevalence Index was high. 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 λ_1 was .867 (SD = .002) and values range from .863 to .871. The mean λ_2 was .752 (SD = .090) and values range from .474 to .826. For condition 3 (85% agreement), the mean Kappa was .674 (SD = .095) and values range from .355 to .756. The mean λ_1 was .801 (SD = .005) and values range from .793 to .810. The mean λ_2 was .683 (SD = .080) and values range from .460 to .755. For condition 4 (80% agreement), the mean Kappa was .621 (SD = .082) and values range from .370 to .696. The mean λ_1 was .737 (SD = .008) and values range from .722 to .750. The mean λ_2 was .625 (SD = .070) and values range from .444 to .691. For all four conditions, Lambda-2 tended to over-correct less than Kappa for high values of the Prevalence Index, converge with Kappa as the Prevalence Index got smaller, and yield nearly identical values to Kappa at the minimum values of the Prevalence Index. λ_1 remained very consistent within each of the simulated conditions.

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 and 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 57 raters, agreement was high (93.33%) for this profile for both the exact and adjacent methods. However, Cohen's Kappa was .000 for the exact method while λ_1 equaled .913 and AC-1 equaled .932. Cohen's Kappa was also .000 for the adjacent method while λ_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 λ_1 equaled .508 and AC-1 equaled .589. Cohen's Kappa was also .000 for the adjacent method while λ_1 equaled .947 and AC-1 equaled .964.

It is important to point out that λ 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. λ is a single coefficient and is agnostic to where in the rating space strictness or leniency occurs. It cannot detect or identify the 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 support, retrain, or recertify 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. Lambda is 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 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. Future research is needed to test Lambda across a wider range of field and simulated data conditions.

References

Altman D.G. (1991). *Practical Statistics for Medical Research.* Chapman and Hall.

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.

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.

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*.

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.

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 perspectivs 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. (1981). *Statistical methods for rates and proportions*. Wiley.

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

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.

Joint Committee on the Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (AERA/APA/NCME, 2014). *The Standards for Educational and Psychological Testing.* American Psychological Association.

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.

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*.

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.*

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. Washington, D.C.: Federal Committee on Statistical Methodology, U.S. Office of Management and Budget. Retrieved 5/3/2021 from https://nces.ed.gov/FCSM/pdf/J4_Xie_2013FCSM.pdf.

Table 1

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

	Exact Agreement			Adjacent Agreement		
	Agreement	Leniency	Strictness	Agreement	Leniency	Strictness
Mean	69.6%	8.1%	22.2%	88.6%	2.4%	9.0%
SD	9.5%	6.5%	11.6%	6.8%	2.8%	7.2%
Minimum	42.0%	0.0%	4.0%	64.0%	0.0%	0.0%
25th percentile	64.0%	3.0%	13.0%	86.0%	0.0%	4.0%
Median	70.0%	8.0%	22.0%	90.0%	2.0%	8.0%
75th percentile	77.0%	12.0%	28.0%	94.0%	5.0%	14.0%
Maximum	88.0%	26.0%	58.0%	100.0%	10.0%	36.0%

Table 2
Chance corrected agreement for the exact agreement condition

	Kappa	Lambda-1	Lambda-2	AC-1	AC-2
Mean	0.497	0.591	0.515	0.618	0.472
SD	0.152	0.131	0.141	0.121	0.092
Minimum	0.121	0.199	0.121	0.273	0.356
25th percentile	0.387	0.514	0.430	0.534	0.359
Median	0.485	0.594	0.514	0.622	0.512
75th percentile	0.611	0.694	0.625	0.713	0.547
Maximum	0.790	0.840	0.798	0.852	0.548

Table 3
Chance corrected agreement for the adjacent agreement condition

	Kappa	Lambda-1	Lambda-2	AC-1	AC-2
Mean	0.686	0.828	0.661	0.861	0.787
SD	0.145	0.104	0.166	0.087	0.052
Minimum	0.315	0.442	0.192	0.532	0.687
25th percentile	0.604	0.787	0.573	0.827	0.763
Median	0.672	0.849	0.674	0.878	0.799
75th percentile	0.769	0.910	0.791	0.928	0.822
Maximum	1.000	1.000	1.000	1.000	0.831

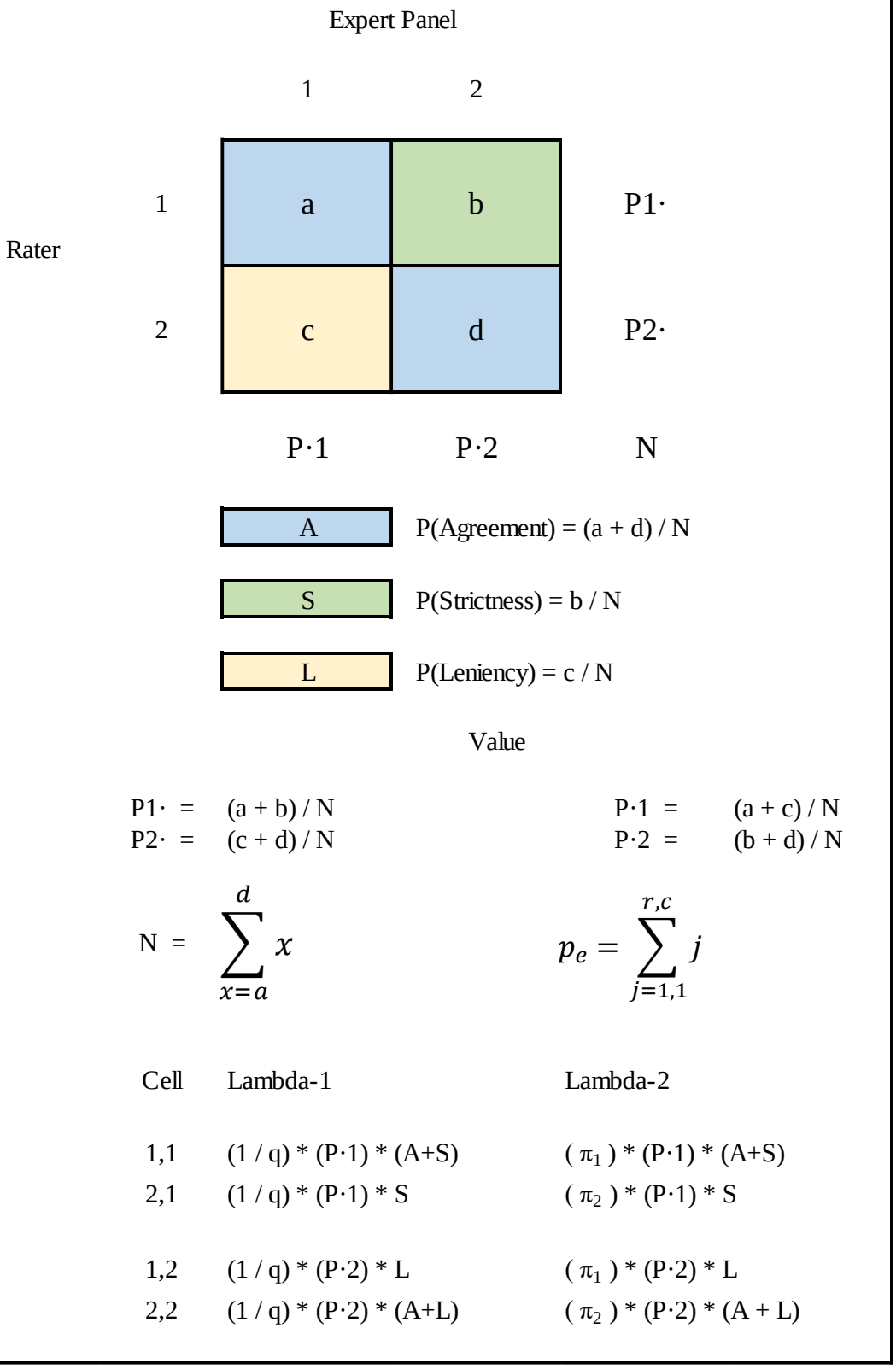

Figure 1. Calculation of Lambda for a 2x2-agreement matrix.

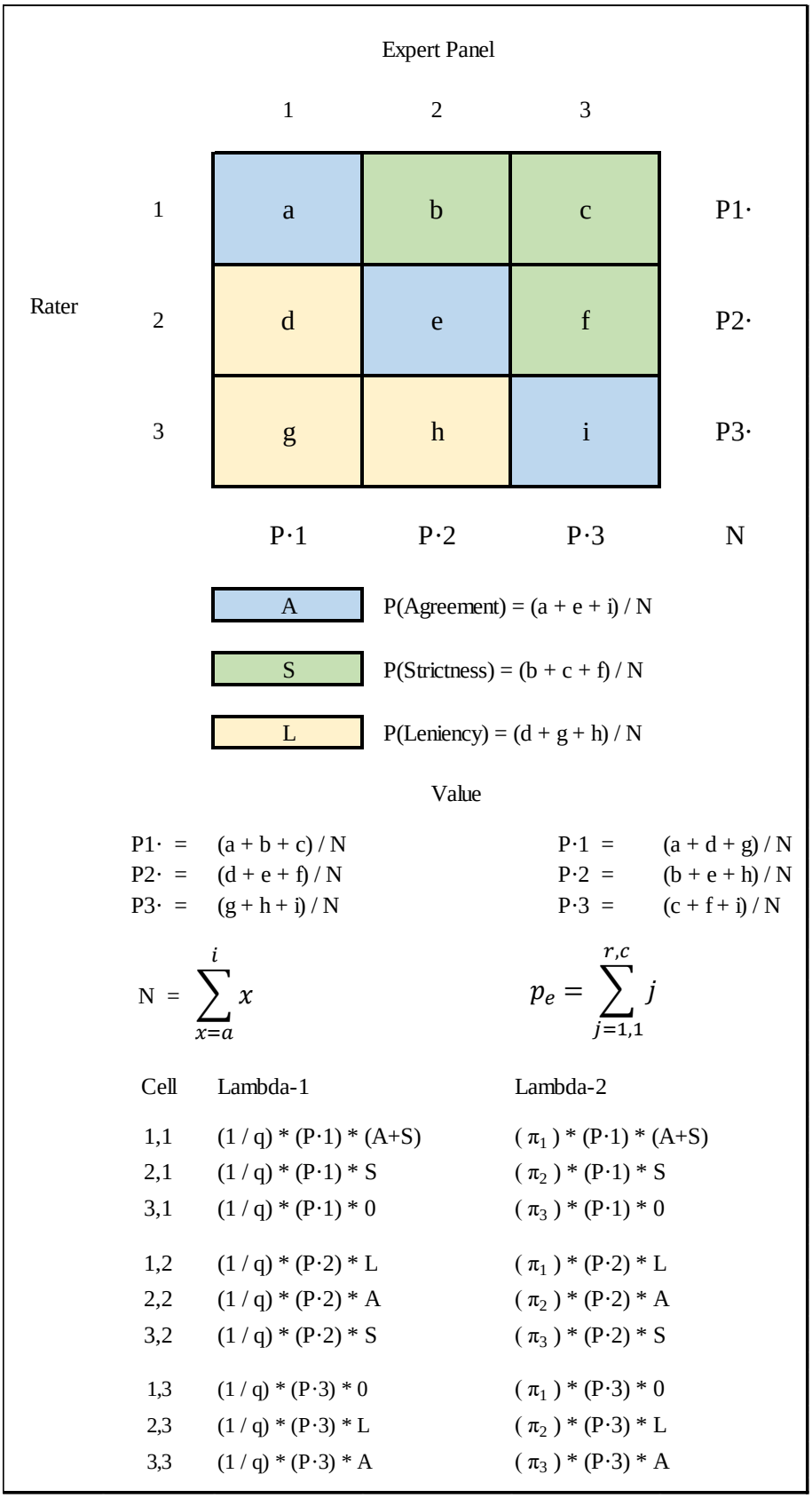

Figure 2. Calculation of Lambda for a 3x3-agreement matrix.

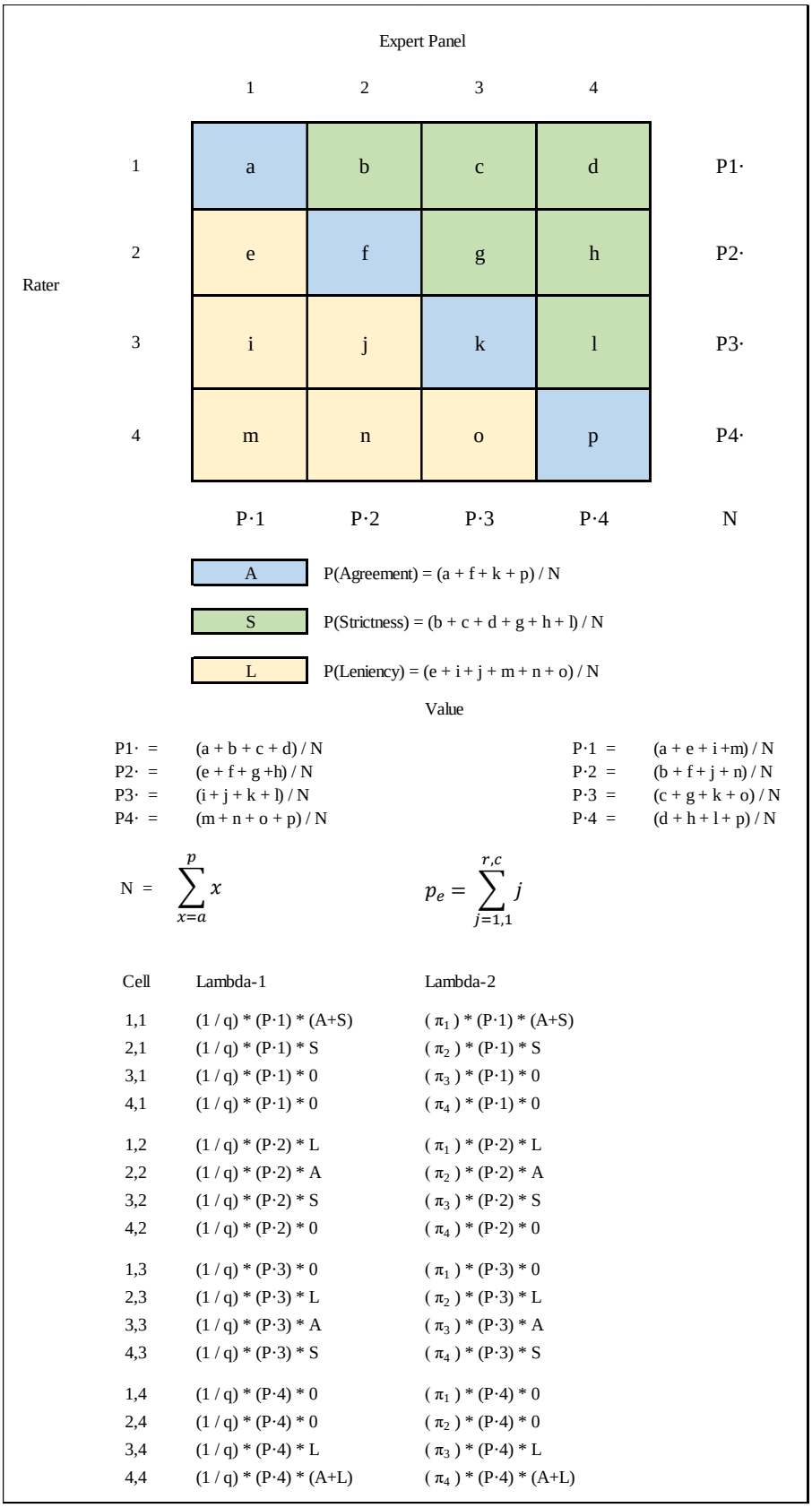

Figure 3. Calculation of Lambda for a 4x4-agreement matrix.

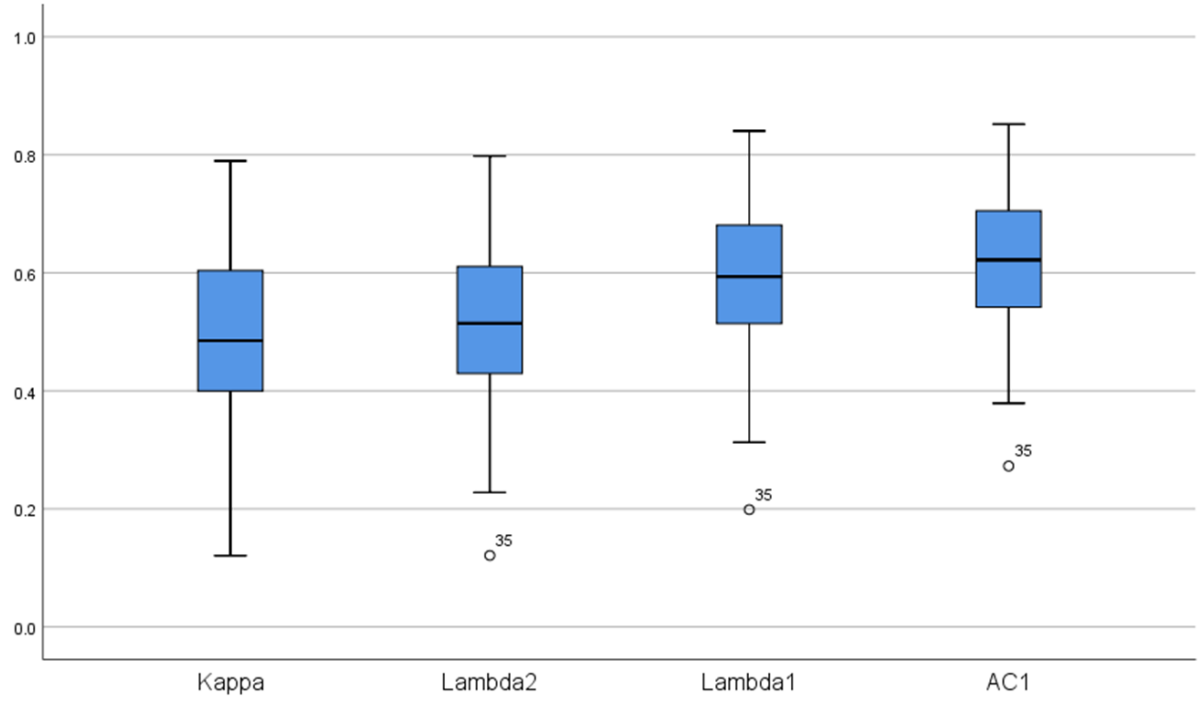

Figure 4. Boxplots of Kappa, Lambda-1, Lambda-2, and AC-1 across all raters in the sample for the exact agreement condition.

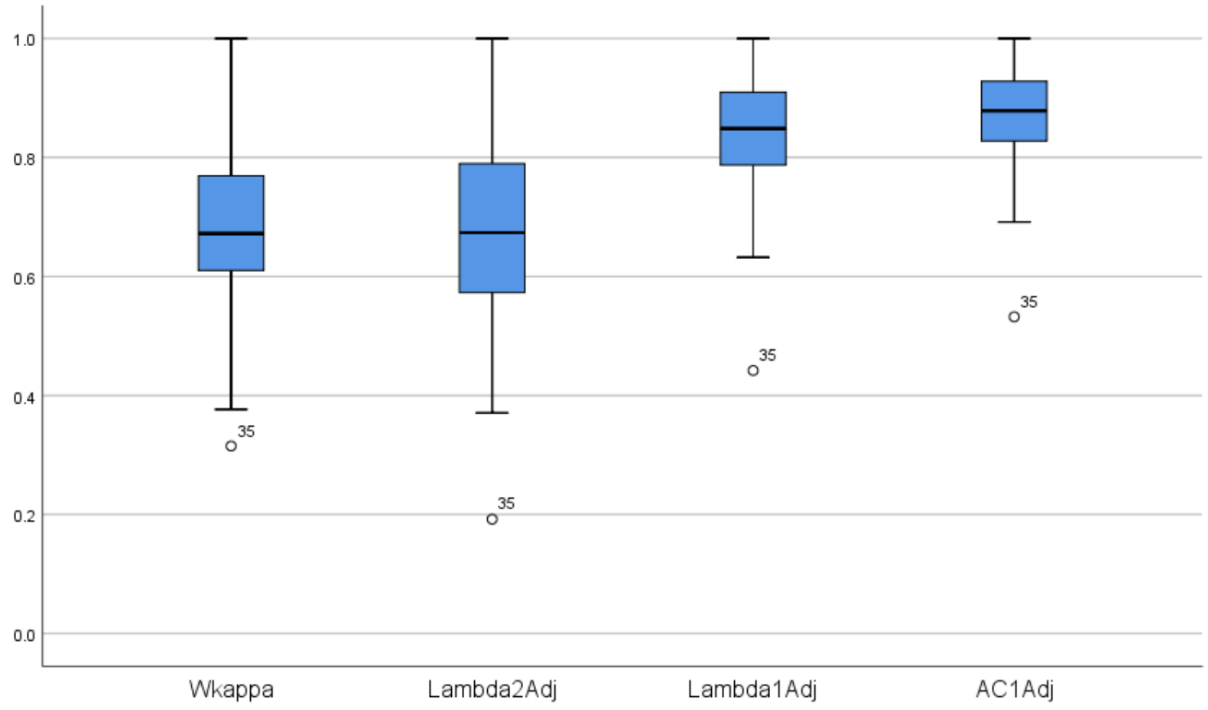

Figure 5. Boxplots of Kappa, Lambda-1, Lambda-2, and AC-1 across all raters in the sample for the adjacent agreement condition.

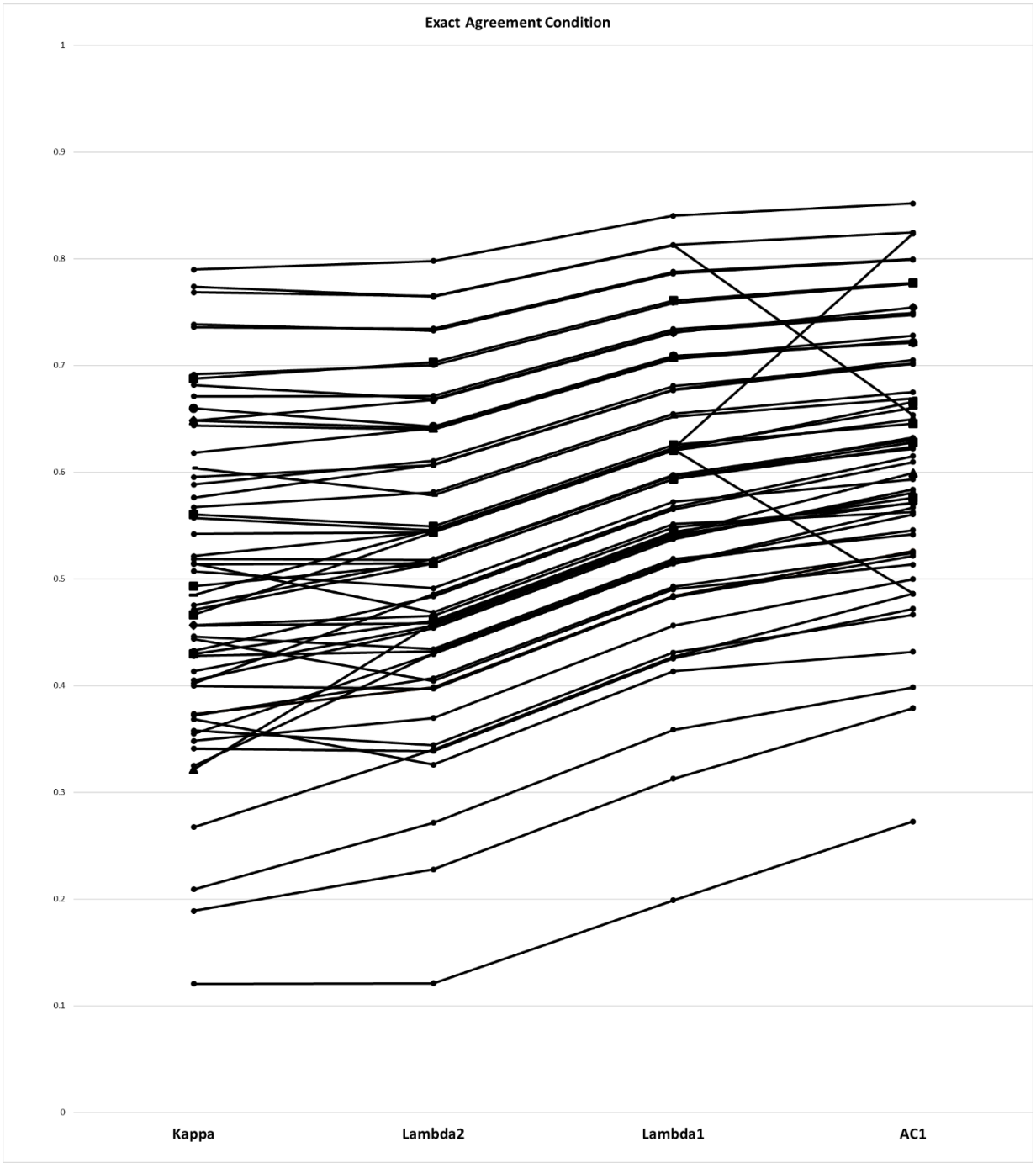

Figure 6. Kappa, Lambda-1, Lambda-2, and AC-1 for each rater under the exact agreement condition.

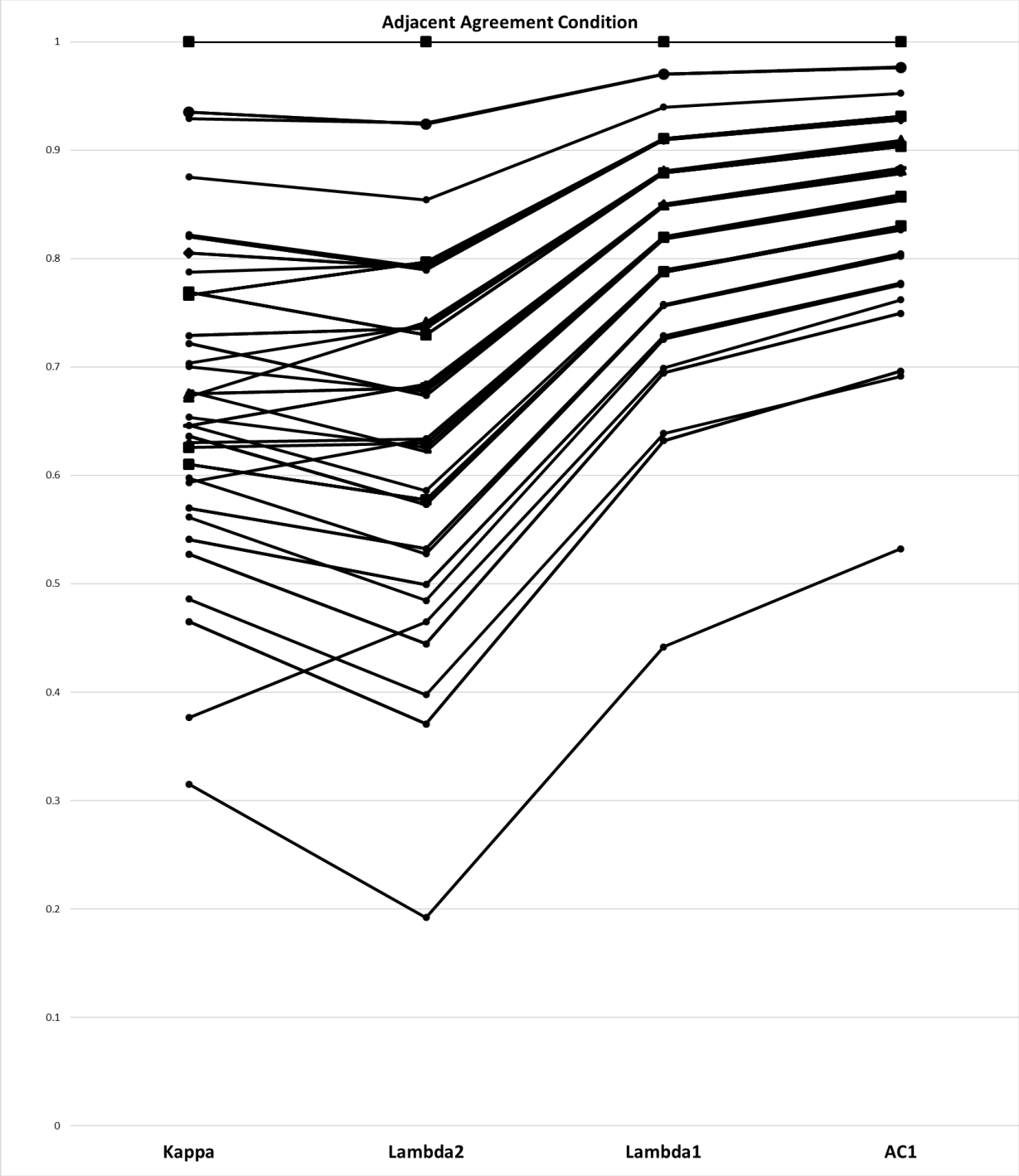

Figure 7. Kappa, Lambda-1, Lambda-2, and AC-1 for each rater under the adjacent agreement condition.

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

Table 5

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

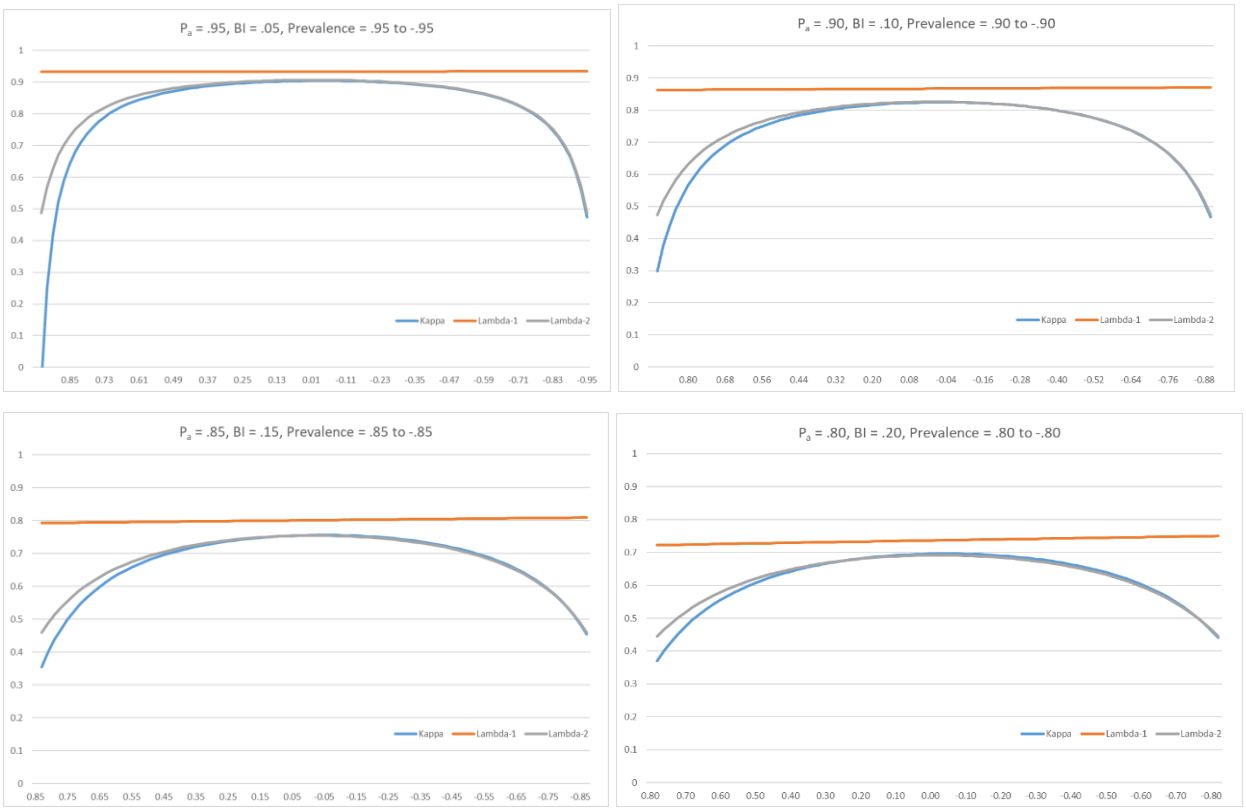

Figure 8. Simulation results. X axis = Prevalence Index, Y axis = Chance-corrected agreement.