LAMBDA COEFFICIENT OF RATER-MEDIATED AGREEMENT: EVALUATION OF AN ALTERNATIVE CHANCE-CORRECTED AGREEMENT COEFFICIENT

by

Timothy Scott Holcomb

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Educational Research, Measurement and Evaluation

University of North Carolina at Charlotte

2022

Approved by:

_____
Dr. Richard Lambert

_____
Dr. Stella Kim

_____
Dr. Carl Westine

_____
Dr. Anne Cash

ABSTRACT

TIMOTHY SCOTT HOLCOMB. Lambda Coefficient of Rater-Mediated Agreement: Evaluation of an Alternative Chance-Corrected Agreement Coefficient. (Under the direction of DR. RICHARD G. LAMBERT)

In this study, the performance of the Lambda Coefficient of Rater-Mediated Agreement was evaluated with other chance-corrected agreement coefficients. Lambda is grounded in rater-mediated assessment theory and was developed as an alternative to Kappa (Cohen, 1960) and other chance-corrected agreement coefficients. Lambda has two variations, a general form that is calculated similarly to how most chance-corrected agreement coefficients are calculated, such as Kappa (Lambert et al., 2021). The general form of Lambda is referred to as Lambda-1. Lambda-2 differs from Lambda-1 in the calculation of the proportion of expected chance agreement. Lambda-2 uses known population proportions when available and applies those proportions in the calculation of expected chance agreement. In total, six coefficients were calculated using generated data by varying the amount and location of agreement and disagreement between ratings across two-, three-, and four-point rating scales. The exact agreement specifications ranged from 75% to 95% across 115 planned data conditions. The simulations adjusted prevalence indices according to exact agreement specifications (Xie, 2013). Results demonstrated the robustness of Lambda-1 and Lambda-2 to data conditions that are problematic for other coefficients. Both variations of Lambda produced benchmark agreement results that maintained meaning that may be diminished by other coefficients.

DEDICATION

I dedicate this dissertation to my wife, Anna Brooks, and our children, Harper, Elle, and Nora. Each of you made countless sacrifices during this endeavor and I could not have made it without your support. Anna Brooks, your encouragement and love made this dissertation possible. Harper, Elle, and Nora, I am so proud to be your dad.

I also dedicate this dissertation to my parents, Tim and Nancy. Thank you for teaching me the importance of leading a purposeful life and for your continued belief in me. To my sisters, Ashley and Amanda, who have inspired me throughout my life.

ACKNOWLEDGEMENTS

Many individuals have contributed to my success in completing this study. I am extremely grateful to Dr. Richard Lambert for serving as my dissertation chair and for taking me in as a graduate assistant in the Center for Educational Measurement and Evaluation in August 2019. The successful completion of this dissertation and many other research opportunities would not have been possible without your teaching and guidance. Your mentorship and leadership have made a tremendous impact on me.

I would like to express my deepest appreciation to Dr. Stella Kim for serving on this dissertation committee. I was fortunate to learn from you in several courses as a student in the ERME program and as a teaching assistant. I appreciate the challenge you brought to courses as a professor and for your advice in shaping this dissertation. I would also like to extend my sincere thanks to Dr. Carl Westine for serving on my dissertation committee and as my advisor. You have extended a great amount of encouragement and patience with me throughout my time in the program and the duration of this study. I sincerely thank Dr. Anne Cash for serving on this dissertation committee. Your thoughtful and constructive feedback greatly improved this dissertation.

I am also grateful for the support of Bryndle Bottoms, my fellow CEME graduate assistant throughout my entire time working in the center. I did not expect to gain a third sister as a doctoral student, but I am so lucky that I did. Finally, I am thankful to have learned from and with all of the CEME staff members, ERME professors, and ERME students during my time at the University of North Carolina at Charlotte.

TABLE OF CONTENTS

LIST OF TABLES xi

LIST OF FIGURES xiii

CHAPTER I: INTRODUCTION 1

Rater-Mediated Assessment Theory 3

Rater Cognition 4

Applications of Rater-Mediated Assessment in Education 5

Formative Assessment 5

Teacher Evaluation 6

Validity and Interrater Reliability 7

Interrater Agreement Coefficients 9

Conclusion 11

Limitations 13

Delimitations 14

CHAPTER II: LITERATURE REVIEW 15

Rater-Mediated Assessment 15

Rater-Mediated Assessment Theory 16

Rubrics 17

Rating Scales 18

Rater Cognition 18

Cut Scores and Standard-Setting 20

Rater Training 21

Rater Accuracy, Agreement, and Consistency 22

Interrater Reliability Coefficients 25

Bennett, Alpert, and Goldstein's *S* 25

Cohen's Kappa 26

Prevalence and Bias Indices 30

Krippendorff's Alpha 31

Gwet's AC1 33

Lambda Coefficient of Rater-Mediated Agreement 35

Summary of IRR Coefficients 37

Applications of Rater Agreement Statistics in Educational Research 38

Formative Assessments 38

Teacher Evaluation 40

Validity and Reliability of Teacher Evaluation Processes 43

Construct Irrelevant Variance in Teacher Evaluation 44

Chance-Corrected Agreement Coefficients: Teacher Evaluation Contexts 45

Conclusion 46

CHAPTER III: METHODOLOGY 49

Simulation Factors 49

Location of Agreement and Disagreement 50

2x2 Agreement Matrix 50

3x3 Agreement Matrix 53

4x4 Agreement Matrix 56

Marginal Distributions 59

Summary 59

Data Generation Conditions 59
2x2 Data Conditions 59
3x3 Data Conditions 62
4x4 Data Conditions 69
Summary 76
Chance-Corrected Agreement Coefficients 77
Calculating Lambda-2 78
Evaluation Criteria 79
Evaluating Research Question 1 79
Evaluating Research Question 2 80
Evaluating Research Question 3 81
Expected Outcomes 82
Summary 83
CHAPTER IV: RESULTS 85
Results: Research Question 1 86
Similarity of Coefficient Values 86
2x2 Coefficient Comparisons 87
3x3 Coefficient Comparisons 89
4x4 Coefficient Comparisons 92
Results: Research Question 2 95
Classification Consistency of Coefficient Values 95
2x2 Classification Consistency 96
3x3 Classification Consistency 99

4x4 Classification Consistency 101

Results: Research Question 3 103

Correlation of Coefficient Values 103

2x2 Coefficient Correlations 103

3x3 Coefficient Correlations 106

4x4 Coefficient Correlations 109

Summary 111

CHAPTER V: DISCUSSION AND IMPLICATIONS 114

Conclusions 115

Findings According to Research Question 1 115

Findings According to Research Question 2 116

Findings According to Research Question 3 117

Limitations 120

Implications for Policy and Practice 121

Directions for Further Research 126

REFERENCES 127

APPENDIX A: AGREEMENT MATRICES WITH PROBABILITY OF A CORRECT GUESS CALCULATIONS (p_e) 141

APPENDIX B: AGREEMENT MATRICES WITH COEFFICIENT CALCULATIONS 148

APPENDIX C: EXAMPLE R SCRIPTS 152

APPENDIX D: BENCHMARK AGREEMENT CLASSIFICATION TABLES (2X2 AGREEMENT MATRICES) 161

APPENDIX E: BENCHMARK AGREEMENT CLASSIFICATION TABLES (3X3 AGREEMENT MATRICES) 170

APPENDIX F: BENCHMARK AGREEMENT CLASSIFICATION TABLES (4X4 AGREEMENT MATRICES) 189

APPENDIX G: CORRELATION TABLES AND FIGURES (2X2 AGREEMENT MATRICES) 213

APPENDIX H: CORRELATION TABLES AND FIGURES (3X3 AGREEMENT MATRICES) 221

APPENDIX I: CORRELATION TABLES AND FIGURES (4X4 AGREEMENT MATRICES) 238

APPENDIX J: COEFFICIENT LINE GRAPHS 260

LIST OF TABLES

Table 1.1: Development of Interrater Agreement Coefficients 10

Table 2.1: 2x2 Agreement Matrix 28

Table 2.2: Benchmark Agreement Levels for Chance Corrected Agreement Indices 29

Table 2.3: Sample Distribution of 100 Ratings 29

Table 2.4: 2x2 Agreement Matrix 31

Table 2.5: Overview of Interrater Reliability Coefficients 38

Table 3.1: 2x2 Data Simulation Conditions 60

Table 3.2: 3x3 Data Simulation Conditions 64

Table 3.3: 4x4 Data Simulation Conditions 70

Table 3.4: Sample of 2x2 Data Generated for Research Question 1, Evaluation Criteria 80

Table 3.5: Sample Benchmark Count for Data Condition 1 (2x2) 80

Table 3.6: Sample Benchmark Percentages for Data Condition 1 (2x2) 81

Table 4.1: 2x2 Mean Coefficient Values 88

Table 4.2: Out of Range Values Between Coefficient Pairs (2x2) 89

Table 4.3: 3x3 Mean Coefficient Values 91

Table 4.4: Out of Range Values Between Coefficient Pairs (3x3) 92

Table 4.5: 4x4 Mean Coefficient Values 93

Table 4.6: Out of Range Values Between Coefficient Pairs (4x4) 94

Table 4.7: Count Across Benchmark Agreement Levels for Data Condition 1 (2x2) 95

Table 4.8: Percentage of Values within Benchmark Levels for Data Condition 1 (2x2) 96

Table 4.9: Count Across Benchmark Agreement Levels for Data Condition 14 (2x2) 98

Table 4.10: Percentage of Values within Benchmark Levels for Data Condition 14 (2x2) 98

Table 4.11: Count Across Benchmark Agreement Levels for Data Condition 23 (3x3) 99

Table 4.12: Percentage of Values withing Benchmark Levels for Data Condition 23 (3x3) 100

Table 4.13: Count Across Benchmark Agreement Levels for Data Condition 53 (4x4) 102

Table 4.14: Percentage of Values within Benchmark Levels for Data Condition 53 (4x4) 102

Table 4.15: Correlation Matrix for Data Condition 5 (3x3) 107

Table 4.16: Correlation Matrix for Data Condition 10 (4x4) 110

Table 4.17: Correlation Matrix for Data Condition 55 (4x4) 110

LIST OF FIGURES

Figure 3.1: 2x2 agreement matrix probability distribution. 52

Figure 3.2: 3x3 agreement matrix probability distribution. 55

Figure 3.3: 4x4 agreement matrix probability distribution. 58

Figure 3.4: 2x2 agreement distribution for data condition 1 from Table 3.1. 61

Figure 3.5: 2x2 agreement distribution for data condition 8 from Table 3.1. 62

Figure 3.6: 2x2 agreement distribution for data condition 15 from Table 3.1. 62

Figure 3.7: 3x3 agreement distribution for data condition 1 from Table 3.2. 65

Figure 3.8: 3x3 agreement distribution for data condition 10 from Table 3.2. 65

Figure 3.9: 3x3 agreement distribution for data condition 12 from Table 3.2. 66

Figure 3.10: 3x3 agreement distribution for data condition 19 from Table 3.2. 66

Figure 3.11: 3x3 agreement distribution for data condition 23 from Table 3.2. 67

Figure 3.12: 3x3 agreement distribution for data condition 26 from Table 3.2. 68

Figure 3.13: 3x3 agreement distribution for data condition 35 from Table 3.2. 68

Figure 3.14: 4x4 agreement distribution for data condition 1 from Table 3.3. 72

Figure 3.15: 4x4 agreement distribution for data condition 18 from Table 3.3. 73

Figure 3.16: 4x4 agreement distribution for data condition 51 from Table 3.3. 74

Figure 3.17: 4x4 agreement distribution for data condition 59 from Table 3.3. 75

Figure 3.18: 4x4 agreement distribution for data condition 65 from Table 3.3. 76

Figure 3.19: Sample correlation matrix for data condition 1 (2x2). 82

Figure 4.1: Replication of Figure 10 from Xie (2013). 86

Figure 4.2: Coefficient Values Under Data Condition 1 (2x2) 97

Figure 4.3: Coefficient Values Under Data Condition 14 (2x2) 98

Figure 4.4: Coefficient Values Under Data Condition 23 (3x3) 100

Figure 4.5: Coefficient Values Under Data Condition 53 (4x4) 102

Figure 4.6: Correlation results under data condition 3 (2x2) 104

Figure 4.7: Correlation results under data condition 8 (2x2) 105

Figure 4.8: Correlation results under data condition 13 (2x2) 106

Figure 4.9: Correlation results under data condition 26 (3x3) 108

Figure 4.10: Correlation results under data condition 29 (3x3) 109

Figure 4.11: Correlation results under data condition 56 (4x4) 111

Figure 5.1: Line graph displaying coefficient performance under data condition 5 (4x4) 119

Figure 5.2: Line graph displaying coefficient performance under data condition 28 (4x4) 119

Chapter I: Introduction

Rater-mediated assessment is defined as "any assessment or test that consists of constructed responses that require a rater, reader, judge, or examiner to interpret the performance and assign a rating based on their judgment" (Engelhard & Wind, 2018, p. 341). Engelhard (2002) introduced the term rater-mediated assessment to capture the concepts related to assessments where human raters were involved in interpreting responses and assigning these performances a rating through complex cognitive processes. In addition to assessments using constructed responses, rater-mediated assessments include various educational performance assessments, tests of language proficiency, and personnel evaluation (Engelhard & Wind, 2019). Assessment developers involve raters because they believe raters will provide relevant insight into the construct of interest as compared to automated scoring methods. However, involving human raters brings about many different imperfect characteristics involved with human ratings. Due to concerns over human idiosyncrasies, researchers suggest the use of several indicators to enhance the psychometric quality of rater-mediated assessments. More specifically, studies on rater-mediated assessment can alleviate issues one may have with the use of human raters by providing information about rater agreement, accuracy, consistency, and errors/biases in scoring and rating scale use (Engelhard & Wind, 2019).

Rater agreement is often measured through calculating the proportion of ratings that match the set of ratings of another rater or against a set of correct ratings. Another important concept is rater consistency; however, consistency alone does not ensure that ratings provided are valid or fair. Raters could be consistently strict or consistently lenient in their rating patterns. As a way to improve information provided from rater-mediated assessments, interrater

agreement and reliability coefficients were developed. Many interrater reliability (IRR) coefficients are measures more stringent than percent agreement values known as chance-corrected agreement coefficients. Kappa is a widely used chance-corrected agreement coefficient, meaning it considers the possibility of raters agreeing due to chance. It is common for a combination of different agreement indices to be reported as part of IRR studies. Graham et al. (2012) recommended overall percentage agreement values should be reported no matter the number of categories or rating levels used on a measure. In the instance a measure uses four or fewer categories, it was recommended to report Cohen's Kappa (Graham et al., 2012). Rating scales with fewer steps tend to be easier for raters to use. In addition, the number of categories used within a scale impacts the calculation of expected chance-agreement for most IRR coefficients. Reliability estimates may be inflated as ratings scales grow in the number of categories, especially when categories on the rating scale are rarely or never used for scoring decisions.

The Lambda Coefficient of Rater-Mediated Agreement (Lambda) is grounded in rater-mediated assessment theory and was developed as an alternative to Kappa (Cohen, 1960) and other chance-corrected agreement coefficients (Lambert et al., 2021). There are well-known issues with the Kappa coefficient of chance-corrected agreement (Cicchetti & Feinstein, 1990; Uebersax, 2002; Xie, 2013); most prominently are issues of presenting low levels of reliability when raters have high-agreement and low-prevalence of category usage. Lambda was developed as part of a study on IRR within preschool teacher evaluations. In the context Lambda was first used, a group of professional evaluators (n = 57) had high-agreement and low-prevalence categories, two common data conditions that result in issues with existing chance-corrected agreement coefficients. Using field data from Lambert et al.'s (2021) study, Lambda yielded

coefficients as compared to Kappa and other chance-corrected agreement coefficients that remained relatively stable across data conditions and did not over-correct for chance agreement (Lambert et al., 2021).

Lambda has two variations, a general form that is calculated similarly to how most chance-corrected agreement coefficients are calculated, such as Kappa (Lambert et al., 2021). The general form of Lambda is referred to as Lambda-1. Lambda-2 differs from Lambda-1 in the calculation of the proportion of expected chance agreement. Lambda-2 uses the population proportions for how frequently points on the rating scale are used and applies those proportions to the calculation to expected chance agreement. For example, one component of the calculation of expected chance agreement for a particular rating scale point is equal to one over the number of categories on the rating scale in the calculation of Lambda-1. In the calculation of Lambda-2 this value would vary according to the population proportion of ratings at each step of the rating scale. This method assumes the population proportions are known and this influences a rater's decision when arriving at a rating selection. There are limited situations in which the calculation of Lambda-2 is possible. The purpose of this study is to test the functioning of Lambda through simulating potential data conditions using two-, three-, and four-point ordinal rating scales posing similar situations as real-world data conditions.

Rater-Mediated Assessment Theory

The current study implements rater-mediated assessment theory (Engelhard, 2002; 2013), which is an adaptation of the lens model (Brunswik, 1952). This theory proposes a conceptual framework for rater judgment and decision-making. This framework aims for a mirroring of the latent variable of interest within observation-based ratings, which requires a close linkage between items, rater, and the rating scale. Rater-mediated assessment consists of raters making

placements on ratings scales. The placements made by raters are a result of a series of decisions made by each rater and can lead to inconsistencies within and between raters. It is possible that raters will tend to use a certain rating category more or less frequently, and even inconsistently apply the scoring guide for a variety of reasons. Therefore, to completely understand how to support raters and enhance the reliability, validity, and fairness of their ratings, it is necessary to examine rater agreement, leniency, and strictness. One way to understand how to support raters is through focusing on the mental processes' raters go through as they make scoring judgments (Myford, 2012).

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). Acknowledging and having 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 for Educational and Psychological Testing* (*Standards*) (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 with 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 according to the rubric or rating scale. Understanding and researching rater cognition are beneficial practices for preventing threats to validity related to rater judgments. An awareness of rater cognition is required for valid, reliable, and fair assessment practices to occur in rater-mediated assessment.

Applications of Rater-Mediated Assessment in Education

Rater-mediated assessment is used in education when rater judgment is able to provide more useful and higher quality information about an examinee's performance on a latent construct as compared to an assessment without human judgement (Engelhard & Wind, 2019). The application of rater-mediated assessment occurs broadly in a variety of educational settings and across content areas. These are most frequently associated with essays and assessments with constructed response items. Two other educational performance assessment areas that are applications of rater-mediated assessment occur in formative assessment processes and in teacher evaluation systems.

Formative Assessment

Assessment can suggest modifications to teaching, guide learning through immediate feedback, and help students self-assess themselves and each other to inform next learning steps (National Council on Measurement in Education [NCME], 2018). Also, assessments provide opportunities to assess young children's strengths, progress, and needs with developmentally appropriate measures. Assessment can be used to make decisions about teaching and learning. Early childhood assessments should be aligned with early learning standards, program goals, and with specific emphases in the curriculum (National Association for the Education of Young Children & National Association of Early Childhood Specialists in State Departments of Education [NAEYC & NAECS-SDE], 2003).

Formative assessment is a process requiring teachers to recognize and interpret demonstrated skills, followed by providing continuous scaffolding and feedback to students (Heritage, 2013). Shepard (2000) described formative assessments as dynamic and on-going, and defined dynamic assessment as a process of "finding out what a student is able to do independently as well as what can be done with adult guidance" (p. 10). Formative assessment can suggest in the moment modifications to teaching, guide student learning through immediate feedback, and aid students in assessing themselves and one another to inform next learning steps (NCME, 2018). The assessment results should produce information that can be utilized for intervention, provide evidence that a learning gap exists and suggest possible means that will successfully close the gap. High-quality formative assessment measures offer valid information that can facilitate the teaching and learning process. Performance assessments, often referred to as authentic assessments, require students to directly demonstrate their ability to perform a skill or task. The goal of such assessments is for students to have opportunities to perform a skill in a natural context (Darling-Hammond, 2017). Performance assessments can be used as formative or summative and can improve teaching and learning.

Teacher Evaluation

When a human rater measures the performance of a student or teacher on a given construct many different rater characteristics and features of the observations can introduce construct-irrelevant variance. Construct irrelevant variance is a term used to describe variance in scores that occurred because of "extraneous factors that distort the meaning of the scores and thereby decrease the validity of the proposed interpretation" (American Educational Research Association [AERA] et al., 2014, p. 217). In these common situations the observed scores are not independent of the timing of the observation. Examples of these structural features of

observations include, and are not limited to, the makeup of students in the classroom, the topic or unit of the lesson, day of the week, and the time of day the observation occurs. Graham et al. (2012) recommended caution when comparing results from observations made at different times. Teacher evaluation systems provide administrators and teachers an idea of teacher effectiveness to improve student learning and educational attainment. Across the United States there are various requirements related to the quantity, length, and type of observations required of each sample of teachers. Due to the high-stakes nature of teacher evaluation, researchers and educators have concerns with the quality, accuracy, and overall fairness of ratings provided by evaluators.

Validity and Interrater Reliability

Validity addresses the extent to which the meaningfulness, appropriateness, and usefulness of specific interpretations of test scores are supported by evidence. The process of validation involves building an accumulation of evidence to support the basis for proposed score interpretation. Interpretations or applications of scores "in a high-stakes environment are vulnerable to many validity threats, such as inadequate construct definition, construct underrepresentation, illogical reasoning…, negative consequences of test score use, and low reliability of test scores" (Haladyna & Downing, 2004, p. 25). According to Lane (2019), the "use of rater-mediated assessments requires the evaluation of the accuracy and consistency of the inferences made by those who interpret examinee performances to ensure the validity of their judgements regarding examinee performances and the use of the examinee scores" (p. 653). Validity evidence supporting rater-mediated assessment should include the recognition and validation of rater cognitive response processes to ensure raters are using the same criteria when making ratings concerning the measured construct (Bejar, 2012).

Reliability is a pre-condition for validity, making it important to address IRR of rater-mediated assessments. In order for validity evidence to be established, there must first be evidence that raters understand the instrument and implement the ratings process with consistency and accuracy. Once evidence for these rater behaviors is established, then the validation process can turn to gathering evidence concerning whether the instrument is in fact measuring the intended construct. According to Shrout & Fleiss (1979), rater reliability is the degree of agreement between raters who are judging a defined construct according to specified rating criteria. Reliability analysis in rater-mediated assessment involves evaluating the consistency of ratings provided by raters, the accuracy of rater judgements, and consistency in the order of a group of raters according to severity (Wind, 2019). The results and methods used to evaluate the reliability of an instrument should be reported clearly by test developers and users (AERA et al., 2014).

Teacher evaluation studies reporting IRR statistics frequently stop at reporting percentage agreement between raters (Casabianca et al., 2015; Hill et al., 2012; Sartain et al., 2010). The *Standards* do not provide suggestions for a specific agreement level or reliability measure, but recommend appropriate measures are reported and calculated while an assessment is in use (AERA et al., 2014; Graham et al., 2012). As Hill et al. (2012) suggested, percentage agreement figures could be overstated through simply having fewer scale points on the rating scale. Concerns about the reliability and validity of ratings are warranted and can start to be addressed through establishing IRR protocols and using an instrument that provides valid and reliable ratings.

In the *Standards* (AERA et al., 2014), fairness is described as “a fundamental validity issue and requires attention throughout all stages of test development and use” (p. 49). Fairness

is essential to the rater scoring process and addressing potential areas of construct-irrelevant variance. According to Wind (2019), rater training should clearly demonstrate how performances differ according to the rating scale in use and "highlight the key construct relevant components of ordered levels of proficiency" (p. 498). This is a key step in ensuring that raters are applying fair judgements of observed performances and in gathering evidence of fairness.

Interrater Agreement Coefficients

There are several methods for calculating IRR, these measures reflect the level of agreement between raters. The most basic measure of IRR is percent agreement between raters and there are more complex methods that take agreement by chance into account. 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. *S* is based 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).

Since the development of Kappa, many other variations of Kappa and additional chance-corrected agreement coefficients were developed to address certain paradoxes of Kappa and designed for more specific contexts of assessing IRR. The main paradoxes of the Kappa

coefficient involve situations where actual agreement is high, and the number of rating categories used is low. These situations result in low Kappa values. Krippendorff's alpha was developed to address the issues prevalent in Kappa and has different calculations according to the level of measurement (Krippendorff, 2011). 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 for this value to range from 0.0 to 1.0 depending on the marginal distribution of ratings (Gwet, 2001).

Table 1.1

Development of Interrater Agreement Coefficients

Coefficient	Description
Bennett, Alpert, & Goldstein's *S* (1954)	Based on proportion of agreement according to the number of categories; slight adjustment from exact agreement; has a constant value for all ratings with same number of categories and level of agreement
Scott's π (1955)	Original chance-corrected coefficient; designed for nominal data in communication studies
Kappa (1960)	Improved the calculation of chance-agreement; developed to account for the possibility that raters guess on some ratings due to uncertainty
Krippendorff's Alpha (1970)	Addressed issues with Kappa, adjusts calculations according to level of measurement; designed for content analysis
Gwet's AC1 (2001)	Resistant to Kappa's known issues, limited pe to 0.5 (Kappa can be as high as 1.0)
Lambda (2021)	Most similar to Gwet's; however, designed for ordinal scales, based on Rater-Mediated Assessment Theory; caps p_e at 0.5

Researchers across many fields have introduced several measures of interrater agreement that provide information beyond simple percentage agreement (Zwick, 1988). The difference between solving for most of these coefficients is found in how each coefficient defines how to solve for agreement due to chance. Early IRR coefficients were each seen as an improvement upon what was considered the standard at the time. *S* was an improvement upon simple percentage agreement, however was based on uniform marginal distributions of ratings. Scott's π improved upon *S* since it corrected for the inclusion of unused categories that inflated the value of *S* (Scott, 1955). Kappa was an enhanced version of Scott's π as it incorporated the actual marginal distribution of ratings as opposed to ignoring marginals as *S* does or assuming marginal homogeneity as is the case in the calculation of Scott's π (Cohen, 1960; Fleiss, 1975; Uebersax, 2006). Chance-agreement is computed by adding the products of the proportional ratings within a category along a row of an agreement matrix and the proportional ratings within a category along the corresponding column of an agreement matrix. An example will be provided in Chapter II of this dissertation. As the use of IRR coefficients expanded across ratings of different scales of measurement and due to the discover of flaws with the Kappa coefficient, several variations of Kappa and new IRR coefficients have been developed. Table 1.1 offers a brief description of several IRR measures.

Conclusion

Therefore, this study is needed and will contribute to existing literature by closely examining a new alternative to Kappa. This study will take a rigorous examination of a new IRR measure. Lambda-1 and Lambda-2 were designed to capture the reliability of scores from raters using rater-mediated assessments more precisely than existing chance-corrected agreement coefficients. Individual levels of reliability from chance-corrected agreement coefficients provide

just one piece of evidence. However, an IRR coefficient designed for a specific use and purpose offers benefits over other coefficients that were solely developed as a means of correcting for the Kappa coefficient's known paradoxes. This study will closely examine the performance of Lambda-1 and Lambda-2 under conditions common in rater-mediated assessments and known to be problematic for other IRR coefficients.

This dissertation contributes to the literature by providing evidence of the validity, reliability, and fairness of ratings produced by raters using rater-mediated assessments, specifically teacher evaluations and formative assessment processes. Evaluation of the validity and reliability of scores produced by raters is essential for maintaining the usefulness of teacher evaluation systems and widely used rater-mediated assessments. This study targets areas in education requiring careful attention to ensure teachers are provided with valid and reliable information to improve their practice and student learning. Teachers could be overrated or underrated by evaluators of their performance. This is problematic as it could result in the misallocation of scarce resources to teachers and schools. Teacher evaluations are often high stakes, a decision around promoting or giving tenure to a teacher is based on evaluation scores. Meaning, there are major consequences for schools and individual teachers if scores from evaluations are unreliable or invalid. Teachers could be unnecessarily retained or pushed out of the profession due to decisions using teacher evaluation measures.

Lambda was initially developed as part of an applied study on the reliability of teacher evaluation scores from a group of raters of early childhood teachers (Lambert et al., 2021). As part of this initial application of Lambda-1 and Lambda-2, four variations of data conditions were evaluated to determine how the new coefficients performed in comparison to Kappa. The four simulated conditions applied conditions known to be problematic for Kappa, high

agreement and low category usage. The current study differs from Lambert et al.'s (2021) study in that it varies substantially more conditions and compares the performance of Lambda-1 and Lambda-2 to several other chance-correct agreement coefficients. The conditions for the current study are further explained in Chapter 3, however they differ from the initial testing of Lambda in that more conditions are adjusted across two-, three-, and four-category rating scales. Lambert et al. (2021) compared the performance of Lambda-1 and Lambda-2 to Kappa, whereas the current study compares those coefficients in addition to the following: *S*, Krippendorff's Alpha, and Gwet's AC1.

In instances where results from rater-mediated assessments are used to inform high-stakes decisions, such as in many teacher evaluation systems, it is critical to report the IRR of raters. In isolation, IRR measures and percentage agreement values provide a limited amount of information. The use of IRR coefficients along with other evidence of reliability and validity of score interpretation processes is necessary to provide a detailed picture of the reliability and agreement between raters. There is a need for the development and validation of a new coefficient due to known limitations and manifestations of problems in teacher evaluation data. Lambda was designed according to rater-mediated assessment theory and specifically for use with ordinal data.

Limitations

The results of this simulation study involve one data-generating mechanism to ensure coverage of scenarios representative of the aims of the study (Morris et al., 2019). In this context, simulation refers to a demonstration of the behavior of a statistic across data conditions and analyzing the performance of a statistic through the creation of a response surface displaying values of each coefficient under each condition. Other data-generating mechanisms, such as the

use of a parametric model to generate data could be applied in future studies on Lambda that consider specific sample sizes or randomization of data conditions. This is unnecessary in the present study because the focus is on the performance of Lambda across a set of conditions commonly found in rater-mediated assessment applications. Factors that varied in the data-generating process were set to closely match what is custom in rater-mediated assessment and commonly found in applications of Kappa, such as teacher evaluation systems. Validation of estimates of uncertainty of the Lambda coefficient are not included in the current study. Another limitation of these coefficients is that they are not useful for detecting rater strictness or leniency. Considering the known issues with Kappa and other chance-corrected agreement coefficients, it is good practice to compute multiple alternative measures in an IRR study.

Delimitations

The focus of this study is on the performance of an alternative to Kappa and other chance-corrected agreement coefficients. The data-generating mechanism focuses on adjusting Bias and Prevalence Indices (Xie, 2013) according to predefined data conditions that are similar to what is found in real-world examples of rater-mediated assessment. Instead of a true Monte Carlo simulation utilizing randomizing and Markov Chain methods, the simulations in this study are a series of preset conditions defined according to the amount and location of agreement and disagreement between a pair of ratings (Morris et al., 2019). The simulated data conditions restrict the rating scales to two-, three-, and four-point matrices.

Chapter II: Literature Review

Current methods of calculating measures of IRR can provide insufficient or inaccurate information about the scores produced by raters. This study investigates the performance of a new IRR coefficient, Lambda, designed for use in rater-mediated assessments that demonstrated robustness to data conditions problematic for existing IRR measures (Lambert et al., 2021). It is important to ensure ratings provided in rater-mediated situations are valid, reliable, and fair. Having a coefficient that provides accurate information is vital to ensuring educators are informed about strategies that extend their own professional growth. Therefore, evaluating Lambda's performance to data conditions that are known to cause issues for other coefficients and that are also similar to real-world IRR agreement values provides supporting evidence that Lambda is a reliable option to measure IRR.

Rater-Mediated Assessment

The term rater-mediated assessment was introduced by Engelhard (2002) to describe assessments including raters "judging the quality of an examinee's responses (e.g. essays and portfolios) become the stimuli that raters must interpret and evaluate to produce ratings" (p. 261). Rater-mediated assessments are often presented with complex situations, such as in teacher evaluation processes. Raters of teacher performance must carefully and skillfully provide scores for an observation, no matter if the use of the score is summative to make a high-stakes decision or formative to provide feedback to the teacher. For teacher evaluation systems based on observational data, the item responses consist of placements on rating scales made by evaluators. These placements result from a series of decisions made by each evaluator and inconsistencies between raters can be common. Engelhard (2002) noted a major concern of rater-mediated

assessments was the potential for raters to have bias. Wilson (2004) listed several reasons inconsistencies may occur: raters may never apply the scoring guide in a correct way due to training differences, there may be differences in rater severity, raters often have natural tendencies to use rating categories more or less frequently, “halo effects”, rater drift, and raters demonstrating inconsistencies themselves for a variety of reasons.

Rater-Mediated Assessment Theory

Rater-mediated assessment was developed according to the judgmental and cognitive process known as the lens model (Brunswik, 1952). Hogarth (1987) further adapted the lens model in cognitive psychology to better understand human judgment, limitations of human judgments, and how to improve decision-making. The lens model is a conceptualization of the cognitive process raters go through when mediating and interpreting a person’s ability using available evidence (referred to as “cues”) to make a judgment of the performance (Engelhard, 2013; Engelhard & Wind, 2018). Cues are things that assessment developers must intentionally develop and train raters to pay close attention to when assigning ratings (Engelhard, 2013). A few examples of cues include rubrics, scoring domains, and rating scales (Engelhard et al., 2018).

Under rater-mediated assessment theory, it is assumed raters are required to have a high level of expertise through training and prior experiences. Also, raters use complex internal response processes to make ratings. This internal response process may be influenced by the rater’s inclination to provide strict or lenient ratings. Strict ratings occur when a rater gives a rating lower than the actual performance of a subject according to a set of correct answers (Lunenberg, 2012). While a lenient rating occurs when a rater provides a higher rating than the test takers actual performance. Additionally, many trained raters begin with an initial rating

starting point in mind when entering a rating situation (Lambert et al., 2021). The completion of this process is aided by the use of defined performance levels on a rubric or rating scale.

Rubrics

One type of assessment tool that is commonly used in formative and classroom assessments is a rubric. According to Brookhart (2018), a rubric consists of criteria matching the purpose of the assessment and performance level descriptions. The criteria are expressed on the rubric as what to look for in the work or performance. While the performance level descriptions describe the criteria at different levels of quality (i.e. low to high) (Andrade, 2000; Brookhart & Chen, 2015). There are several ways to categorize rubrics, one that is pertinent to rater-mediated assessments is the distinction between analytic and holistic rubrics (Johnson et al., 2009). The discrepancy here is in how information is evaluated by the rater and feedback is disseminated to the ratee.

An analytic rubric is often paired with a formative assessment where the intention of the assessment is to provide feedback that informs and improves future performance (Brookhart, 2018). Analytic rubrics assess individual dimensions of an overall assessment. Whereas when using a holistic rubric, the rater makes an overall judgment about the performance or work as a whole (Jonsson & Svingby, 2007; Moskal, 2000). While a key disadvantage is that holistic rubrics often lack specific, individualized feedback and some advantages include that they can be less cognitively demanding and time consuming for raters (Brookhart, 2018). Some researchers use the term rubric to capture the term for any type of evaluative tool used by a rater in an assessment situation. While others make a distinction between rubrics, checklists (yes/no), and rating scales (Brookhart, 2018; Brookhart & Chen, 2015).

Rating Scales

Rating scales are used by raters to make scoring decisions about an observed performance. With rating scales a clear description of the level by level performance is not always described as well as it is within a rubric. Rating scales are often a numerical scale using a range of levels (i.e., 1-4), evaluative scale describing the performance in relation to the task (i.e., below expectations to beyond expectations), or a frequency scale describing how often a behavior was demonstrated (i.e., never to always). No matter the type of rating scale in use, it is important the scale have a number of rating categories that allow for enough differentiation between levels of performance and not too few that categories are indistinguishable (Johnson et al., 2009; Lane & Stone, 2006). Rating scales should be introduced during the design phase of assessment development (AERA et al., 2014). In addition, the *Standards* (2014) suggest that a test's scoring specifications should describe the qualifications, training, and monitoring process when scoring is completed by human raters.

Rater Cognition

While there has been a great deal of effort to understand the cognitive processes raters of constructed response items undertake when making ratings decisions, there is not as much known about the process raters use to assign ratings according to observational data and performance assessments (Qi et al., 2018). The mental processes raters go through when assigning ratings to a performance is known as rater cognition. According to Kimball & Milanowski (2009), rater cognition is "important to explore in studies of evaluation decision making because of (raters) logical connection to cognitive processes and enacted behaviors" (p. 40).

The addition of contextual factors to certain rater-mediated assessment situations makes understanding the cognitive process raters go through even more complex. In research applications, the ratee and rater of an observation are often not acquaintances. However, in a classroom setting where a teacher is rating a student's performance, or an administrator is evaluating a teacher's performance this is not the case. Dynamics of the existing relationship add a new challenge to the rating situation, and this could be reflected in the ratings assigned (Ho & Kane, 2013; Qi et al., 2018). The addition of added factors to a rating situation makes it important for researchers to consider how raters arrive at ratings and the implications raters' cognitive processes has on the reliability and validity of assigned scores (Bell et al., 2018; Qi et al., 2018).

According to Suto (2012), cognitive interviews, or think-aloud exercises, are commonly used to study rater behavior. This process allows raters to reveal their thoughts while completing an assigned task, such as arriving at a rating decision (Lewis et al., 2020). In a study on the validity of teacher evaluation ratings provided by school administrators, researchers discovered that most evaluators in the study did not consciously think about the methods in which the employ when making ratings decisions when asked in retrospective cognitive interviews (Kimball & Milanowski, 2009). Another important finding was uncovered in that teacher evaluators often enter an observation scenario knowing they are deciding between two distinct categories on the rating scale and not entering in the rating situation absentmindedly or randomly. A principal interviewed in Kimball & Milanowski's (2009) study directly stated "my teachers are all very experienced, so I would never walk in thinking they are a Level 1. To me it is a matter of, OK, what types of things will I see now that will have me decide

between Level 2 and Level 3" (p. 59). This statement demonstrates an important component of rater-mediated assessment theory, raters go through a complex process when arriving at a final placement on a rating scale through synthesizing information from previous ratings and evidence during new observations.

It is consistent throughout the literature on numerous performance assessments that raters across various backgrounds do not begin using observational protocols absent of prior knowledge, experiences, and expectations of performance (Bell et al., 2018; Kraft & Gilmour, 2016; Nijveldt et al., 2009; Qi et al., 2018). In a two-year study spanning a training year and the initial implementation year of a new teacher evaluation process, researchers found administrators relied on evidence during the observation in the training year when arriving at a rating decision according to think-aloud exercises (Qi et al., 2018). Whereas in the following implementation year, after administrators went through a year of rating practice and training, it was revealed through similar cognitive, think-aloud exercises the administrators were relying on a specific rating score they already had in mind when entering the rating situation. Variations in the cognitive process raters go through make it necessary to examine evidence of reliability and scores produced from ratings assigned in rater-mediated assessments.

Cut Scores and Standard-Setting

While cut scores and standard-setting processes apply to most any type of assessment, they are key components in rater-mediated assessments. Cut scores are locations on a scale that are used to establish minimal competency on a given construct (AERA et al., 2014). Scores at or beyond the cut score point are interpreted and applied differently from scores that fall below that point. The process of systematically establishing cut scores is known as standard setting

(McClarty et al., 2013). Often this involves the formal organization of a group of qualified experts in the group of skills and content covered by the assessment (Peterson et al., 2011).

According to Standard *5.21*:

> "If a judgmental standard-setting process is followed, the method employed should be described clearly, and the precise nature and reliability of the judgments called for should be presented whether those are judgments of persons, of item or test performances, or of other criterion performances predicted by tests scores" (AERA et al., 2014, p. 108).

Peabody and Wind (2019) suggested cut scores are arbitrary in some instances and panelists involved in standard-setting procedures must have the necessary expertise and qualifications required to make qualified judgments. The number of panelists involved in developing cut scores should be large enough to ensure expert recommendations are reliable and would not vary if the standard setting process were repeated by a similar panel of experts (AERA et al., 2014). While raters do not need to understand specific details of the standard setting process, the process has implications for how raters are trained. Panelists make suggestions during the standard setting process partly under the assumption that raters will apply scoring rubrics or rating scales in a manner that is consistent with their own interpretation. It is critical for raters to be well trained and monitored to ensure they are in fact making accurate and consistent ratings decisions.

Rater Training

Human raters are used in rater-mediated assessments because it is believed raters can add value to the scoring and feedback process through their experiences and high level of expertise (Lane, 2019). After raters are selected according to their qualifications, they undergo a training

process to ensure the ratings they provide are valid and reliable. To address issues of rater inconsistency it is important to have a process for training raters and a monitoring system that tracks the consistency of raters over time (Johnson et al., 2008; Wilson, 2004). Wilson (2004) recommended five components to include in rater-mediated assessment training programs. Raters should have:

1. An understanding of the assessment or construct.
2. Opportunities to examine a large, representative sample of responses from the construct of interest.
3. Opportunities to have cognitive discussions with other raters on overlapping work.
4. Feedback provided to raters centering on how well they rate responses.
5. A system of rater calibration steps that result in raters passing training or being identified as having a need for further support.

Additionally, Wilson (2004) recommended a pre-developed monitoring system that could involve co-observations or re-ratings done by experts over a sample of a caseload. Using reference ratings of some sort allows for raters to see how consistent a sample of their ratings have been over time. The training and monitoring process of raters is critical to obtaining accurate and consistent scores.

Rater Accuracy, Agreement, and Consistency

Rater training, developing cut scores, and standard setting processes feed into setting up raters with the necessary cues to make valid and reliable ratings. Research on rater agreement has utilized various methods for examining the consistency and accuracy of ratings in performance assessment. From a rater-mediated perspective, consistency refers to the extent to which an

examinee's performance would be categorized into the same rating category over repeated instances using the same measure. Whereas accuracy refers to the degree actual ratings or classifications made by raters agrees with a correct rating as determined by an expert panel. The focus of the current study involves assessing rater accuracy, comparing ratings made by a single rater in comparison to a correct answer.

One way to ascertain the accuracy of a rater is to produce an index of the proportion of ratings of exact agreement. Exact agreement refers to the degree of agreement between a rater and a correct answer or can refer to agreement between raters evaluating the same assessment. There are methods that can be developed related to adjacent agreement that can be adjusted based on the rating scale used for the construct of interest. Most commonly, adjacent agreement refers to the percentage of ratings within one level of each other on a performance assessment rubric (Koslow, 2018). Using adjacent agreement calculations can produce overly positive results, especially in the case where there are a small number of rating categories available. Adjacent agreement values are useful to report in low-stakes situations, such as when differences in scores across certain ratings points are treated as equivalent by test users. In many teacher evaluation systems states report the percentage of teachers that received "satisfactory" or "proficient" ratings on summative evaluations (Kraft & Gilmour, 2017). In these instances where the need for nuanced analysis of ratings is not used, adjacent agreement could be used. It is common to produce proportion of strict and lenient ratings along with the exact or adjacent agreement values.

One step beyond reporting exact or adjacent percentage agreement measures involves using IRR coefficients that adjust scores based on chance-agreement. When there are five or fewer rating levels on a scale it is better to use a more stringent measure than exact or adjacent

agreement (Graham et al., 2012). There are several chance-corrected agreement coefficients in use, and each has its own advantages and limitations. There is not a single measure that is best in all situations, making it necessary to calculate and report multiple measures of reliability (AERA et al., 2014; Graham et al., 2012). According to the *Standards* (AERA et al., 2014), the three main areas where evidence is necessary to support the use and interpretation of an assessment are validity, reliability, and fairness.

The *Standards* are not prescriptive for a single classification of an assessment, however there are standards that have direct implications for the reliability of judgements made by raters of rater-mediated assessments. Standards for reliability are grouped into eight clusters and it is possible to connect how rater accuracy, agreement, and consistency relates to each cluster. Both standards within the Reliability/Precision Cluster 3 – "Reliability/Generalizability Coefficients" have a more direct alignment to covering issues related to rater accuracy, agreement, and consistency. Standard 2.6 states: "A reliability or generalizability coefficient (or standard error) that addresses one kind of variability should not be interpreted as interchangeable with indices that address other kinds of variability, unless their definitions of measurement error can be considered equivalent" (AERA et al., 2014, p. 44). Standard 2.7 states:

> "When subjective judgement enters into test scoring, evidence should be provided on both interrater consistency in scoring and within-examinee consistency over repeated measurements. A clear distinction should be made among reliability data based on (a) independent panels of raters scoring the same performances or products, (b) a single panel scoring successive performances or new products, and (c) independent panels scoring successive performances or new products" (AERA et al., 2014, p. 44).

While these principles apply to "testing" as broadly defined, their application also includes a variety of assessment situations when rater or observer judgment is involved, not just for essay exams. These standards emphasize the importance of using a variety of indicators of reliability of an assessment and each indicator requires an appropriate interpretation.

Interrater Reliability Coefficients

Bennett, Alpert, and Goldstein's *S*

S was an early attempt at accounting for the percentage of agreement expected by chance as opposed to relying on simple agreement (Bennett et al., 1954). In the initial application of *S*, Bennett et al. (1954) examined the consistency of ratings from a group of 16 undergraduate students completing an interview and a 30-item limited-response poll utilizing four categories for each of the items. Results found the measures of consistency (0.46 to 1.0) provided by *S* were greater than expected chance of agreement, which is equal to one over the number of categories in this example (0.25) (Bennett et al., 1954).

S is constant across all scenarios with the same agreement level and number of rating categories, distinguishing it from chance-corrected agreement coefficients.

The formula for *S* is:

$$S = \frac{(q \cdot p_a) - 1}{q - 1} \tag{2.1}$$

Where:

q = number of rating scale points or categories

p_a = proportion of exact agreement

Under all scenarios, *S* assumes uniform marginal distribution according to the number of categories on the measure in use, so it does not use actual proportions to estimate the expected proportions of ratings to agree by chance. Since *S* makes this assumption, it tends underestimate IRR values and can be artificially increased by including unused rating categories (Warrens, 2012; 2014; Xie, 2013; Zwick, 1988).

Cohen's Kappa

According to Xie (2013), Kappa is the most widely used summary measure of IRR with ten times more citations than the next most referenced agreement index, Scott's π. Kappa was developed to measure the agreement between two raters classifying the same units into the same number of mutually exclusive categories (Cohen, 1960). Kappa came about just after Scott's π, which was developed for application in content analysis in survey research in which it was assumed the proportional use of category ratings was equal across raters (Scott, 1955). The original development of Kappa was centered around the context of clinical psychology in situations in which there was not a "correct" judgment, and the two raters were qualified to make judgments using a scale of interest (Cohen, 1960). For simplicity, examples throughout this study refer to hypothetical situations where correct judgments, or ratings, are available. It is common practice in IRR certification processes and training exercises that raters are evaluated against a correct set of ratings when calculating agreement coefficients. A Kappa value of 1.0 indicates exact agreement between raters, while a Kappa of 0.0 indicates the agreement between raters is the same as agreement due to chance. It is possible for Kappa to have a negative value in the case the agreement between two raters is less than the chance agreement. Cohen (1960) claimed Kappa to be "directly interpretable as the proportion of joint judgements in which there is agreement, after chance agreement is excluded" (p. 46).

An agreement table is used to analyze the exact (p_a) and expected chance-agreement (p_e) proportions. For a two-point scale, a 2x2 agreement matrix (Table 2.1) would be used to calculate the marginal distribution that is required to calculate p_e. The chance-agreement proportion (p_e) in the formula for the Kappa coefficient is calculated as follows in Equation 2.2 using:

$$p_e = \sum_{x=1}^{q}(\mathrm{P} \cdot \mathrm{x})(\mathrm{Px} \cdot) \tag{2.2}$$

Where:

x = the specific rating scale point or category

P · x = proportion of observations in column x

Px · = proportion of observations in row x

Using Table 2.1 as a point of reference, cell *a* represents the number of ratings where the observer agreed with the correct answer using the first point on the rating scale. Cell *b* represents the number of ratings where the observer decided to place the performance at the first point but the correct rating was the second point. This is an instance when the rating was a "strict" rating, or the observer underscored the actual performance. Cell *c* represents the number of ratings where the observer placed the rating at the second point but the correct rating was the first point. This is an instance when the rating was a "lenient" rating, or the observer overestimated the actual performance. Cell *d* represents the number of ratings where the observer agrees with the correct answer using the second point on the rating scale. This agreement matrix could also be used for two different observers in the case a correct decision is not known.

Table 2.1

2x2 Agreement Matrix

		Correct Ratings		
		1	2	
Observer	1	*a*	*b*	P1·
	2	*c*	*d*	P2·
		P·1	P·2	N

The total number of observations (N) can be calculated in a two-point agreement matrix as the sum of cell *a*, cell *b*, cell *c*, and cell *d*. In the two-point agreement matrix, P·1 would be calculated as the sum of counts in cell *a* and cell *c* divided by the total number of observations (N). Next, P1· is calculated as the sum of counts in cell *a* and cell *b* divided by the total number of observations (N). P·2 would be calculated as the sum of counts in cell *b* and cell *d* divided by the total number of observations (N). Finally, P2· is calculated as the sum of counts in cell *c* and cell *d* divided by the total number of observations (N).

In the Table 2.1, p_e would be calculated by taking the sum of the marginal distribution of the rating distribution for the two columns. In this example, this means combining the probability the rater and correct answer used the first rating point with the sum of the probability the rater and correct answer used the second rating point. Which is calculated as follows for a two-point scale in Equation 2.3:

$$p_e = (\mathrm{P} \cdot 1)(\mathrm{P1} \cdot) + (\mathrm{P} \cdot 2)(\mathrm{P2} \cdot) \tag{2.3}$$

Kappa is calculated as seen in Equation 2.4:

$$\kappa = \frac{p_a - p_e}{1 - p_e} \tag{2.4}$$

This statistic estimates chance agreement through assuming ratings from different raters are completely random.

There is not a unified explanation of how to interpret agreement with any individual IRR coefficient. However, Landis and Koch (1977), Fleiss (1981), and Altman (1991) provided suggestions for how to interpret agreement with the Kappa coefficient (see Table 2.2).

Table 2.2

Benchmark Agreement Levels for Chance Corrected Agreement Indices

	< 0.00	0.00 – 0.20	0.21 – 0.40	0.41 – 0.60	0.61 – 0.80	0.81 – 1.00
Landis & Koch (1977)	No agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
Fleiss (1981)	Poor	Poor	Poor	Fair	Good	Excellent
Altman (1991)	Poor	Poor	Fair	Moderate	Good	Very Good

Kappa has demonstrated two well-known paradoxes. The first issue results in low Kappa values when exact agreement between raters is high. The second issue is the reliance Kappa places on the marginal distribution of ratings. In instances when there is low-prevalence category use or an imbalance in category use between two raters or a single rater and a correct answer, Kappa can be abnormally far from the actual agreement levels. For example, presume an observer made 100 ratings as shown in Table 2.3:

Table 2.3

Sample Distribution of 100 Ratings

		Correct Ratings		
		1	2	Total
Observer	1	*65*	*28*	93
	2	*4*	*3*	7
	Total	69	31	100

The sample rater in Table 2.3 had an overall exact agreement (p_a) of 68% (cell *a* and cell *d*), with 28% ratings in the strict cell (cell *b*), and 4% of ratings in the lenient cell (cell *c*). This example

illustrates a situation where agreement is high (68%) and there is an imbalance in the category use between the observer and the correct answer (see row totals as compared to column totals). Using the calculation of p_e for the Kappa coefficient, the Kappa value for this example would be solved as follows:

$$p_e = (.93)\ (.69) + (.07)\ (.31)$$

$$p_e = .64 + .02$$

$$p_e = .66$$

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

$$\kappa = \frac{.68 - .66}{1 - .66}$$

$$\kappa = \frac{.02}{.34}$$

$$\kappa = .06$$

This provides an example of an instance of acceptable exact agreement and an imbalance in category use prevalence, the resulting Kappa value is low (.06) even though actual agreement was much higher (.68).

Prevalence and Bias Indices

According to Byrt et al. (1993) these "difficulties occur because Kappa not only measures agreement but is also affected in complex ways by the presence of bias between observers and by the distribution of data across the categories that are used ('prevalence')" (p. 423). There is bias between raters when the marginal distribution is not equal. For example, in a 2 x 2 rating table (see Table 2.4) the "Bias Index" (BI) would be the difference in the proportions of "Yes" ratings and estimated by (P1· / N) – (P·1 / N).

Table 2.4
2x2 Agreement Matrix

		Correct Ratings		
		Yes	No	Total
Observer	Yes	*a*	*b*	P1·
	No	*c*	*d*	P2·
	Total	P·1	P·2	N

What Byrt et al. (1993) referred to as the "Prevalence Index" (PI) is the difference between the probability of the ratings on the diagonal. In a 2 x 2 rating table this would be estimated by the difference between (P1· + P·1) / 2 and (P2· + P·2) / 2. The limitations associated with Kappa gave rise to the development of several other alternatives (Gwet, 2008; Krippendorff, 2011; Lambert et al., 2021; Xie, 2013).

Krippendorff's Alpha

Krippendorff's Alpha was developed and first used in content analysis in counseling and survey research when coding open-ended interview data but can be applied in other contexts as a reliability coefficient (Krippendorff, 2011). Krippendorff (1970) noted the initial form of this coefficient was not easily calculated and stated, "since most of these assessments are done by computers there is no reason to give preference to a simpler solution which yields less information" (p. 70). The initial form of Krippendorff's Alpha was designed to improve the measurement of the reliability of interval data over Kendall's *W* (1948) and other non-parametric correlation coefficients (Krippendorff, 1970).

Since the initial form was first developed, Krippendorff's Alpha has established variations that are generalizations of other reliability indices. There are different calculations for Krippendorff's Alpha, and each variation has applications no matter the number of observers, number of categories, or type of variable being measured (nominal, ordinal, interval, or ratio). This is a criticism of Krippendorff's Alpha, it is a term used for several different coefficients that

are used for different purposes. One disadvantage of this coefficient is that most variations of Krippendorff's Alpha are quite complex to calculate (Stemler & Tsai, 2008). Also, there is not a minimum sample size, and Krippendorff's Alpha can be calculated with missing data (Krippendorff, 2011). Krippendorff Alpha's general equation is shown in Equation 2.3:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{2.5}$$

Where:

D_o = the observed disagreement

De = the disagreement expected when attributed to chance

The version of Krippendorff's Alpha most closely related to Kappa is the second calculation in Krippendorff (2011). This version of Krippendorff's Alpha is most similar to Kappa because it is designed for nominal data and evaluation of ratings between two observers (or a single rater against a set of correct ratings). In this case, Krippendorff's Alpha is calculated through what is shown in Equation 2.4:

$$\alpha_{nominal} = 1 - \frac{D_o}{D_e} = \frac{p_a - p_e}{1 - p_e} = \frac{(n-1)\sum_c o_{cc} - \sum_c n_x(n_x-1)}{n(n-1) - \sum_c n_x(n_x-1)} \tag{2.6}$$

Where:

n = the total number of ratings for two observers (or a single rater and correct answers)

n_x = the number of ratings across individual categories

o_{cc} = the observed coincidences of agreement

Gwet's AC1

Gwet's AC1 was introduced in 2001 (Gwet, 2001) and developed as an alternative to Kappa to assess IRR when there are multiple raters of the same performance. This IRR coefficient has demonstrated the ability to overcome issues related to the marginal distribution that Kappa collapses under (Gwet, 2008; Walsh et al., 2014; Xie, 2013). AC1 has the same range of potential values as Kappa (-1.0 to 1.0), however it is a more stable measure especially in the presence of high agreement or low-prevalence category use. While it is unlikely in practice for a chance-corrected agreement coefficient value to fall below zero in practice, it is possible (McHugh, 2012). AC1 produces coefficients close to the percentage of agreement while still accounting for the random chance of agreement. For AC1 both the overall agreement probability and the chance-corrected agreement probability need to be calculated. AC1 is "the conditional probability that two randomly selected raters might agree given that there is no agreement by chance" (Gwet, 2001). AC1 was designed to use with any number of raters using categorical rating systems. Formulas used to calculate AC1 are shown in Equations 2.7 – 2.9 below:

$$\pi_x = \frac{1}{n}\sum_{i=1}^{n} \frac{R_{ix}}{R} \tag{2.7}$$

Where:

R_{ix} = the number of raters who classified the ith object into the qth category

i ranges from 1 to *n*

x ranges from 1 to q

R = the total number of raters

$$p_e = \frac{1}{q-1}\sum_{x=1}^{q} \pi_x(1-\pi_x) \quad (2.8)$$

$$p_a = \frac{1}{n}\sum_{i=1}^{n}[\sum_{x=1}^{q}\frac{R_{ix}(R_{ix}-1)}{R(R-1)}] \quad (2.9)$$

Similar to the nominal Krippendorff's Alpha, Gwet's AC1 follows the same basic equation as Kappa, as shown in Equation 2.10:

$$\text{AC1} = \frac{p_a-p_e}{1-p_e} \quad (2.10)$$

Both overall agreement probability and chance agreement probability were estimated for AC1. The initial equation (solving for π_q) is used to calculate the probability that a rater classifies an object into a specific category, where p_a is the overall agreement probability and p_{ey} is the proportion of agreement by chance considering a random rating.

In clinical psychology applications, it was recommended to use AC1 in place of or at least along with Kappa (Wongpakaran et al., 2013) because it was a more stable measure than Kappa and simpler to calculate as compared to Krippendorff's Alpha. Another distinguishing trait of AC1 is that it does not require the assumption of independence between raters is met (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).

Lambda Coefficient of Rater-Mediated Agreement

Lambda was developed based on the theoretical framework of rater-mediated assessment theory (Engelhard & Wind, 2018) and first developed for use in teacher performance evaluations. Lambda was developed as an alternative to Kappa and was built on specific theory as opposed to simply addressing known issues with the Kappa coefficient. While most chance-corrected agreement coefficients were designed for nominal scales, Lambda is designed for ordinal rating scales, like those used for teacher evaluations. According to Lambert et al. (2021), Lambda follows a set of assumptions about raters and the complex cognitive process used by raters to select a rating. Lambda follows a similar basic formulaic structure as Kappa, yet Lambda adjusts the formula to incorporate the theoretical assumptions. While other chance-corrected agreement coefficients acknowledge that raters are most likely experts and provide qualified judgments (Cohen, 1960), this is built into Lambda's theoretical framework and calculations. Further distinguishing Lambda from other chance-corrected agreement coefficients, it can be used with rank order scales and categorical data.

The calculation for both Lambda 1 and Lambda 2 follows the same steps and is shown in Equations 2.11 – 2.13:

$$\lambda = \frac{p_a - p_e}{1 - p_e} \tag{2.11}$$

$$p_e = \sum p_s \, p_c p_f \tag{2.12}$$

$$\sigma_\lambda = \sqrt{\frac{p_a(1 - p_a)}{n\,(1 - p_e)^2}} \tag{2.13}$$

Where:

Σ = sum across all cells from $r = 1$, $c = 1$ to $r = q$, $c = q$

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

p_c = proportion of ratings for which the given category (q) 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 difference between the calculations in Lambda-1 and Lambda-2 is in the formula for p_s. In Lambda-1, this value is equal to one divided by the number of steps on the rating scale (1/q). Whereas in Lambda-2, the probability of use of each category is based on population values, when available. In the current study, the population values are from actual teacher evaluation data of all teachers rated in North Carolina in a given year on a four-point ordinal scale. The North Carolina Teacher Evaluation Process (NCTEP) implements a performance rating scale with four levels and a not demonstrated option across five teaching standards and 25 elements. The initial rating point is “Developing”, followed by “Proficient”, “Accomplished”, and “Distinguished” The population probability of using “Developing” (the first step on the ordinal scale) is 0.05, 0.65 for “Proficient” (the second step), 0.25 for “Accomplished” (the third step), and 0.05 for “Distinguished” (the fourth step on the NCTEP). This value for Lambda-2, “assumes the rater is uncertain about which rating to give, uses guessing as a means to arrive at a starting point for their deliberation, and their internal guessing process weights the points on the rating scale according to how frequently they use each point” (Lambert et al., 2021, p. 11).

In order to address the issue of providing valid, reliable, and fair ratings through rater-mediated assessments it is important to have a coefficient designed to match the purpose of the assessment being used. In contrast to other coefficients used and reported in rater-mediated assessments, Lambda was designed based on theory behind rater-mediated assessment (Lambert et al., 2021). Lambda is based on the complex rating process raters go through when using ordinal scales. It is necessary to test these theoretical propositions across a wide range of data conditions in order to address the coefficient's viability as an alternative to Kappa and other chance-corrected agreement coefficients.

Summary of IRR Coefficients

IRR coefficients were initially designed to provide information about the reliability of ratings beyond what is available through the use of percentage agreement values. Agreement values and chance-corrected agreement values provide information that is relevant for test users and test takers. Individual coefficients cannot provide information about the measurement properties of rater-mediated assessments. However, chance-corrected agreement coefficients provide information about the accuracy and quality of judgmental processes of raters (Engelhard & Wind, 2018). When selecting an IRR coefficient, it is imperative to consider if the coefficient being used in the rater-mediated assessment situation was designed for a similar purpose. Table 2.5 provides a quick summary of the initial purpose, use, and limitations of the IRR coefficients described in this section.

Table 2.5
Overview of Interrater Reliability Coefficients

Coefficient	Purpose and Use	Limitations
Cohen's Kappa	Originally proposed for use with nominal rating scales with two raters. Designed to account for the possibility of raters guessing when making ratings decisions. *	May produce overly low levels of reliability. When raters are well trained and guessing is unlikely, it is best to use and report another measure.
S	Adjusts the proportion of agreement according to the number of categories used. Improved measure than simple agreement between raters	Does not correct for chance-agreement. Can be inflated when categories are not used.
Krippendorff's Alpha	For any measurement scale, can handle missing data. designed as an alternative to Kappa, corrects for Kappa's known paradoxes	Values over 0.8 are preferred, 0.667 is lowest acceptable level**
Gwet's AC1	Designed as an alternative to Kappa, resistant to Kappa's known paradoxes	May under-correct for chance-agreement in areas Kappa over-corrects
Lambda-1	Designed according to rater-mediated assessment theory. For use with ordinal scales in rater-mediated assessments.	Limited applications and research evidence
Lambda-2	Proposed as an alternative to Lambda-1, for use when the population proportion of rating category use on an ordinal scale are known.	Limited applications and research evidence

Note. *McHugh (2012). **Krippendorff (2004)

Applications of Rater Agreement Statistics in Educational Research

Formative Assessments

One category of widely used rater mediated assessment are authentic formative assessments. Often formative assessments such as literacy skills assessments and early childhood developmental assessments include rater-mediated components (Wang et al., 2017). Typically,

these assessments require an evaluator, who is often the classroom teacher, to judge student performance on a construct according to their perceptions and interpretations of the assessed skill. By definition many formative assessment processes fall under the umbrella of rater-mediated assessment. The validity, reliability, and fairness of the scoring decisions on these types of rater-mediated assessments based on the performance ratings given to students are important and require careful attention.

Learning progressions offer developmental models for teachers to use to trace learning paths students may follow within a specific skill or concept (Trumball & Lash, 2013). It is unclear if and to what extent principles of measurement theory should be used with formative assessments (Bennett, 2011). However, measurement theory can be useful in some formative assessment situations, especially those requiring a teacher to make placements on a learning progression. It is important to consider how raters of student performance make quality inferences about what a student can or cannot do based on a collection of evidence. This can help provide evidence the inferences made by teachers are valid, reliable, and fair.

Researchers found in school settings implementing classroom-based formative assessments, “teachers developed deep expertise that translates into shared judgments and common mental models of what constitutes acceptable student performance on complex types of learning” (Darling-Hammond, 2017, p. 49). In a project titled Building Educator Assessment Literacy, nearly all teachers participating in the study claimed that participating in the scoring process deepened their understanding of the curriculum, assessment system, and standards (Daro & Wei, 2015). Successful implementation of rater-mediated formative assessment practices can offer high-quality professional learning opportunities for teachers, provide a way for teachers to

engage in reflective practices, and can allow teachers to have a greater sense of their students' learning trajectories (Maier et al., 2020).

Teacher Evaluation

In educational research and evaluation, there is a strong emphasis placed on the reliability and validity of student achievement outcomes. These outcomes have commonly been used as part of the teacher evaluation process in many states since the adoption of Race to the Top (RTT) policies. RTT placed an emphasis on the inclusion of performance pay and value-added measures in teacher evaluation processes (Bleiberg et al., 2021; Kraft & Gilmour, 2017; Murphy et al., 2013; Rodriguez & Hunter, 2021). However, there is not a similar focus on the validity, reliability, and fairness of teacher observation data. It is common for teacher evaluation systems to have unclear, or even absent, requirements related to the validity and reliability of scores from these observational instruments altogether (Herlihy et al., 2014). Previous research identified heavily context-specific factors that make scores from teacher evaluations valid and reliable (Cohen & Goldhaber, 2016; Darling-Hammond et al., 2012; Herlihy et al., 2014).

Several researchers claimed that raters are the largest source of error in evaluation systems (Casabianca et al., 2013; Cohen & Goldhaber, 2016; Hill et al., 2012). However, in all of these studies the rater was almost exclusively a school-level administrator. Some school districts and states require the use of an external evaluator or master rater during at least one observation (Adnot et al., 2017; Herlihy et al., 2014; Rockoff & Speroni, 2010). Administrators often are faced with the choice of offloading many of their duties to other staff members or allocating time to complete observations and provide teachers with necessary support and feedback (Firestone, 2014). Administrators may even rate teachers artificially higher to avoid

what they view as “disincentives” when having to conduct extra observations or provide extra support to lower rated teachers (Kraft & Gilmour, 2017).

Evaluation systems can be content and context specific, making the process of using an observation instrument complex and requiring a level of expertise to minimize subjectivity and enhance reliability of scores. These are measures suggested to enhance the overall quality of placements made across components of a given observation instrument. Steinberg & Sartain (2015) suggested the use of highly trained raters to conduct observations led to greater improvement in overall teacher quality. Other reports and studies debated who should be conducting observations and the best route to effectively train raters (Dee & Wyckoff, 2015; Hill & Herlihy, 2011; Sartain et al., 2010).

In teacher evaluation systems, rater accuracy refers to the ability of a rater to provide accurate scores from an observation against a set of ratings provided by an expert panel or master rater. It is typical in calibration training of raters that passing a rating certification training process involves raters being able to provide accurate scores on a given teacher evaluation instrument (Cash et al, 2012; Hill et al. 2012). Whether accuracy is of the most interest in a given rating of a teacher observation is dependent upon the purpose of the observation. In situations where the observations can have high stakes for teachers, a rater’s ability to provide accurate ratings of a teacher’s performance is critical.

Information from teacher evaluations can be used to provide formative feedback to help teachers grow their practice and direct a teacher’s professional development plan for improvement. Firestone (2014) added the feedback can also provide policymakers an understanding of the necessary conditions to facilitate good teaching. Teacher growth and

intentional professional development are two major factors in maintaining the best possible teacher workforce (Adnot et al., 2017; Almy, 2011; Herlihy et al., 2014; Hill & Herlihy, 2011).

A required evaluator (or rater) certification and recertification process is among the suggestions offered as necessary pieces supporting the validity and reliability of teacher evaluation scores (Zepeda & Jimenez, 2019). An essential part of periodic training and recertification of raters of teachers involves investigation of IRR. This training and certification process looks different across grade levels, subjects, and states. In some cases, raters co-rate lessons during observations with a certified rater and compare scores. However, in many states the decision regarding who observes and how the rater is credentialed to conduct observations is left up to a local school district. Regardless of these decisions most states and local districts do not attend to multiple issues related to teacher evaluation systems. Almost all states omit calculating IRR rates and statistics as a measure of reliability of scores produced by teacher evaluators (Herlihy et al., 2014). This is not done in practice because in most districts and states the majority of evaluations are conducted by a single observer due to logistical and financial constraints.

The analysis and interpretation of rating quality beyond agreement percentages are necessary parts of teacher evaluation systems. The rating decisions made by teacher evaluators play a critical role in determining the effectiveness of a teacher. The use of chance-corrected agreement coefficients supplements efforts to provide valid and reliable scores of teacher performance. Teacher performance evaluation ratings should be invariant to rater effects, meaning the ratings a teacher receives should not be dependent upon which evaluator conducts the observations.

Validity and Reliability of Teacher Evaluation Processes

There is an extensive body of literature questioning the validity of teacher evaluation processes due to issues with value-added measures being incorporated into overall evaluation ratings (Amrein-Bearsley, 2008; Bitler et al., 2021; Darling-Hammond et al., 2012; Haertel, 2013; Hill et al., 2011; Papay. 2010). However, the focus of this section is on the validity and reliability of scores from teacher evaluations conducted by school administrators or other teacher evaluators. As teacher evaluations became increasingly more high-stakes, evidence of the validity of ratings provided by evaluators has grown in importance (Kimball & Milanowski, 2009). Previous iterations of changes in the teacher evaluation landscape across the country dissuaded principals from making low ratings of teacher performance as it required principals to spend a great breadth of time developing extensive, and in some cases multiple, improvement plans for teachers rated as ineffective (Kraft & Gilmour, 2016; Weisberg et al., 2009).

One study focused on identifying teacher evaluators as providing more or less valid teacher evaluation scores according to the relationship between ratings, qualitative data from retrospective cognitive interviews, and student achievement (Kimball & Milanowski, 2009). This study provided evidence that there were differences in how certain evaluators used evaluation rubrics in more analytical ways than other evaluators (Kimball & Milanowski, 2009). In this study, all evaluators were school principals or assistant principals. Kimball and Milanowski (2009) found evaluators with more training and experience provided more valid ratings of teacher quality. Other important evidence researchers used to constitute evaluators ratings as valid included the importance of the rater having a positive attitude about the evaluation system, and other rater behavior such as gathering extensive evidence when making ratings decisions, taking notes throughout observations, and having an open working environment with teachers.

Researchers concluded many raters "constructed their own meaning (of the teacher evaluation process) by adapting an evaluation process that is acceptable to them, their teachers, and their school environments" (Kimball & Milanowski, 2009, p. 62) and suggested extensive rater training to standardize the use of evaluation instruments in order to ensure ratings are consistent and fair across an evaluation system.

In a study aimed at addressing whether principals were effectively promoting teacher development through the implementation of a new teacher evaluation process, researchers discovered principals did not always view or implement the evaluation system as articulated in their trainings (Kraft & Gilmour, 2016). The differences in perspectives raters may have of the purpose of an evaluation system and the variability in their implementation practices are validity concerns. In this context, a majority of the evaluators were using the evaluation system as opportunities to help teachers improve their practice while other evaluators deemed the purpose of the same evaluation system to drive out lower-performing teachers. This nuanced view in the purpose of the same evaluation process calls into question the validity and reliability of scores produced by certain evaluators.

Construct Irrelevant Variance in Teacher Evaluation

As stated in the *Standards*, construct irrelevant variance refers to the amount "scores may be systematically influenced to some extent by processes that are not part of the construct" (AERA et al., 2014, p. 13). Construct irrelevant variance can negatively impact the quality of teacher evaluation scores. Among many factors related to the existence of variance in teacher evaluation systems, some primary factors contributing to construct irrelevant variance can include the lesson observed, the rater, and the observational instrument in use (Hill et al., 2012).

Some studies have attempted to address this issue through the use of multiple raters. In the case of a teacher evaluation system, construct-irrelevant variance may be added to the situation according to the time of day the observation occurs, the subject the lesson the teacher is focused on, the group of students the teacher is working with during that particular lesson and/or subject, etc. The list of factors in the area of construct-irrelevance involved with teacher evaluation systems and classroom observation is countless (Garrett & Steinberg, 2015; Steinberg & Garrett, 2016; Whitehurst et al., 2014) and have great consequences for establishing and maintaining a positive school climate (Kraft & Gilmour, 2016).

Chance-Corrected Agreement Coefficients: Teacher Evaluation Contexts

Jimenez and Zepeda (2020) examined the interrater reliability of 42 principals and assistant principals from one school district in a southeastern state. Each participant rated four teachers using videos of actual lessons in a rater calibration training exercise. Each teacher was rated using the same rubric across 25 teaching elements, meaning each rater provided a total of 100 ratings used to calculate each reliability coefficient. While the administrators only rate teachers from grades taught within their school buildings, for the purpose of the study each participant rated two elementary school teachers, one middle school teacher, and one high school teacher. The researchers wanted to determine if percent exact agreement and variations of Kappa were adequate measures of IRR. Also, Jimenez and Zepeda (2020) tested if known issues with Kappa were prevalent in their findings and how Gwet's AC1 statistic performed in comparison to overall agreement and Kappa values in a teacher evaluation context. Findings from this study emphasized there is a need to include measures of IRR in teacher evaluation contexts and recommended the use of AC1 in place of Kappa in teacher evaluation studies (Jimenez & Zepeda, 2020).

In the initial application of the Lambda Coefficient of Rater-Mediated Agreement, a group of 57 professional evaluators rated ten online teacher profiles across five progressions for a total of 2,850 ratings (Lambert et al., 2021). Similar to Jimenez and Zepeda's (2020) study, all evaluators were using the same instrument to rate teachers. Also, Lambert et al.'s (2021) study calculated and analyzed the exact agreement percentage, Kappa, and AC1 statistics. Two important distinctions are the Lambert et al. (2021) study (1) focused on professional evaluators of early childhood educators and (2) added calculations and interpretations for discrepant rating percentages, Gwet's AC2, Lambda-1, and Lambda-2. This study confirmed the shortcomings of Kappa in a teacher evaluation context and the robustness of AC1 to overcome shortcomings associated with Kappa. In this introductory application of Lambda, Lambda-1 demonstrated robustness to data conditions that are problematic for Kappa (Lambert et al., 2021).

Conclusion

This study targets areas in education that require careful attention to ensure teachers are provided with valid and reliable information to improve their practice and student learning. Evaluation of the validity and reliability of scores produced by raters is essential for maintaining the usefulness of teacher evaluation systems and other widely used rater-mediated assessments. There are many factors to consider and steps that go into ensuring the quality of ratings and scores provided in a rater-mediated assessment. It is important that assurances are made to test users that information from these assessments are valid, reliable, and fair. As education systems continue to improve current rater-mediated assessments, and adopt new systems, they should focus on empirical findings and establish rater-mediated assessment processes that point to strategies to strengthen their evaluation systems.

Problems related to traditionally used chance-corrected agreement coefficients highlight the need for further development and evaluation of new methods that are designed to handle the complexities of rater-mediated assessment. The purpose of this study was to test the functioning of Lambda through evaluating potential data conditions using two-, three-, and four-point rating scales posing similar situations as real-world data conditions. Rating scales of these sizes were selected because it is necessary to evaluate this new statistical method to ensure it works in the scenarios for which it was designed (Lambert et al., 2021; Morris et al., 2019). According to Morris et al. (2019), the current study can be categorized as an "absolute evaluation" and "comparative evaluation" study (p. 2075). An absolute evaluation occurs when a new method is examined using a data-generating mechanism to verify the statistic works for the intended purposes and to compare how it performs relative to other statistical methods (Morris et al., 2019). A common pairing to an absolute evaluation is the comparative evaluation, which is a comparison of the new technique to other similarly applied and developed statistical methods. In this study, the data-generating mechanism of interest involved making repeated adjustments of preselected and determined data conditions. Issues of sampling variability, random sampling of ratings, sampling error, or sample size are not addressed in this study. This was unnecessary in this application because in this absolute and comparative evaluation the performance of coefficients at specified level of agreement is examined and is agnostic to the rater. This was part of the simulation design so that the behavior of the statistic was isolated and not dependent upon varying rater characteristics. These methods followed the simulation procedure that evaluated the AC1 statistic in comparison to Kappa (Xie, 2013).

This study examines the following research questions:

1. Under what conditions do Cohen's Kappa, *S*, Krippendorff's Alpha, Gwet's AC1, Lambda-1, and Lambda-2 produce similar values?
2. To what extent are the IRR coefficients, calculated using Cohen's Kappa, *S*, Krippendorff's Alpha, Gwet's AC1, Lambda-1, and Lambda-2, placed in the same classification categories according to well-known taxonomies (i.e., almost perfect, substantial, moderate, fair, slight, and no agreement) under the same levels of agreement?
3. To what extent do Cohen's Kappa, *S*, Krippendorff's Alpha, Gwet's AC1, Lambda-1, and Lambda-2 measure agreement in similar ways?

Again, this study was performed to evaluate the performance of Lambda as compared to other chance-corrected agreement coefficients under varied data conditions similar to what is found in rater-mediated assessment applications.

Chapter III: Methodology

Previous research demonstrates the need for further investigation of an alternative to existing IRR coefficients (Gwet, 2008; Lambert et al., 2021; Warrens, 2012; Xie, 2013). Lambda was designed according to rater-mediated assessment theory and empirical results from the first application of this coefficient were promising (Lambert et al., 2021). Further evaluation of this new statistic was required in order to provide validity evidence of its potential use. Specifically, comparing the performance of Lambda-1 and Lambda-2 to common IRR coefficients in conditions similar to what are commonly found in rater-mediated assessment was a next step in assessing the usability of Lambda-1 and Lambda-2.

The performance of Lambda-1 and Lambda-2 under data conditions similar to real-world data scenarios and conditions known to be problematic for other commonly used chance-corrected agreement coefficients were assessed through utilizing a data generation mechanism. The data generation mechanism in this study involved adjusting the size of the rating scale, amount of agreement, location of agreement, amount of disagreement, and location of disagreement. Data generated allowed for the comparison of the performance of Lambda-1 and Lambda-2 to Kappa, *S*, Gwet's AC1, and Krippendorff's Alpha. This allowed for the analysis of coefficient values and assessed where differences were found according to the size of the rating scale, amount of agreement/disagreement, or location of agreement/disagreement.

Simulation Factors

To address known issues with commonly used and reported chance-corrected agreement coefficients and the recently proposed Lambda coefficient (Lambert et al., 2021), the percentage of agreement, location of agreement, percentage of disagreement, and location of disagreement

were varied as simulation factors. Hypothetical distributions of observed agreement rates between raters were planned to create ratings distributions according to two-category (2x2), three-category (3x3), and four-category (4x4) agreement matrices. Proportions were used in place of specific counts of ratings since calculations of chance-agreement coefficients reduce expected and exact agreement to proportions. High levels of agreement ranging from 75% to 95% (Altman, 1991; Fleiss, 1981; Landis & Koch, 1977) were used as the first variation in data condition. This variation was followed by adjusting the location of agreement and location of disagreement across specific cells within the applicable rating matrix. The calculation of the marginal distributions for each data condition was adjusted as a result of varying the location of agreement and disagreement within each matrix. The following section provides an overview and justification for the adjustments made in each data condition.

Location of Agreement and Disagreement

2x2 Agreement Matrix. The 2x2 agreement matrix, used for a two-point rating scale, was used to vary the agreement in increments of five percentage points between 75-95% across cells *a* and *d* (see Figure 3.1, all agreement matrix figures are adapted from Lambert et al., 2021) for each data condition generated. Cell *a* in Figure 3.1 refers to the count of ratings where the rater decided on the first rating scale point (labeled as "1") and agreed with the correct answer of "1". Cell *b* is a location of disagreement between the rater and the correct rating. This cell refers to instances when the rater scores the performance using the second rating scale point, but the correct rating was the first rating scale point. This is an example of a strict rating, where the rater underestimated the actual performance. Cell *c* provides another count of disagreement between the rater and the correct answers. In *c* the rater scored the performance at the second rating scale point, but the correct answer was the first scale point. This cell (*c*) is an example of a lenient

rating, where the rater overestimated the actual performance. Cell *d* is a count of correct ratings in this context. This cell is the count of correct ratings made by the rater using the second point on the rating scale (labeled as “2”).

The overall amount of agreement (75%-95%), specific rating scale points where agreement occurs (cells *a* and *d*), and location of disagreement (in cell *b* and/or *c*) were adjusted according to the data simulation conditions. A total of 15 data simulation conditions were performed using the 2x2 agreement matrix. These are explained in more detail and examples are provided later in this chapter.

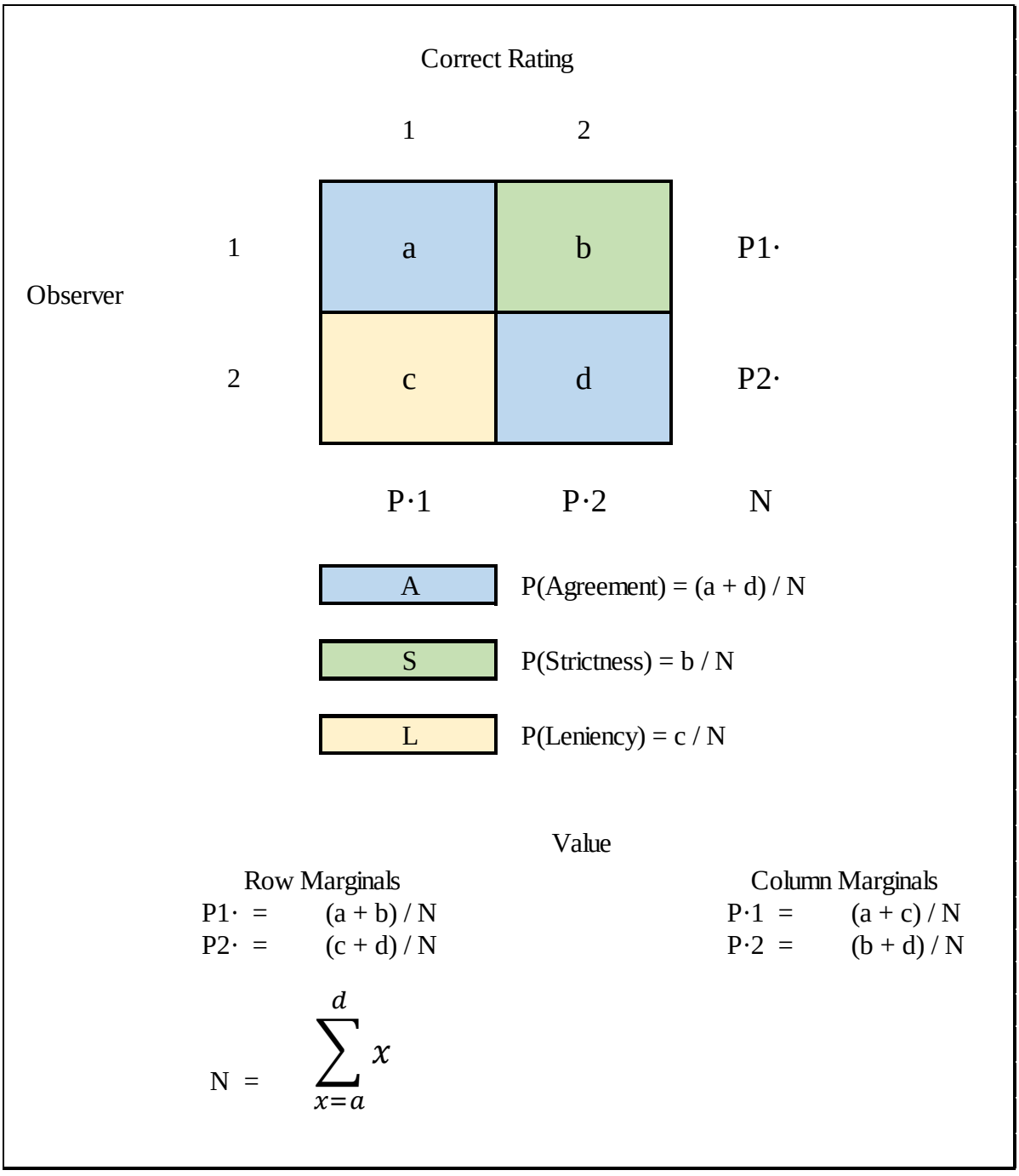


Figure 3.1. 2x2 agreement matrix probability distribution.

3x3 Agreement Matrix. The 3x3 agreement matrix, used for a three-point rating scale, were used to vary the agreement in increments of five percentage points between 75-95% across cells *a*, *e*, and *i* (see Figure 3.2) for each generated data condition. These three cells (*a*, *e*, and *i*), shaded in blue in Figure 3.2, are locations of correct or accurate ratings made by the rater. Cell *a* refers to the count of ratings where the rater decided on the first rating scale point (labeled as "1") and agreed with the correct answer of "1". Cell *e* is a count of correct ratings made by the rater using the second point on the rating scale (labeled as "2"). Cell *i* is the count of correct ratings made using the third point on the rating scale (labeled as "3").

Cell *b*, *c*, and *f* are locations of disagreement between the rater and the correct rating. These cells (shaded in green in Figure 3.2) refer to instances where the rater provided strict ratings of the performance. Cell *b* counts occur when a rater uses the first scoring point but the correct answer was the second rating point. Cell *c* provides another count of disagreement between the rater and the correct answers. In *c* the rater scored the performance at the first rating scale point, but the correct answer was the third scale point. Cell *f* provides another count of disagreement between the rater and the correct answers. In *f* the rater scored the performance at the second rating scale point, but the correct answer was the third scale point. Each of these three cells (*b*, *c*, and *f*) are examples of "strict" ratings, where the rater underestimated the actual performance.

Cell *d*, *g*, and *h* are also locations of disagreement between the rater and the correct rating. These cells (shaded in yellow in Figure 3.2) refer to instances where the rater provides lenient ratings of the performance. Cell *d* counts occur when a rater uses the second scoring point, but the correct answer was the first rating point. Cell *g* provides another count of disagreement between the rater and the correct answers. In *g* the rater scored the performance at

the third rating scale point, but the correct answer was the first scale point. Cell *h* provides another count of disagreement between the rater and the correct answers. In *h* the rater scored the performance at the third rating scale point, but the correct answer was the second scale point. Each of these three cells (*d*, *g*, and *h*) are examples of "lenient" ratings, where the rater overestimated the actual performance.

The overall amount of agreement (75%-95%), specific rating scale points where agreement occurs (cells *a*, *e*, and *i*), and location of disagreement (in cells *b*, *c*, *d*, *f*, *g*, and/or *i*) were adjusted according to the data simulation conditions. In total, there were 35 data condition variations under the 3x3 agreement matrix. These are explained in more detail and examples are provided later in this chapter.

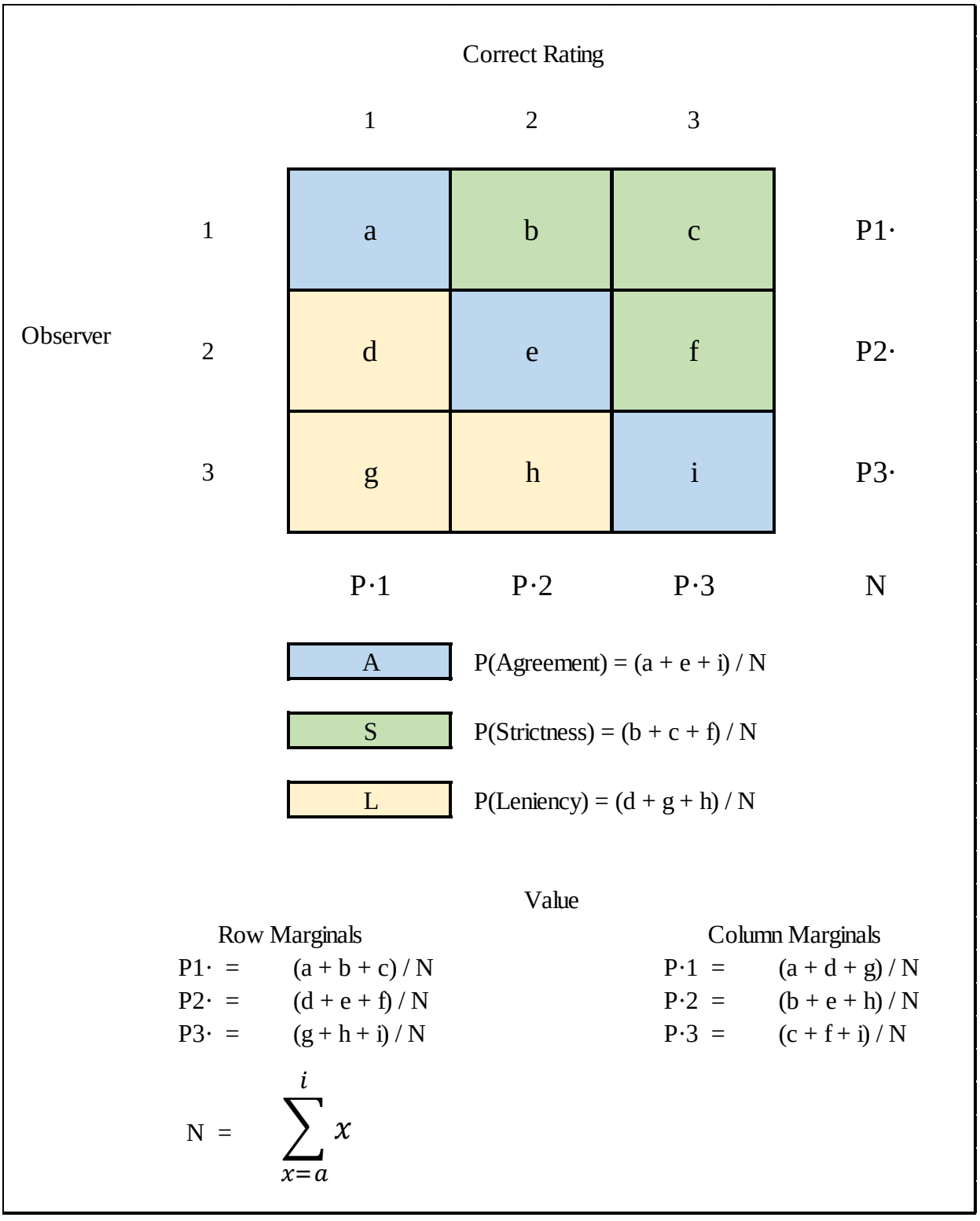


Figure 3.2. 3x3 agreement matrix probability distribution.

4x4 Agreement Matrix. The 4x4 agreement matrix, used for a four-point rating scale, were used to vary the agreement in increments of five percentage points between 75-95% across cells *a*, *f*, *k*, and *p* (see Figure 3.3) for each generated data condition. These four cells (*a*, *f*, *k*, and *p*), shaded in blue in Figure 3.3, are locations of correct or accurate ratings made by the rater. Cell *a* refers to the count of ratings where the rater decided on the first rating scale point (labeled as "1") and agreed with the correct answer of "1". Cell *f* is a count of correct ratings made by the rater using the second point on the rating scale (labeled as "2"). Cell *k* is the count of correct ratings made using the third point on the rating scale (labeled as "3"). Cell *p* is the count of correct ratings made using the fourth point on the rating scale (labeled as "4").

Cell *b*, *c*, *d*, *g*, *h*, and *l* are locations of disagreement between the rater and the correct rating. These cells (shaded in green in Figure 3.3) refer to instances where the rater provided strict ratings of the performance. Cell *b* counts occur when a rater uses the first scoring point, but the correct answer was the second rating point. In *c* the rater scored the performance at the first rating scale point, but the correct answer was the third scale point. Cell *d* provides a count of disagreement where the rater scored the performance at the first rating scale point, but the correct answer was the fourth scale point. Cell *f* counts occur when a rater uses the second scoring point but the correct answer was the third rating point. In *h* the rater scored the performance at the second rating scale point, but the correct answer was the fourth scale point. Cell *l* provides a count of disagreement where the rater scored the performance at the third rating scale point, but the correct answer was the fourth scale point. Each of these cells (*b*, *c*, *d*, *g*, *h*, and *l*) are examples of "strict" ratings, where the rater underestimated the actual performance.

Cell *e*, *i*, *j*, *m*, *n*, and *o* are locations of disagreement between the rater and the correct rating. These cells (shaded in yellow in Figure 3.3) refer to instances where the rater provided

lenient ratings of the performance. Cell *e* counts occur when a rater uses the second scoring point, but the correct answer was the first rating point. In *i* the rater scored the performance at the third rating scale point, but the correct answer was the first scale point. Cell *j* provides a count of disagreement where the rater scored the performance at the third rating scale point, but the correct answer was the second scale point. Cell *m* counts occur when a rater uses the fourth scoring point, but the correct answer was the first rating point. In *n* the rater scored the performance at the fourth rating scale point, but the correct answer was the second scale point. Cell *o* provides a count of disagreement where the rater scored the performance at the fourth rating scale point, but the correct answer was the third scale point. Each of these cells (*e*, *i*, *j*, *m*, *n*, and *o*) are examples of "lenient" ratings, where the rater overestimated the actual performance.

The overall amount of agreement (75%-95%), specific rating scale points where agreement occurs (cells *a*, *f*, *k*, and *p*), and location of disagreement (in cells *b*, *c*, *d*, *e*, *f*, *g*, *h*, *i*, *j*, *l*, *m*, *n*, and *o*) were adjusted according to the data simulation conditions. A total of 65 data conditions were simulated using the 4x4 agreement matrix (see Figure 3.3). Simulations were performed according to agreement percentage cell location, agreement amount, and location of the disagreement. These are explained in more detail and examples are provided later in this chapter.

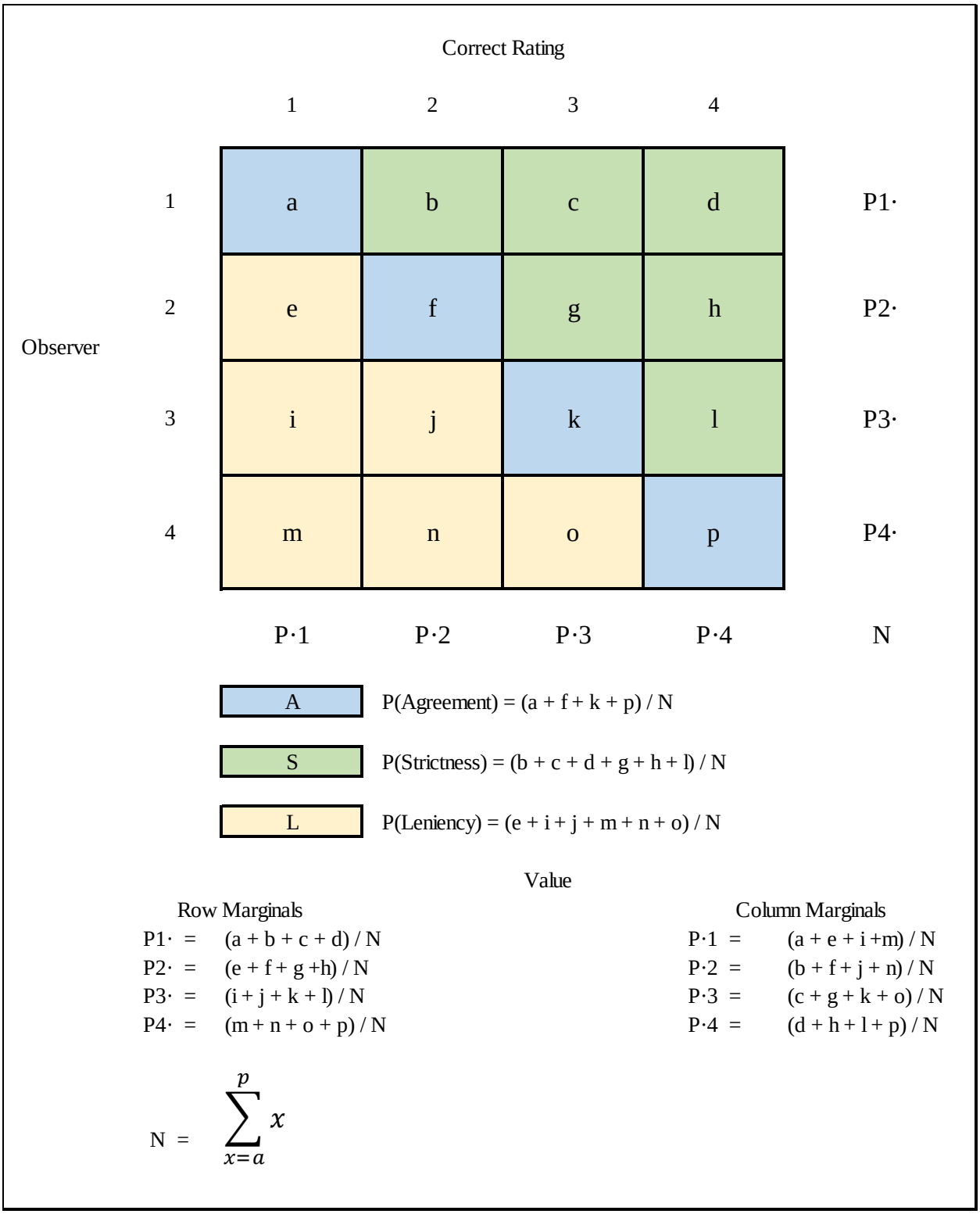


Figure 3.3. 4x4 agreement matrix probability distribution.

Marginal Distributions

Through adjusting the amount and location of agreement and disagreement in each of the two-, three-, and four-point matrices the marginal distribution were altered under each data condition. The calculation of row and column marginals is shown in Figures 3.1, 3.2, and 3.3. The row and column marginal products were used in the calculation of the expected chance agreement (p_e) portion of the chance-corrected agreement coefficients used in this study. Kappa relies exclusively on these marginals in the calculation of p_e (Xie, 2013).

Summary

Across the three different rating scales a total of 115 data conditions were produced and analyzed. Some data conditions repeated estimates from previous data conditions and duplicate calculations were removed from the full analysis, this caused the number of final data conditions to slightly decrease in the final study (from 115 to 95).

Data Generation Conditions

2x2 Data Conditions. Table 3.1 provides a full overview of the two-point rating scale data conditions. Table 3.1 should be used with Figure 3.1, which illustrates the location of cells according to each rating made by individual raters. In the 2x2 data conditions the amount of agreement will vary across two cells (“a” and “d”). For example, the row labeled “1” specifies the amount of agreement as 95% and the amount of disagreement as 5%. The location of the agreement will capture all possible combinations totaling the amount of agreement (95%) across the two agreement cells. The location and amount of disagreement is specified in the columns labeled “Strictness” and “Leniency” in Table 3.1. Continuing the example from row “1”, all disagreement would be located in cell *b* and the leniency in this condition would be equal to zero. The number of simulations within each row of data conditions was dependent upon the

number of possible combinations of the agreement value. In the example provided in Table 3.1 for row "1" this was 95, 0; 94, 1; 93, 2…to 1, 94; 0, 95. The value in cell *b* was fixed at 5% and cell *c* was fixed at 0%. Figure 3.4 provides an example of the distribution of agreement proportions according to the first data condition shown in Table 3.1.

Table 3.1

2x2 data simulation conditions

	Magnitude		Location of Agreement and Disagreement		
	Agreement (%)	Disagreement (%)	Agreement Distribution	Strictness	Leniency
1	95	5	varies from a = 95, d = 0 to a = 0, d = 95	b	-
2	90	10	varies from a = 90, d = 0 to a = 0, d = 90	b	-
3	85	15	varies from a = 85, d = 0 to a = 0, d = 85	b	-
4	80	20	varies from a = 80, d = 0 to a = 0, d = 80	b	-
5	75	25	varies from a = 75, d = 0 to a = 0, d = 75	b	-
6	95	5	varies from d = 0, a = 95 to d = 95, a = 0	-	c
7	90	10	varies from d = 0, a = 90 to d = 90, a = 0	-	c
8	85	15	varies from d = 0, a = 85 to d = 85, a = 0	-	c
9	80	20	varies from d = 0, a = 80 to d = 80, a = 0	-	c
10	75	25	varies from d = 0, a = 75 to d = 75, a = 0	-	c
11	95	5	varies from a = 95, d = 0 to a = 0, d = 95	b	c
12	90	10	varies from a = 90, d = 0 to a = 0, d = 90	b	c
13	85	15	varies from a = 85, d = 0 to a = 0, d = 85	b	c
14	80	20	varies from a = 80, d = 0 to a = 0, d = 80	b	c
15	75	25	varies from a = 75, d = 0 to a = 0, d = 75	b	c

Conditions listed in rows 1-5 in Table 3.1 adjusted the total amount of agreement and disagreement as indicated. In the data conditions in rows 1-5 strictness was constant in cell *b* as the total amount of disagreement and leniency constant in cell *c* at 0%. The data conditions in Table 3.1 rows 6-10 followed the same pattern as rows 1-5 for agreement totals, disagreement totals, and agreement distribution. The location and amount of disagreement in data conditions 6-10 were all in the lenient cell *c* and the strictness cell was equal to 0% in these conditions. Figure 3.5 shows the distribution of agreement proportions for data condition 8 from Table 3.1. Finally, the conditions listed in Table 3.1 rows 11-15 followed the same pattern as rows 1-10 for

agreement totals, disagreement totals, and agreement distribution. The location and amount of disagreement was equally distributed across cells *b* and *c*. Figure 3.6 shows the distribution of agreement proportions for data condition 15 from Table 3.1.

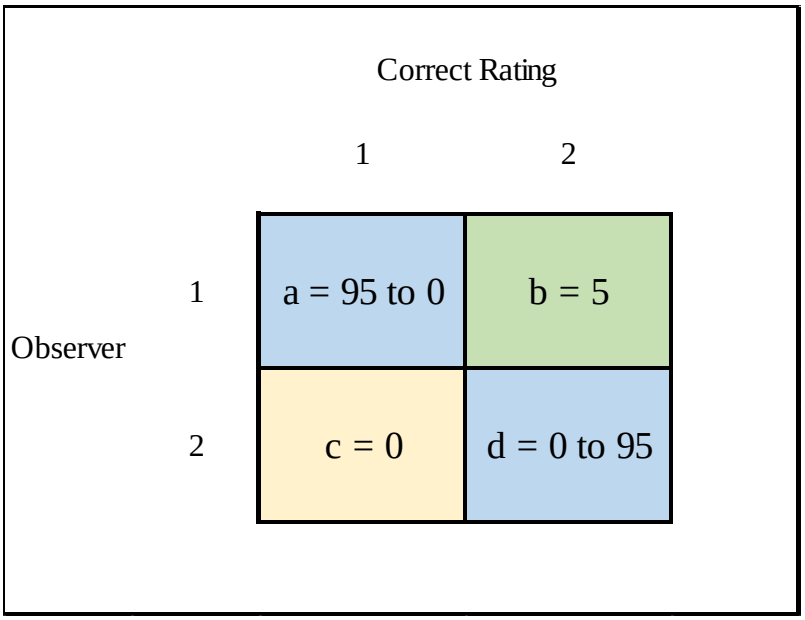


Figure 3.4. 2x2 agreement distribution for data condition 1 from Table 3.1.

These conditions were selected because they capture situations where raters have high accuracy values and low prevalence of category use. These conditions are known to offer complications for commonly used chance-corrected agreement coefficients. The repeated adjustments of the amount of agreement, amount of disagreement, location of agreement, and location of disagreement were the first component in calculating the coefficient values in the current study.

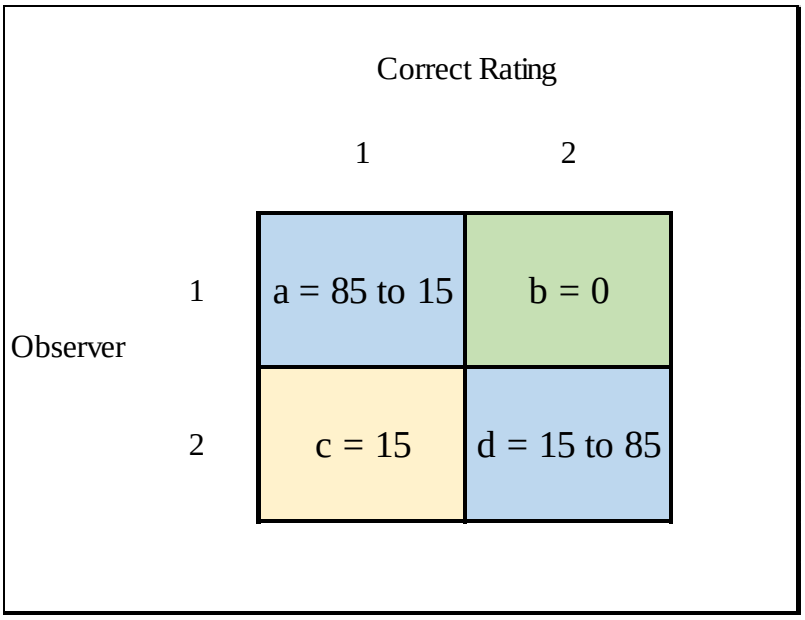


Figure 3.5. 2x2 agreement distribution for data condition 8 from Table 3.1.

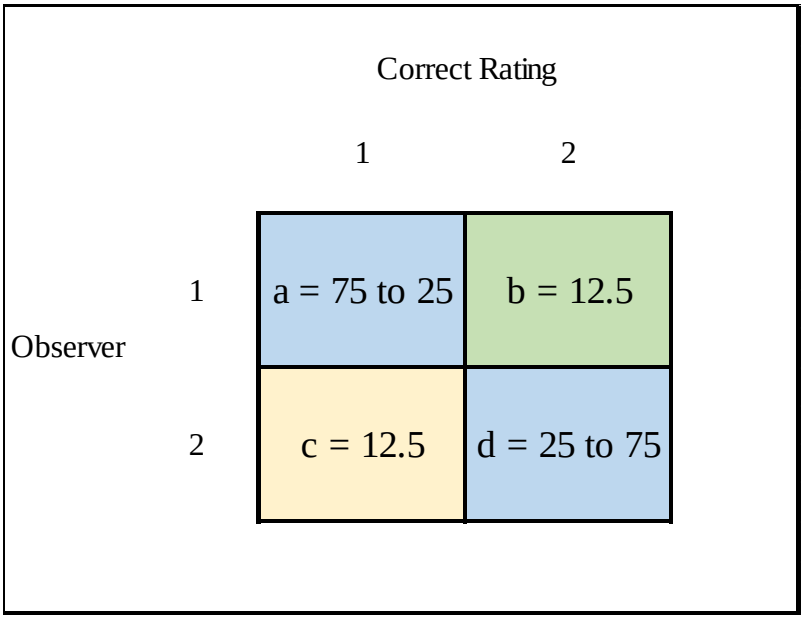


Figure 3.6. 2x2 agreement distribution for data condition 15 from Table 3.1.

3x3 Data Conditions. Table 3.2 provides a full overview of the data conditions using the three-point rating scale. Table 3.2 should be used with Figure 3.2, which illustrates the location

of cells according to each rating made by individual raters. The 3x3 data conditions were more involved than the 2x2 conditions given there were three agreement cells and six cells showing disagreement. As shown in Table 3.2, the amount of agreement was adjusted similar to how it was varied in the 2x2 conditions, however the location of the agreement varied across two cells in some instances (in data condition rows 1-20) and across all three agreement cells in other conditions (data conditions 21-35). Similarly, the amount of disagreement in the 3x3 conditions varied similarly to how it was adjusted in the 2x2 conditions, however the location of disagreement varied from being located in a single strict or lenient cell (conditions 1-20), two strict or lenient cells (conditions 21-30), or up to four strict or lenient cells (conditions 31-35). Again, these repeated adjustments were made because the modifications resulted in providing a variety of row and column marginals that impacted the value of the IRR coefficients.

Table 3.2

3x3 data simulation conditions

	Magnitude		Location of Agreement and Disagreement		
	Agreement (%)	Disagreement (%)	Agreement Distribution	Strictness	Leniency
1	95	5	varies from a = 95, e = 0 to a = 0, e = 95	b	-
2	90	10	varies from a = 90, e = 0 to a = 0, e = 90	b	-
3	85	15	varies from a = 85, e = 0 to a = 0, e = 85	b	-
4	80	20	varies from a = 80, e = 0 to a = 0, e = 80	b	-
5	75	25	varies from a = 75, e = 0 to a = 0, e = 75	b	-
6	95	5	varies from a = 95, e = 0 to a = 0, e = 95	-	d
7	90	10	varies from a = 90, e = 0 to a = 0, e = 90	-	d
8	85	15	varies from a = 85, e = 0 to a = 0, e = 85	-	d
9	80	20	varies from a = 80, e = 0 to a = 0, e = 80	-	d
10	75	25	varies from a = 75, e = 0 to a = 0, e = 75	-	d
11	95	5	varies from e = 95, i = 0 to e = 0, i = 95	f	-
12	90	10	varies from e = 90, i = 0 to e = 0, i = 90	f	-
13	85	15	varies from e = 85, i = 0 to e = 0, i = 85	f	-
14	80	20	varies from e = 80, i = 0 to e = 0, i = 80	f	-
15	75	25	varies from e = 75, i = 0 to e = 0, i = 75	f	-
16	95	5	varies from e = 95, i = 0 to e = 0, i = 95	-	h
17	90	10	varies from e = 90, i = 0 to e = 0, i = 90	-	h
18	85	15	varies from e = 85, i = 0 to e = 0, i = 85	-	h
19	80	20	varies from e = 80, i = 0 to e = 0, i = 80	-	h
20	75	25	varies from e = 75, i = 0 to e = 0, i = 75	-	h
21	95	5	varies from a = 95, e = 0, i = 0 to a = 0, e = 95, i = 0 to a = 0, e = 0, i = 95	b and f	-
22	90	10	varies from a = 90, e = 0, i = 0 to a = 0, e = 90, i = 0 to a = 0, e = 0, i = 90	b and f	-
23	85	15	varies from a = 85, e = 0, i = 0 to a = 0, e = 85, i = 0 to a = 0, e = 0, i = 85	b and f	-
24	80	20	varies from a = 80, e = 0, i = 0 to a = 0, e = 80, i = 0 to a = 0, e = 0, i = 80	b and f	-
25	75	25	varies from a = 75, e = 0, i = 0 to a = 0, e = 75, i = 0 to a = 0, e = 0, i = 75	b and f	-
26	95	5	varies from a = 95, e = 0, i = 0 to a = 0, e = 95, i = 0 to a = 0, e = 0, i = 95	-	d and h
27	90	10	varies from a = 90, e = 0, i = 0 to a = 0, e = 90, i = 0 to a = 0, e = 0, i = 90	-	d and h
28	85	15	varies from a = 85, e = 0, i = 0 to a = 0, e = 85, i = 0 to a = 0, e = 0, i = 85	-	d and h
29	80	20	varies from a = 80, e = 0, i = 0 to a = 0, e = 80, i = 0 to a = 0, e = 0, i = 80	-	d and h
30	75	25	varies from a = 75, e = 0, i = 0 to a = 0, e = 75, i = 0 to a = 0, e = 0, i = 75	-	d and h
31	95	5	varies from a = 95, e = 0, i = 0 to a = 0, e = 95, i = 0 to a = 0, e = 0, i = 95	b and f	d and h
32	90	10	varies from a = 90, e = 0, i = 0 to a = 0, e = 90, i = 0 to a = 0, e = 0, i = 90	b and f	d and h
33	85	15	varies from a = 85, e = 0, i = 0 to a = 0, e = 85, i = 0 to a = 0, e = 0, i = 85	b and f	d and h
34	80	20	varies from a = 80, e = 0, i = 0 to a = 0, e = 80, i = 0 to a = 0, e = 0, i = 80	b and f	d and h
35	75	25	varies from a = 75, e = 0, i = 0 to a = 0, e = 75, i = 0 to a = 0, e = 0, i = 75	b and f	d and h

The data conditions in rows 1-20 of Table 3.2 varied the agreement across just two of the cells, and the strictness or leniency in a single cell. These conditions were selected because they provided a ratings distribution where the hypothetical rater had high agreement and limited category use. In data conditions 1-20, the rater would hypothetically be using two out of the three possible ratings categories. Figures 3.7 – 3.10 provide sample agreement distributions for specific data conditions from the 3x3 agreement matrix.

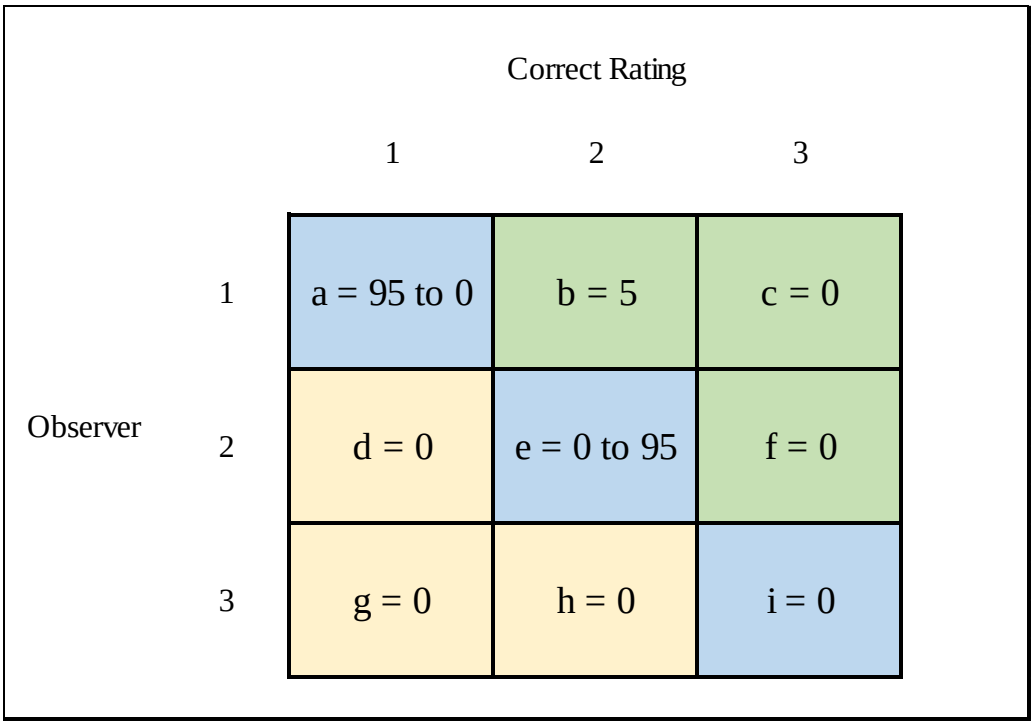


Figure 3.7. 3x3 agreement distribution for data condition 1 from Table 3.2.

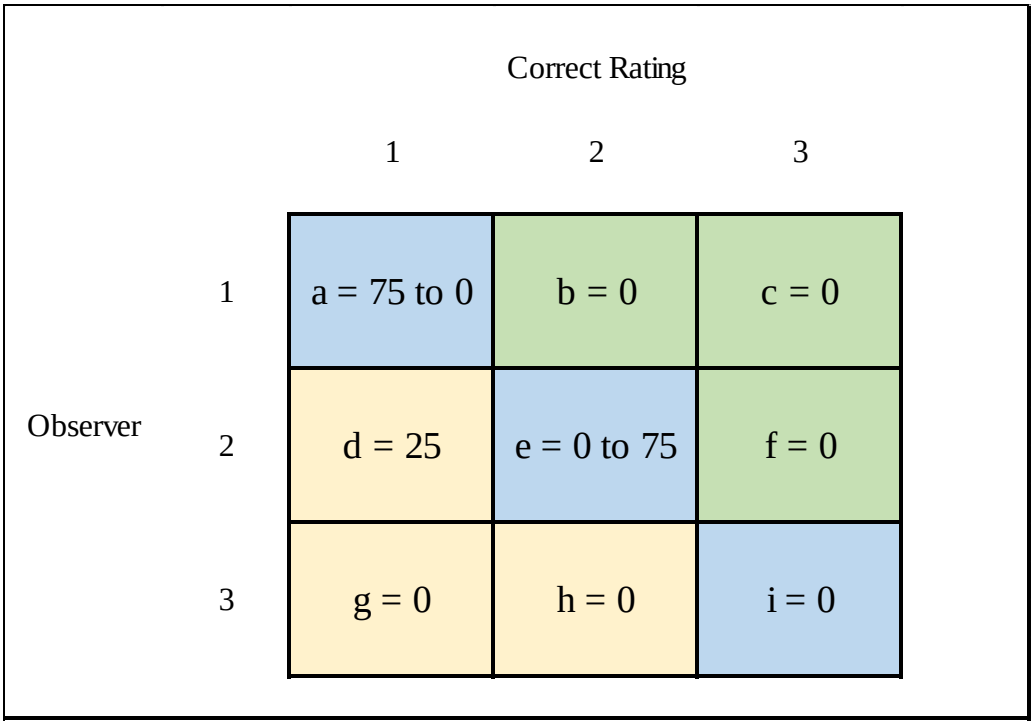


Figure 3.8. 3x3 agreement distribution for data condition 10 from Table 3.2.

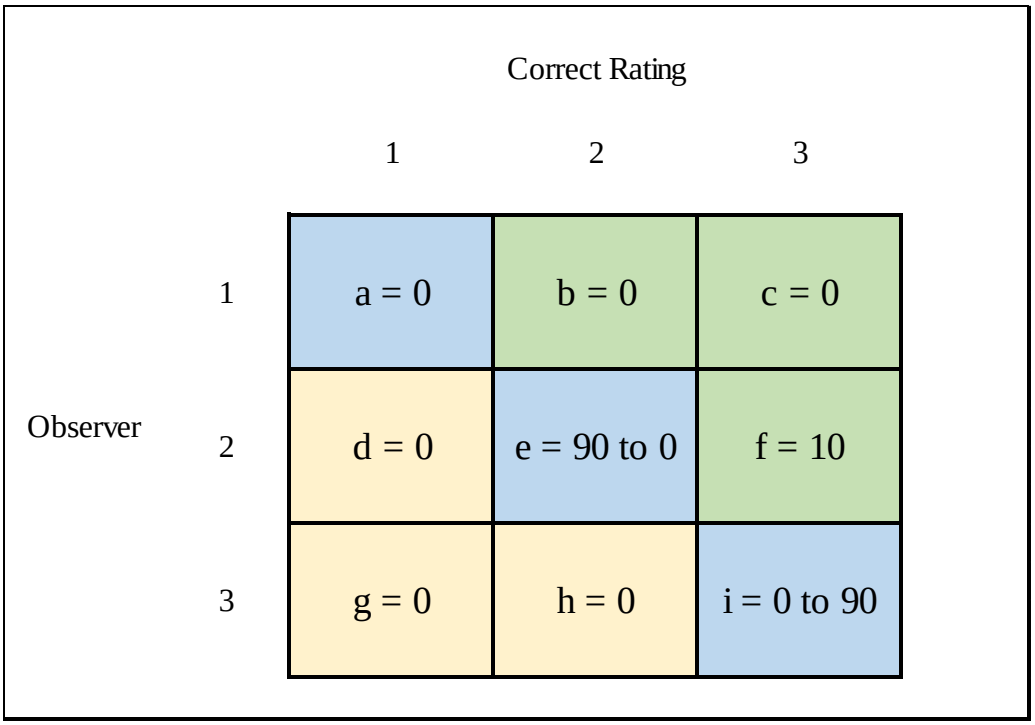


Figure 3.9. 3x3 agreement distribution for data condition 12 from Table 3.2.

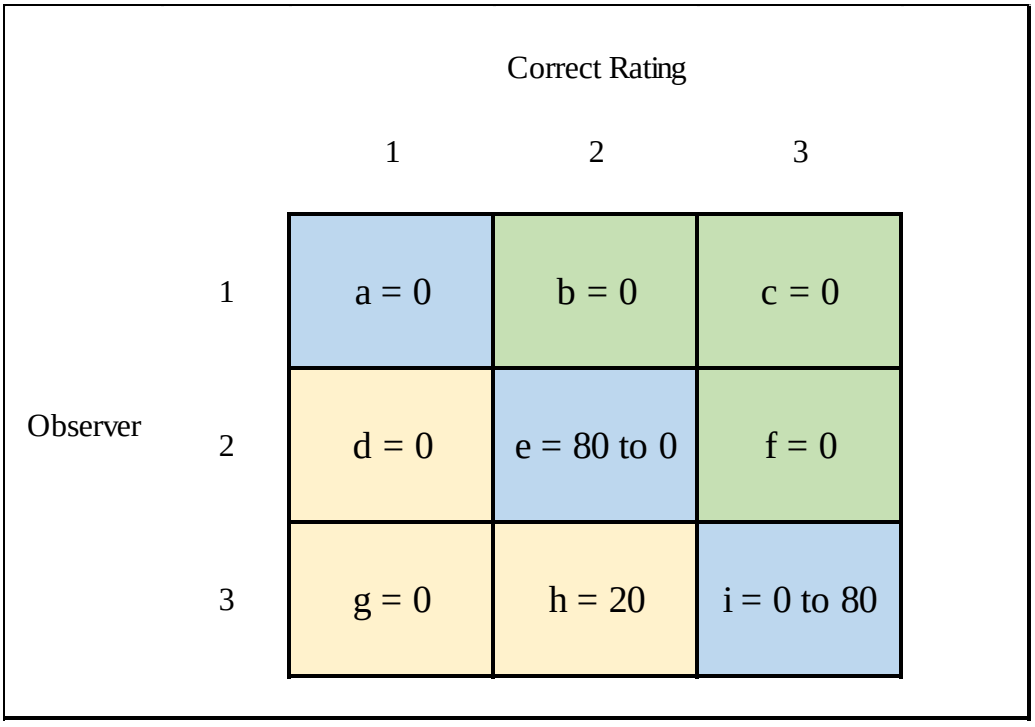


Figure 3.10. 3x3 agreement distribution for data condition 19 from Table 3.2.

The data conditions in rows 21-35 of Table 3.2 varied the agreement across all three agreement cells (*a*, *e*, and *i*). In conditions 21-25 the disagreement was evenly distributed across cells *b* and *f*. Conditions 26-30 moved the disagreement to two lenient cell locations (*d* and *h*). The final group of conditions (31-35) for the three-point rating scale evenly distributed disagreement across cells *b*, *d*, *f*, and *h*. These conditions were selected because they provided a ratings distribution where the rater had high agreement and greater overall category use than in conditions 1-20. Figures 3.11 – 3.13 provide sample agreement distributions for specific data conditions from the 3x3 agreement matrix.

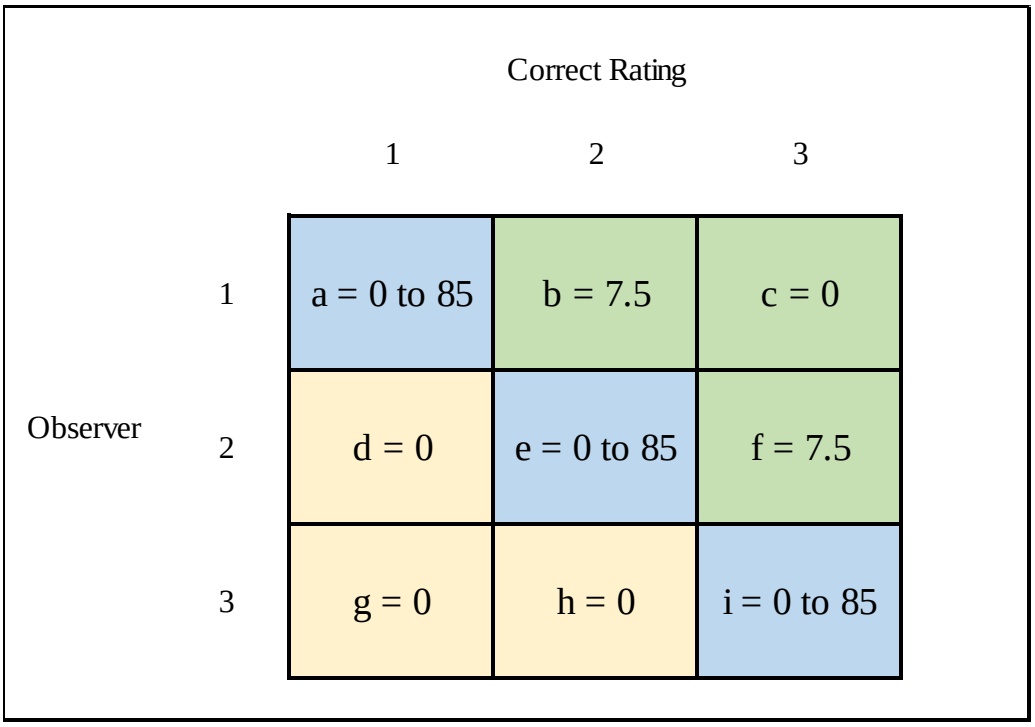


Figure 3.11. 3x3 agreement distribution for data condition 23 from Table 3.2.

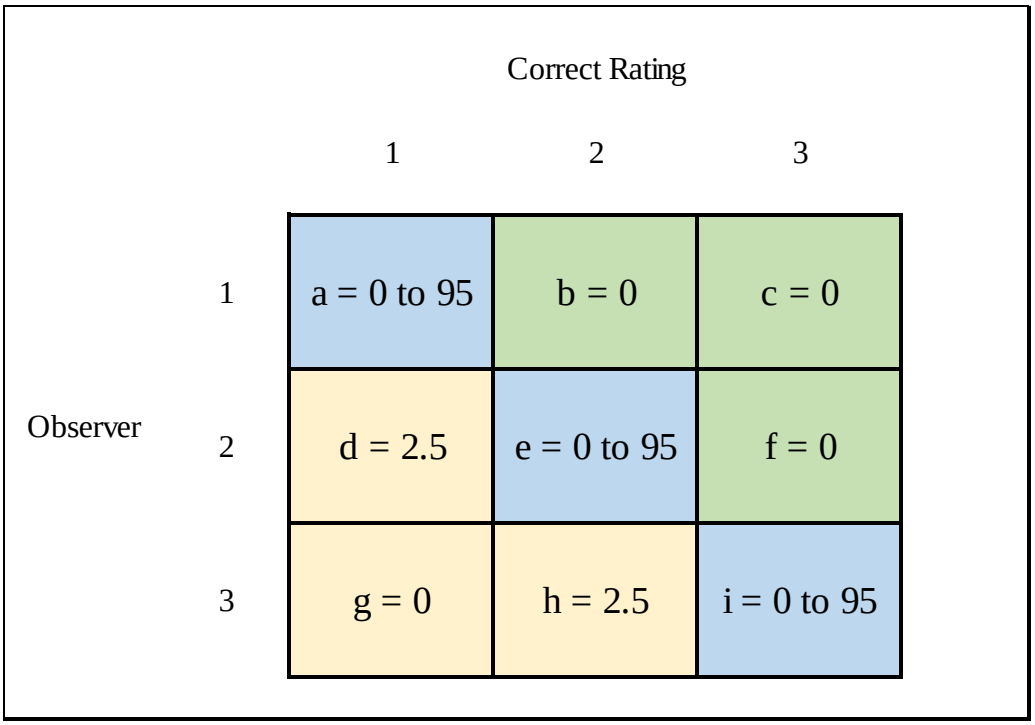


Figure 3.12. 3x3 agreement distribution for data condition 26 from Table 3.2.

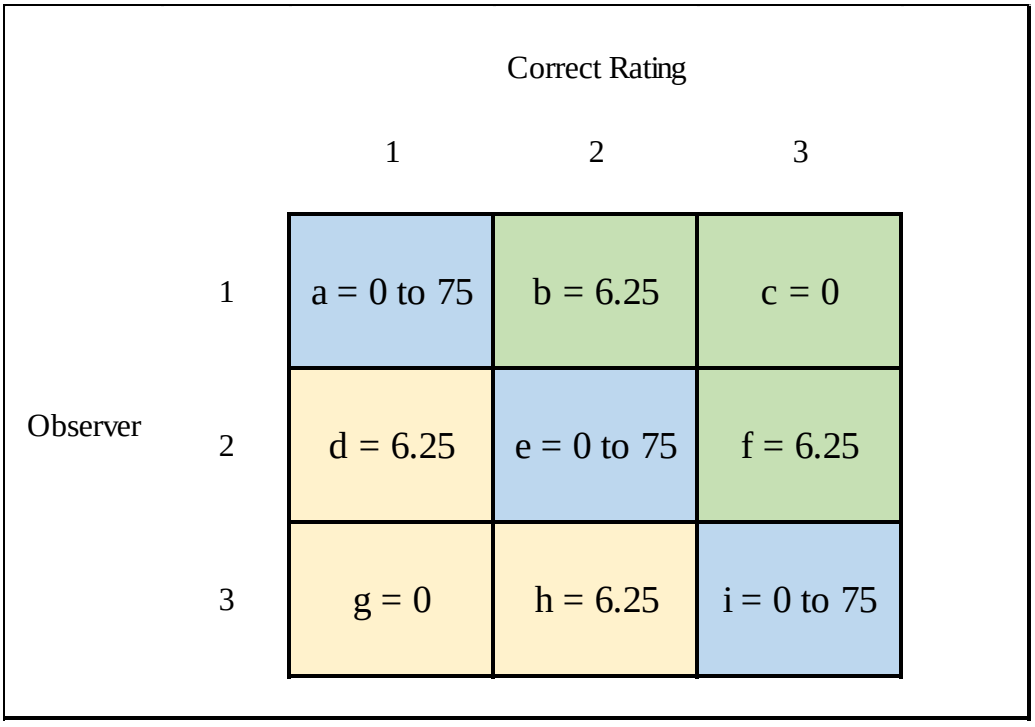


Figure 3.13. 3x3 agreement distribution for data condition 35 from Table 3.2.

4x4 Data Conditions. Table 3.3 lists the simulations and data condition variations according to the 4x4 rating matrix. Figure 3.3 corresponds with the data conditions explained in Table 3.3. The magnitude of agreement was distributed across two or four cells in each condition. The conditions adjusted whether the disagreement occurred in a strict cell, lenient cell, or was balanced across the strict and lenient cells. When multiple strict or lenient cells hold the disagreement, the disagreement was equally distributed across the indicated cells.

Data conditions 1-50, shown in Table 3.3, limited the distribution of agreement to a combination of two cells. Within conditions 1-50 a total of 30 conditions (conditions 1-15, 21-25, 31-35, and 41-45) limited the disagreement to a single strict or lenient cell. These 30 conditions were used to evaluate the performance of the IRR coefficients of interest in situations where the rater had high agreement and low category use, see Figure 3.14 for an example of the ratings distribution for one of these scenarios. The remaining 20 conditions (16-20, 26-30, 36-40, and 46-50) from rows 1-50 in Table 3.3 located the disagreement across either a combination of three strict cells or three lenient cells located one point above or below the correct ratings. These scenarios explored the performance of the coefficients in instances of high agreement and moderate levels of category use. Figure 3.15 displays the ratings distribution for data condition 18 from Table 3.3.

Table 3.3

4x4 data simulation conditions

	Magnitude		Location of Agreement and Disagreement		
	Agreement (%)	Disagreement (%)	Agreement Distribution	Strictness	Leniency
1	95	5	varies from a = 95, p = 0 to a = 0, p = 95	d	-
2	90	10	varies from a = 90, p = 0 to a = 0, p = 90	d	-
3	85	15	varies from a = 85, p = 0 to a = 0, p = 85	d	-
4	80	20	varies from a = 80, p = 0 to a = 0, p = 80	d	-
5	75	25	varies from a = 75, p = 0 to a = 0, p = 75	d	-
6	95	5	varies from a = 95, p = 0 to a = 0, p = 95	-	m
7	90	10	varies from a = 90, p = 0 to a = 0, p = 90	-	m
8	85	15	varies from a = 85, p = 0 to a = 0, p = 85	-	m
9	80	20	varies from a = 80, p = 0 to a = 0, p = 80	-	m
10	75	25	varies from a = 75, p = 0 to a = 0, p = 75	-	m
11	95	5	varies from f = 95, k = 0 to f = 0, k = 95	g	-
12	90	10	varies from f = 90, k = 0 to f = 0, k = 90	g	-
13	85	15	varies from f = 85, k = 0 to f = 0, k = 85	g	-
14	80	20	varies from f = 80, k = 0 to f = 0, k = 80	g	-
15	75	25	varies from f = 75, k = 0 to f = 0, k = 75	g	-
16	95	5	varies from f = 95, k = 0 to f = 0, k = 95	b, g, l	-
17	90	10	varies from f = 90, k = 0 to f = 0, k = 90	b, g, l	-
18	85	15	varies from f = 85, k = 0 to f = 0, k = 85	b, g, l	-
19	80	20	varies from f = 80, k = 0 to f = 0, k = 80	b, g, l	-
20	75	25	varies from f = 75, k = 0 to f = 0, k = 75	b, g, l	-
21	95	5	varies from f = 95, k = 0 to f = 0, k = 95	-	j
22	90	10	varies from f = 90, k = 0 to f = 0, k = 90	-	j
23	85	15	varies from f = 85, k = 0 to f = 0, k = 85	-	j
24	80	20	varies from f = 80, k = 0 to f = 0, k = 80	-	j
25	75	25	varies from f = 75, k = 0 to f = 0, k = 75	-	j
26	95	5	varies from f = 95, k = 0 to f = 0, k = 95	-	e, j, o
27	90	10	varies from f = 90, k = 0 to f = 0, k = 90	-	e, j, o
28	85	15	varies from f = 85, k = 0 to f = 0, k = 85	-	e, j, o
29	80	20	varies from f = 80, k = 0 to f = 0, k = 80	-	e, j, o
30	75	25	varies from f = 75, k = 0 to f = 0, k = 75	-	e, j, o
31	95	5	varies from k = 95, f = 0 to k = 0, f = 95	g	-
32	90	10	varies from k = 90, f = 0 to k = 0, f = 90	g	-
33	85	15	varies from k = 85, f = 0 to k = 0, f = 85	g	-
34	80	20	varies from k = 80, f = 0 to k = 0, f = 80	g	-
35	75	25	varies from k = 75, f = 0 to k = 0, f = 75	g	-
36	95	5	varies from k = 95, f = 0 to k = 0, f = 95	b, g, l	-
37	90	10	varies from k = 90, f = 0 to k = 0, f = 90	b, g, l	-
38	85	15	varies from k = 85, f = 0 to k = 0, f = 85	b, g, l	-
39	80	20	varies from k = 80, f = 0 to k = 0, f = 80	b, g, l	-
40	75	25	varies from k = 75, f = 0 to k = 0, f = 75	b, g, l	-

(continued)

	Magnitude		Location of Agreement and Disagreement		
	Agreement (%)	Disagreement (%)	Agreement Distribution	Strictness	Leniency
41	95	5	varies from k = 95, f = 0 to k = 0, f = 95	-	j
42	90	10	varies from k = 90, f = 0 to k = 0, f = 90	-	j
43	85	15	varies from k = 85, f = 0 to k = 0, f = 85	-	j
44	80	20	varies from k = 80, f = 0 to k = 0, f = 80	-	j
45	75	25	varies from k = 75, f = 0 to k = 0, f = 75	-	j
46	95	5	varies from k = 95, f = 0 to k = 0, f = 95	-	e, j, o
47	90	10	varies from k = 90, f = 0 to k = 0, f = 90	-	e, j, o
48	85	15	varies from k = 85, f = 0 to k = 0, f = 85	-	e, j, o
49	80	20	varies from k = 80, f = 0 to k = 0, f = 80	-	e, j, o
50	75	25	varies from k = 75, f = 0 to k = 0, f = 75	-	e, j, o
51	95	5	varies from a = 47.5, f = 0, k = 47.5, p = 0 to a = 0, f = 47.5, k = 0, p = 47.5	b, c, d, g, h, l	-
52	90	10	varies from a = 45.0, f = 0, k = 45.0, p = 0 to a = 0, f = 45.0, k = 0, p = 45.0	b, c, d, g, h, l	-
53	85	15	varies from a = 42.5, f = 0, k = 42.5, p = 0 to a = 0, f = 42.5, k = 0, p = 42.5	b, c, d, g, h, l	-
54	80	20	varies from a = 40.0, f = 0, k = 40.0, p = 0 to a = 0, f = 40.0, k = 0, p = 40.0	b, c, d, g, h, l	-
55	75	25	varies from a = 37.5, f = 0, k = 37.5, p = 0 to a = 0, f = 37.5, k = 0, p = 37.5	b, c, d, g, h, l	-
56	95	5	varies from a = 47.5, f = 0, k = 47.5, p = 0 to a = 0, f = 47.5, k = 0, p = 47.5	-	e, i, j, m, n, o
57	90	10	varies from a = 45.0, f = 0, k = 45.0, p = 0 to a = 0, f = 45.0, k = 0, p = 45.0	-	e, i, j, m, n, o
58	85	15	varies from a = 42.5, f = 0, k = 42.5, p = 0 to a = 0, f = 42.5, k = 0, p = 42.5	-	e, i, j, m, n, o
59	80	20	varies from a = 40.0, f = 0, k = 40.0, p = 0 to a = 0, f = 40.0, k = 0, p = 40.0	-	e, i, j, m, n, o
60	75	25	varies from a = 37.5, f = 0, k = 37.5, p = 0 to a = 0, f = 37.5, k = 0, p = 37.5	-	e, i, j, m, n, o
61	95	5	varies from a = 47.5, f = 0, k = 47.5, p = 0 to a = 0, f = 47.5, k = 0, p = 47.5	b, c, d, g, h, l	e, i, j, m, n, o
62	90	10	varies from a = 45.0, f = 0, k = 45.0, p = 0 to a = 0, f = 45.0, k = 0, p = 45.0	b, c, d, g, h, l	e, i, j, m, n, o
63	85	15	varies from a = 42.5, f = 0, k = 42.5, p = 0 to a = 0, f = 42.5, k = 0, p = 42.5	b, c, d, g, h, l	e, i, j, m, n, o
64	80	20	varies from a = 40.0, f = 0, k = 40.0, p = 0 to a = 0, f = 40.0, k = 0, p = 40.0	b, c, d, g, h, l	e, i, j, m, n, o
65	75	25	varies from a = 37.5, f = 0, k = 37.5, p = 0 to a = 0, f = 37.5, k = 0, p = 37.5	b, c, d, g, h, l	e, i, j, m, n, o

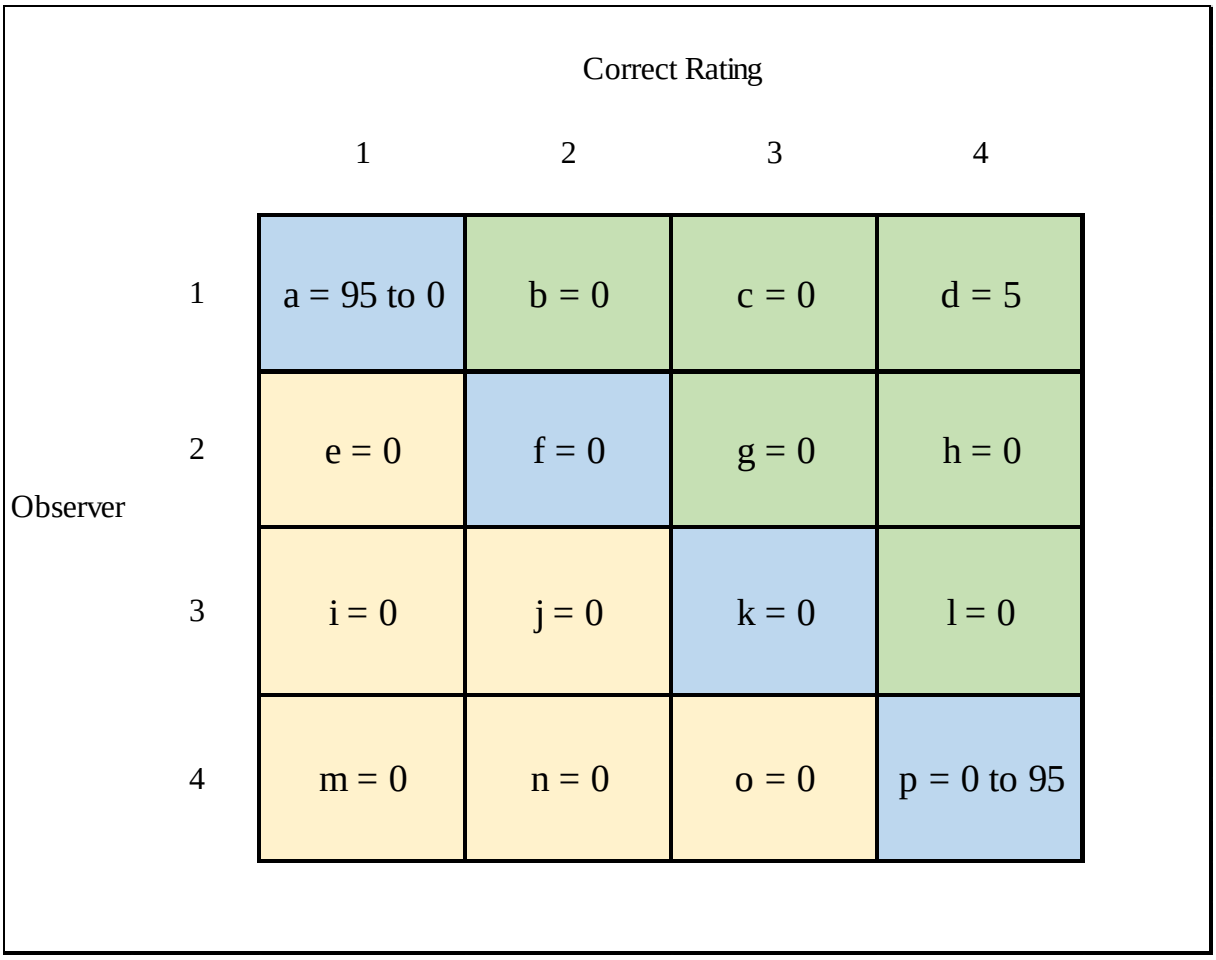


Figure 3.14. 4x4 agreement distribution for data condition 1 from Table 3.3.

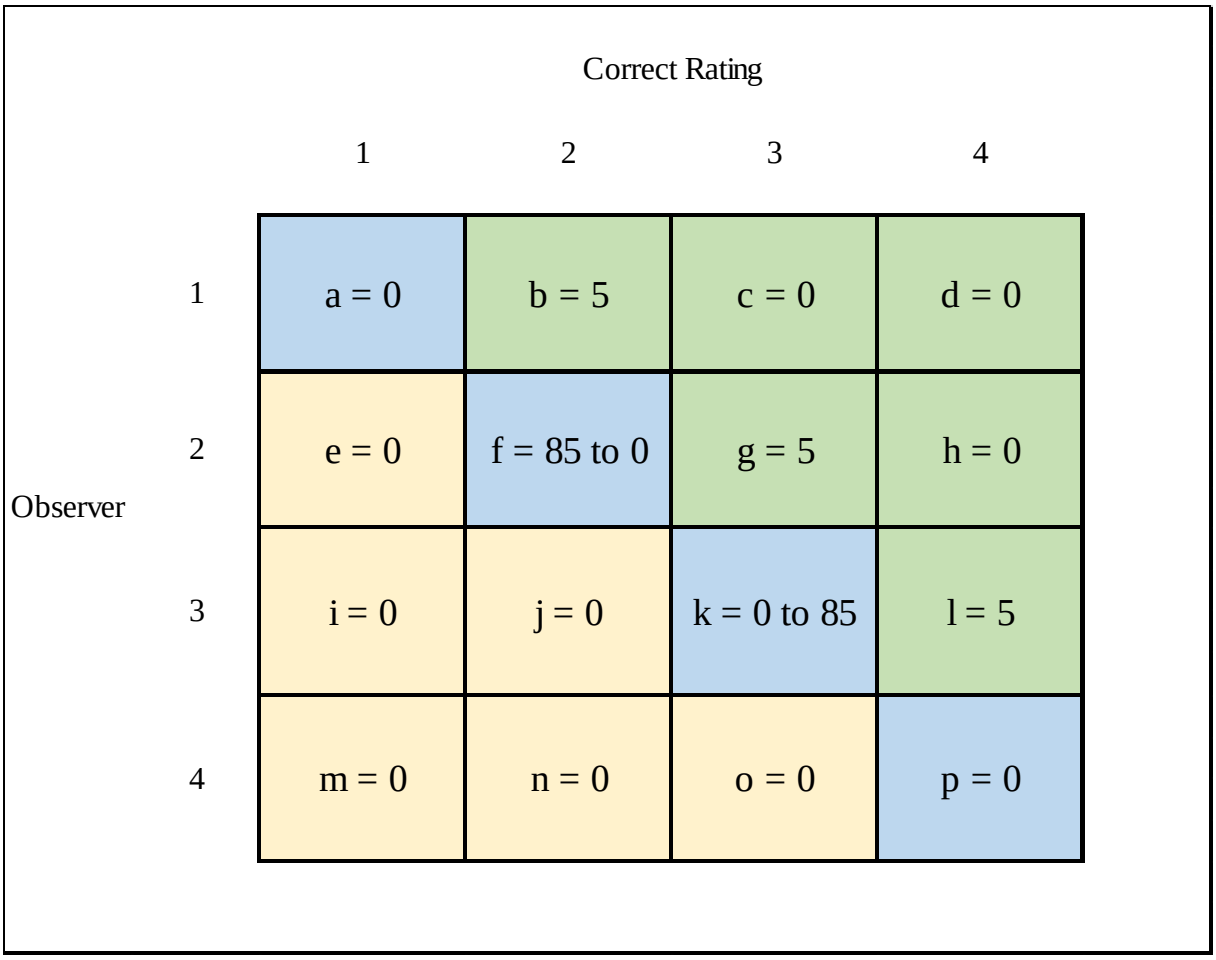


Figure 3.15. 4x4 agreement distribution for data condition 18 from Table 3.3.

The final 15 conditions (51-65) in Table 3.3 allowed for the proportion of agreement to be allocated across all four possible agreement cells. Conditions 51-55 evenly distributed the proportion of disagreement across all six strict cells, see Figure 3.16 for a distribution from this group of conditions. Data conditions 56-60 placed disagreement evenly across all six lenient cells, see Figure 3.17 for an example distribution. Finally, conditions 61-65 evenly distributed disagreement across all 12 discrepant (strict and lenient) cell locations. Figure 3.18 provides the proportion distribution for data condition 65.

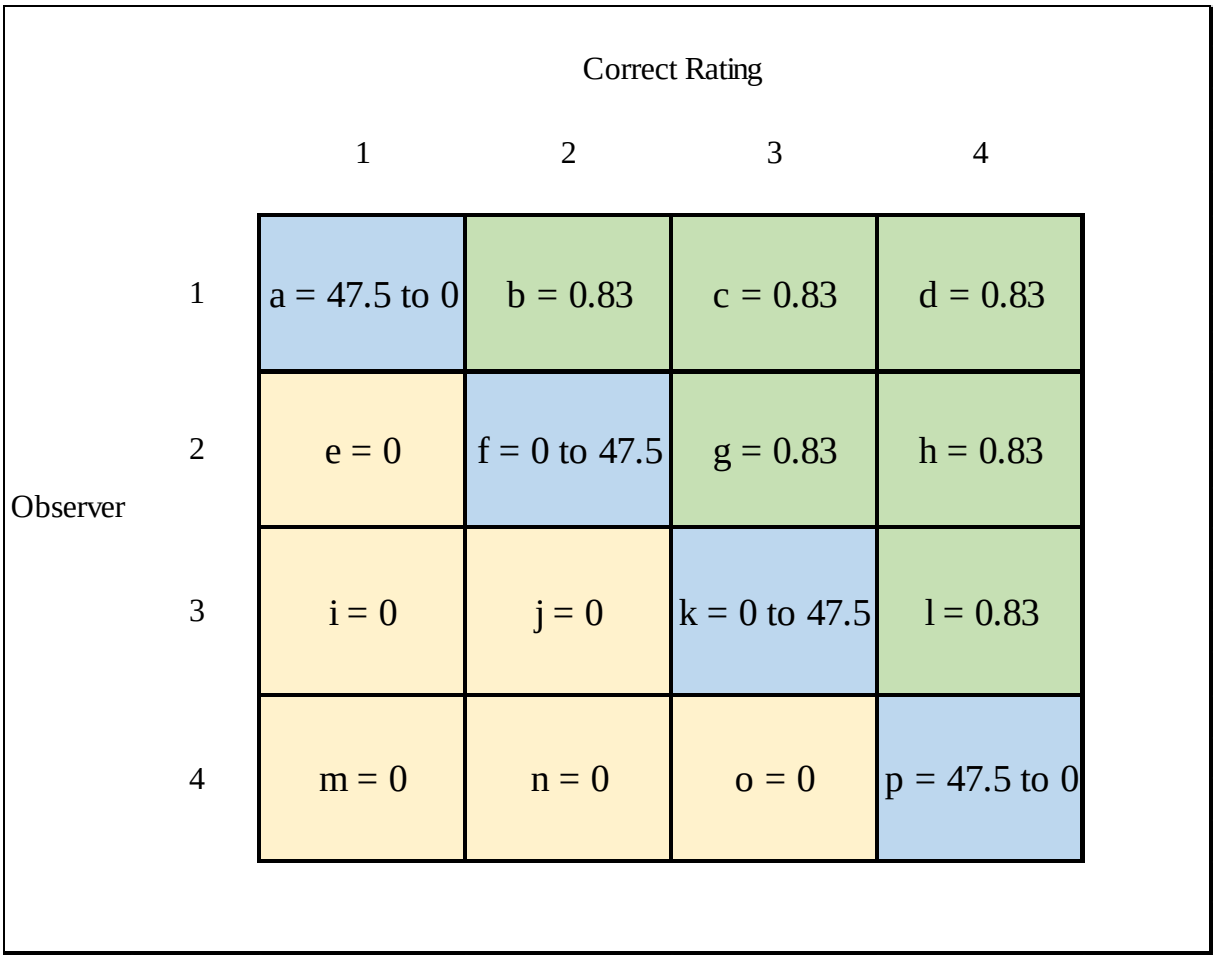


Figure 3.16. 4x4 agreement distribution for data condition 51 from Table 3.3.

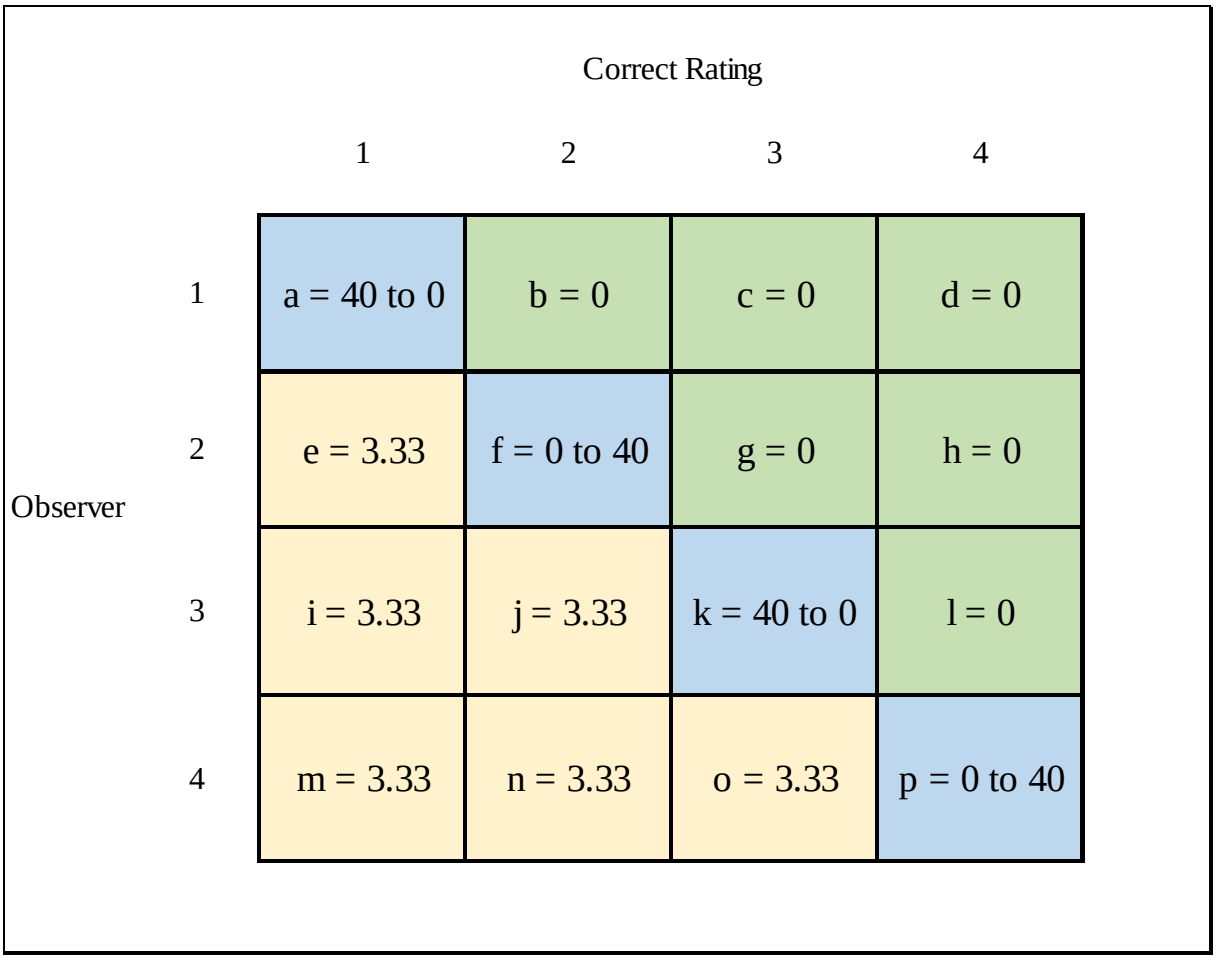


Figure 3.17. 4x4 agreement distribution for data condition 59 from Table 3.3.

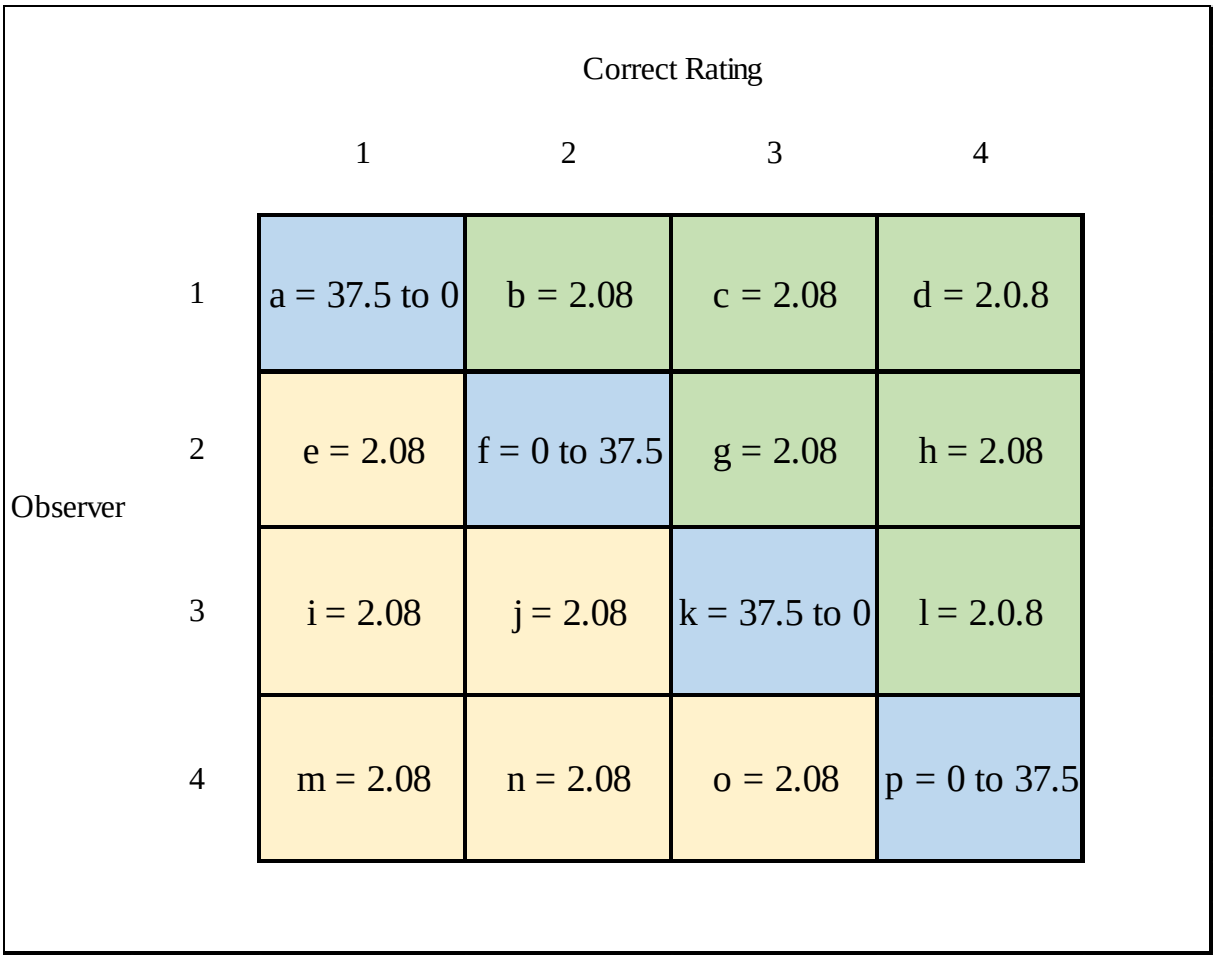


Figure 3.18. 4x4 agreement distribution for data condition 65 from Table 3.3.

Summary

The conditions explained across the 2x2, 3x3, and 4x4 matrices were selected to systematically provide a sample of ratings that have demonstrated issues in the usefulness of IRR coefficients. Certain data conditions were selected to provide information about the performance of Lambda-1 and Lambda-2 that are not as problematic for certain coefficients applied in this study. Again, the adjustments were not intended to be exhaustive of all possible ratings combinations. These adjustments allowed for the calculation of row and column marginals that

are a required calculation for parts of the expected chance agreement (p_e) portion in the calculations of Lambda-1, Lambda-2, AC1, and Kappa.

Chance-Corrected Agreement Coefficients

The IRR coefficients selected for this study were calculated using data generated from the conditions listed in Tables 3.1 – 3.3. The primary coefficients of interest in this study were Lambda-1 and Lambda-2. Evaluation of the performance of these coefficients under certain conditions was part of the process in determining the capacity these coefficients have for being used along with, or in place of, other IRR coefficients. Lambda-1 and Lambda-2 were calculated and evaluated along with Cohen's Kappa, *S*, Krippendorff's Alpha, and Gwet's AC1. An explanation of the calculation of each of these coefficients is located in Chapter 2. The calculation of each of these coefficients was predicated on the rating distributions generated for each data condition listed in Tables 3.1 – 3.3. These distributions were used to calculate the value of each coefficient across all scenarios.

Cohen's Kappa is the most widely used and cited chance-corrected agreement coefficient (Xie, 2013). Despite well-known issues with the Kappa coefficient, it remains popular. There are two well-known issues with Kappa that led to the development of alternative chance-corrected agreement coefficients, like Krippendorff's Alpha and Gwet's AC1. Krippendorff's Alpha is often compared to Kappa and percent agreement (Stevens et al., 2014). It was included in the current study since the calculation of Krippendorff's Alpha can be adjusted to have application in any scale of measurement. Lambda was designed for ordinal rating scales, while Krippendorff's Alpha was not specifically designed for ordinal data alone, it was the only other measure included in this study with consideration for this scale of measurement in its design. Since 2001, Gwet's AC1 coefficient has been frequently used as an alternative IRR coefficient due to well-

known issues with the Kappa coefficient (Wongpakaran et al., 2013, Xie, 2013). *S* only considers the number of categories and the overall percentage of agreement between raters. *S* was included in the current study because it has demonstrated that it is a more consistent and easily interpretable coefficient. The coefficient calculations for Krippendorff's Alpha and the *S* coefficient are located in Appendix A. A full illustration of the calculations of chance correct agreement (p_e) using Kappa, Gwet's AC1, Lambda-1, and Lambda-2 are located in Appendix B.

Calculating Lambda-2

The difference in calculations for Lambda-2 involve accounting for the probability of a rater's starting point on a given rating scale. The probability of using a given category in these examples can be adjusted to match the scenario in which Lambda-2 is calculated, which is the population proportion (π). The general form of Lambda-2 uses the probability of use of each category based on population values when available. For demonstration purposes, the category starting point used in the examples in this chapter are based on category use from the population values from teacher evaluations in North Carolina, which uses a four-point scale teacher evaluation rubric. The values used to calculate Lambda-2 in the 4x4 matrix set the probability of selecting category 1 (π_1) at 0.05, category 2 (π_2) at 0.65, category 3 (π_3) at 0.25, and category 4 (π_4) at 0.05. These values reflect what is found in practice when evaluators of teacher performance used a four-point scale. For the three-point and two-point data conditions, the categories were combined to calculate the category starting point probability. The 3x3 matrix adjusted the π_2 to combine the original middle two ratings, making π_2 equivalent to 0.90. In the calculation of Lambda-2 in a three-point scale both π_1 and π_3 were each set at 0.05. For the 2x2 matrix, π_1 remained at 0.05 and π_2 is 0.95.

Evaluation Criteria

In order to address the evaluation criteria for the current study it is important to revisit the research questions:

1. Under what conditions do Cohen's Kappa, *S*, Krippendorff's Alpha, Gwet's AC1, Lambda-1, and Lambda-2 produce similar values?
2. To what extent are the IRR coefficients, calculated using Cohen's Kappa, *S*, Krippendorff's Alpha, Gwet's AC1, Lambda-1, and Lambda-2, placed in the same classification categories according to well-known taxonomies (i.e., almost perfect, substantial, moderate, fair, slight, and no agreement) under the same levels of agreement?
3. To what extent do Cohen's Kappa, *S*, Krippendorff's Alpha, Gwet's AC1, Lambda-1, and Lambda-2 measure agreement in similar ways?

Evaluating Research Question 1

The first research question was addressed by comparing the mean values of each coefficient within each generated data condition following guidelines suggested by Raadt et al. (2021). The difference between reliability coefficients compared in Raadt and his colleague's study considered differences between mean coefficient values ≤ 0.10 as "small" differences (Raadt et al., 2021). To provide a frame of reference, data condition 1 using the 2x2 matrix (Table 3.1) was generated and mean values for each coefficient are shown in Table 3.4 to provide an example of how results will be displayed and evaluated. Under data condition 1 from the 2x2 agreement matrix, only Kappa and AC1 had a difference falling outside of the recommended range, differing by 0.12. All other possible pairs of coefficients were within the acceptable criteria value (differences ≤ .10). The ratio column in Table 3.4 identifies the number

of coefficients with differences outside the criteria across all possible pairwise comparisons of agreement coefficients, which is 1 out of 15 in data condition 1.

Table 3.4
Sample of 2x2 Data Generated for Research Question 1, Evaluation Criteria

Data Condition	n	Lambda-1	Lambda-2	Kappa	*S*	AC1	Kripp.'s Alpha	Ratio
1	96	0.90	0.85	0.80	0.90	0.92	0.85	1
2								
3								
4								
…								

Note. Ratio refers to the number of pairs of coefficients falling outside of the acceptable range of ≤ 0.10 out of a possible 15 coefficient pairs per data condition. n = combinations of ratings generated under current data condition.

Evaluating Research Question 2

The second research question was addressed by using Landis & Koch's (1977) recommended benchmark agreement level descriptions. Landis & Koch (1977) referred to IRR values of < 0.00 as having "no agreement", 0.00 – 0.20 as "slight", 0.21 – 0.40 as "fair", 0.41 – 0.60 as "moderate", 0.61 – 0.80 as "substantial", and 0.81 to 1.00 as "almost perfect" agreement. The count and percentage of coefficient values within benchmark agreement level generated were produced and interpreted.

Table 3.5
Sample Benchmark Count for Data Condition 1 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	<0.00	0.00 – 0.20	0.21 – 0.40	0.41 – 0.60	0.61 – 0.80	0.81 – 1.00
Lambda-1	0	0	0	0	0	96
Lambda-2	0	0	0	4	16	76
Kappa	0	2	2	6	16	70
S	0	0	0	0	0	96
AC1	0	0	0	0	0	96
Kripp. Alpha	0	0	0	4	14	78

Note. Using Landis & Koch (1977) benchmark categories.

Using data condition 1 from the 2x2 agreement matrix as an example, the majority of coefficient values were in the "almost perfect" agreement category. Kappa produced values in the "almost perfect" range 72.9% of the time in data condition 1 from the 2x2 agreement matrix,

while 100% of Lambda-1, *S*, and AC1 values fell in the highest benchmark category. Results from 2x2 agreement matrix data condition 1 are displayed as counts in Table 3.5 and percentages in Table 3.6. This analysis allowed for interpretations about classification consistency for the same data conditions across the six coefficients analyzed in this study.

Table 3.6

Sample Benchmark Percentages for Data Condition 1 (2x2)

	No Agreement <0.00	Slight 0.00 – 0.20	Fair 0.21 – 0.40	Moderate 0.41 – 0.60	Substantial 0.61 – 0.80	Almost Perfect 0.81 – 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	4.2%	16.7%	79.2%
Kappa	0.0%	2.1%	2.1%	6.3%	16.7%	72.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	4.2%	14.6%	81.3%

Note. Using Landis & Koch (1977) benchmark categories.

Evaluating Research Question 3

The third research question was addressed by calculating Pearson correlation coefficients to assess how similarly coefficients were measuring agreement. The stability and consistency of the rank order of each generated value provided insight into the degree to which each coefficient was measuring agreement in similar ways. The *S* coefficient is a constant and did not have any variance within each data condition. So, *S* was not included in evaluating the third research question. Correlation tables were generated and analyzed for each data condition. This information provided insight into how, and whether, Lambda-1 and Lambda-2 measured agreement similar to each other and the other coefficients under varied proportional distributions of agreement and disagreement.

Results from 2x2 agreement matrix data condition 1 are displayed in Figure 3.19. This is an example of how correlation matrices may be displayed. The darker red text identifies high, negative correlations. While the darker blue text identifies high, positive correlations. The lighter

the text appears, the lower the correlation. The near invisible or blank sections of the correlation matrix are insignificant correlation values. This analysis and visual display allowed for interpretations about what extent of conclusions about IRR would be made for the same data condition across coefficients.

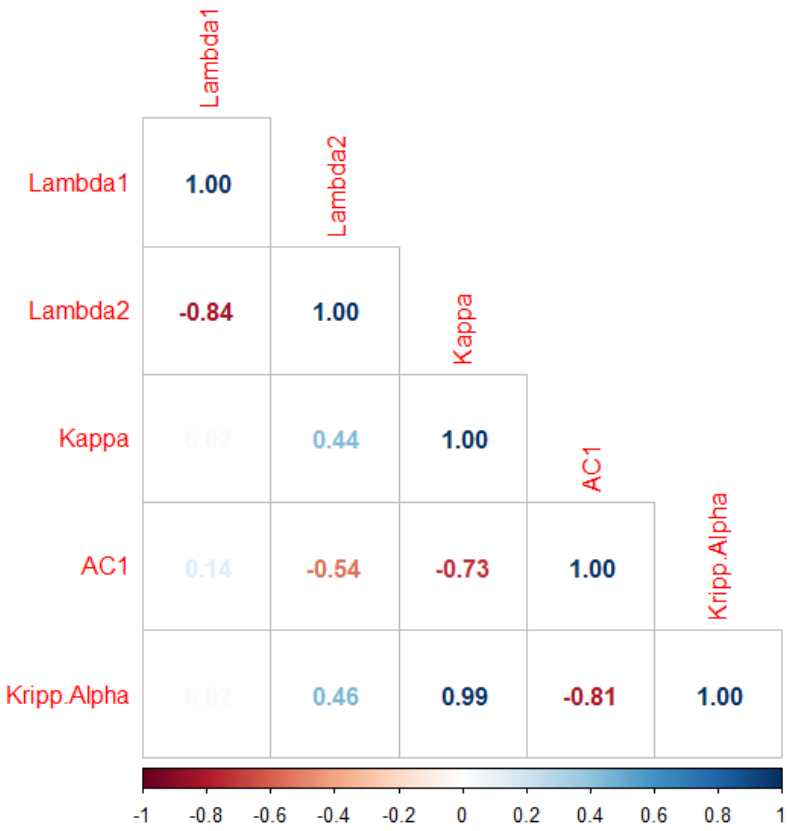


Figure 3.19. Sample correlation matrix for data condition 1 (2x2).

Expected Outcomes

While it was not possible to exactly predict how all coefficients performed relative to the prespecified conditions, it is an important component of a simulation study to hypothesize findings. In the initial application of Lambda-1 and Lambda-2, these coefficients were calculated using actual teacher evaluation data and simulated conditions (Lambert et al., 2021). Both the actual teacher evaluation data and simulated conditions utilized four-point rating scales in this

study. Across both types of data coefficient values for Lambda-2 closely matched the values for Kappa. Lambda-1 did not overcorrect for chance-agreement as much as Kappa or Lambda-2. In most situations Lambda-1 produced slightly higher and more consistent coefficient values. Lambda-1 closely matched the performance of AC1 using empirical teacher evaluation data. AC1 was not utilized in the simulated conditions. Ideally both coefficients will produce more consistent results that are not as dependent upon the amount and location of agreement and disagreement as the Kappa coefficient.

Based on findings from the initial application of Lambda-1 and Lambda-2, it is expected that the two coefficients will perform similar across new data conditions. Given the formula for calculating Lambda-1, it most likely will remain consistent and stable as the rating scale size changes. Whereas Lambda-2 varies the weights assigned to each category according to population proportion values. So, Lambda-2 will most likely have more unpredictable behavior as certain data conditions are adjusted, specifically the size of the rating scale. If differences are found between coefficient performances, it would indicate that certain coefficients are more stable and are consistently measuring agreement within a data condition.

Summary

The methodology of this study aimed to evaluate the functioning of a new statistical method. Lambda-1 and Lambda-2 are rater-mediated chance-corrected agreement coefficients. To understand the performance of Lambda-1 and Lambda-2 in comparison to common IRR coefficients, adjustments were made to the amount and location of agreement and disagreement across 2x2, 3x3, and 4x4 agreement according to the aforementioned simulation design. Varying these conditions resulted in adjustments to the marginal distribution of ratings and expected chance agreement calculations for the IRR coefficients. From this data generation mechanism,

the values for Lambda-1, Lambda-2, Kappa, *S*, Krippendorff's Alpha, and Gwet's AC1 were calculated and used to answer the three research questions. In addition to the evaluation criteria listed above, depictions of graphical indicators will be generated for each data condition. The graphs allowed for the visual inspection and comparison of each coefficient's performance under the same fixed conditions. Investigating results from this study helps inform researchers about the performance of selected chance-corrected agreement coefficients under various conditions.

CHAPTER IV: RESULTS

Analysis was performed on results from 15 data conditions for a two-point rating scale, 35 data conditions for the three-point scale, and 65 data conditions for the four-point scale. The data conditions were systematically specified to reflect agreement figures commonly found in actual ratings applications. Also, the selected conditions have shown to be problematic for chance-corrected agreement coefficients. The performance of Lambda-1 and Lambda-2 in comparison to the performance of Kappa, *S,* AC1, and Krippendorff's Alpha under the same condition was analyzed according to evaluation criteria explained in the previous chapter. In the current study, population proportion of ratings values from NCTEP data were applied to calculate Lambda-2. Findings from Xie's (2013) study comparing AC1, S, and Kappa, were replicated to simulate exact same conditions and findings. This served as an accuracy check for the data generation process, specified conditions, and coefficient calculations. Figure 4.1 illustrates the results of data condition 2 (2x2) which matches Figure 10 in Xie (2013).

Data for each of the various 115 data conditions were generated in R using the data.table package (Dowle, 2021). Analysis was performed using the tidyverse (Wickham, 2021), dplyr (Wickham, 2022), kim (Kim, 2021), ggplot (Wickham et al, 2021), and corrplot (Wei & Simko, 2021) R packages. The R script for generating and analyzing data conditions for each of the 2x2, 3x3, and 4x4 agreement matrices can be found in Appendix C. A total of 20 out of the 65 planned data conditions for the 4x4 matrices produced repeated agreement distributions. Data conditions 31 – 40 produced the same agreement distribution as data conditions 11 – 20 for the 4x4 agreement matrix. Data conditions 41 – 50 produced the same agreement distribution as data conditions 21 – 30 for the 4x4 agreement matrix. Since distributions for these conditions were

identical, data conditions 31 – 50 were removed from further analysis. In total, a final group of 95 data conditions were analyzed. Results are organized and explained according to research question and each individual rating scale size.

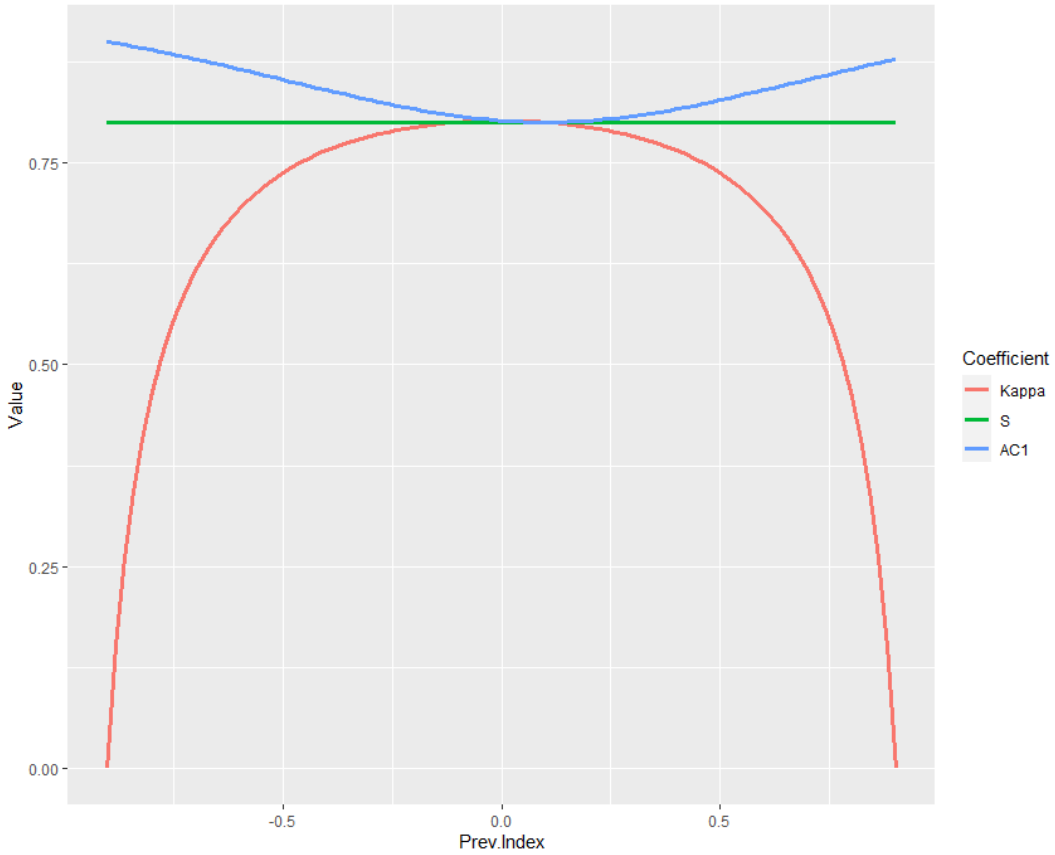


Figure 4.1. Replication of Figure 10 from Xie (2013).

Results: Research Question 1

Similarity of Coefficient Values

The first research question identified the conditions where the coefficients produced similar values. The mean values from each generated condition for each coefficient were compared. The count of the number of pairs outside of the range were tabulated for each combination of coefficients. Differences in values less than or equal to .10 were considered similar when comparing reliability values produced by each data condition (Raadt et al., 2021).

The aim of this research question was to provide evidence the coefficients, specifically Lambda-1 and Lambda-2, were producing values similar to other coefficients under the same data conditions. Results are organized and explained according to rating scale size. The issue of variability in overall coefficient values produced was addressed by the second research question.

2x2 Coefficient Comparisons

The higher coefficient values were produced where the data conditions had the highest specified values of agreement. Data conditions 1, 6, and 11 set agreement at 95% and resulted in 96 different rating combinations each across cells *a*-*d*. These three data conditions had the highest values for each coefficient and produced the most similar results according to the criteria value (differences ≤ .10). Each of these three conditions had one pair of differences between coefficients fall outside the range, in all three cases this occurred between the values for Kappa and AC1. Data conditions 3 – 5 produced the highest number of differences between pairs of coefficients. Data conditions 3 and 5 produced eight pairs of mean values outside of the .10 cutoff and data condition 4 produced nine pairs outside of the criteria. Full results of mean coefficient values across the 15 2x2 agreement matrices data conditions are provided in Table 4.1. For further explanation about individual 2x2 data conditions see Table 3.1.

For the 2x2 data conditions, the mean value of Lambda-1 was similar to *S*, AC1, and Krippendorff's Alpha across all 15 data conditions. Lambda-1 differed from Lambda-2 across data conditions 3 – 5. Lambda-1 differed from the Kappa coefficient in 12 out of 15 scenarios. The three conditions where Lambda-1 and Kappa met the criteria, the difference between the two coefficients was at the upper limit of the acceptable difference range (at exactly .10). Lambda-2 produced mean values most similar to Lambda-1 and *S*. Lambda-2 fell outside of the range with those two coefficients in the results from data conditions 3 – 5. There were five situations where

Lambda-2 was outside of the acceptable criteria when compared to Kappa and Krippendorff's Alpha. Lambda-2 had the highest number of differences with the mean values for AC1, occurring in seven out of 15 possible occasions.

Table 4.1
2x2 Mean Coefficient Values

Data Condition	n	Lambda-1	Lambda-2	Kappa	*S*	AC1	Kripp.'s Alpha	Ratio
1	96	0.90	0.85	0.80	0.90	0.92	0.85	1
2	91	0.80	0.70	0.67	0.80	0.84	0.76	3
3	86	0.71	0.53	0.57	0.70	0.75	0.69	8
4	81	0.61	0.36	0.48	0.60	0.67	0.64	9
5	76	0.52	0.17	0.40	0.50	0.58	0.59	8
6	96	0.90	0.88	0.80	0.90	0.92	0.85	1
7	91	0.80	0.77	0.67	0.80	0.84	0.76	3
8	86	0.71	0.69	0.57	0.70	0.75	0.69	5
9	81	0.61	0.61	0.48	0.60	0.67	0.64	5
10	76	0.52	0.54	0.40	0.50	0.58	0.59	4
11	96	0.90	0.87	0.80	0.90	0.92	0.85	1
12	91	0.80	0.74	0.66	0.80	0.84	0.76	4
13	86	0.70	0.63	0.55	0.70	0.75	0.69	4
14	81	0.60	0.52	0.44	0.60	0.66	0.64	6
15	76	0.50	0.42	0.34	0.50	0.57	0.59	6

Note. Ratio refers to the number of pairs of coefficients falling outside of the acceptable range of ≤ 0.10 out of a possible 15 coefficient pairs per data condition. n = combinations of ratings generated under current data condition.

Overall, Kappa accounted for the highest number of differences between pairs of coefficients. Kappa fell out of acceptable range with AC1 across all 15 data conditions, 12 out of 15 conditions when compared to Lambda-1, nine out of 15 conditions when compared to both *S* and Krippendorff's Alpha, and five out of 15 conditions in comparison to Lambda-2. All possible pairs of Lambda-1, *S*, AC1, and Krippendorff's Alpha were within the criteria across all 15 data conditions. Table 4.2 provides an overview of the counts of pairs outside the acceptable range of ≤ .10.

Table 4.2

Out of Range Values Between Coefficient Pairs (2x2)

	1	2	3	4	5	6
1 - Lambda-1	-	3	12	0	0	0
2 - Lambda-2	3	-	5	3	7	5
3 - Kappa	12	5	-	9	15	9
4 - *S*	0	3	9	-	0	0
5 - AC1	0	7	15	0	-	0
6 - Kripp.'s Alpha	0	5	9	0	0	-

3x3 Coefficient Comparisons

Similar to results from the 2x2 agreement matrices, higher coefficient values were produced where the data conditions had the highest specified values of agreement. Data conditions 21 – 35 produced the greatest number of possible rating combinations since the agreement was set to vary across all three agreement cells (*a*, *e*, and *i*) in these scenarios. The total number of ratings provided under these conditions ranged from 2,926 to 4,656 across agreement percentages of 75% to 95%. Whereas the number of ratings provided in instances where agreement was distributed across two categories ranged from 76 to 96 ratings. The substantial increase in ratings led to less variation in the mean coefficient values. For data conditions 21 – 23, 26 – 28, and 31 – 33 all mean coefficient values were within the acceptable range of differences ≤ .10. These data conditions had the highest level of agreement, ranging from 85% to 95% agreement. Data conditions 24 – 25, 29 – 30, and 34 – 35 resulted in just one difference between coefficients per condition. All of these differences were between the Kappa and AC1 coefficients. Data conditions 24 – 25, 29 – 30, and 34 – 35 were had high moderate levels of agreement specifications, 75% and 80%. For further explanation about individual 3x3 data conditions see Table 3.2.

Data conditions 1 – 20 varied agreement levels across two out of three agreement cells. These conditions with a limited number of rating combinations produced a greater amount of

variation in the mean coefficient values under each data condition as compared to data conditions 21 – 35. Adjusting the amount of agreement and disagreement in data conditions 1 – 20 had two primary results. The first finding was the lower agreement levels produced lower coefficient values in each cluster of five data conditions. For example, data conditions 1 – 5 make up a single cluster because they follow the same pattern for agreement and disagreement locations. Also, the conditions with a higher specified agreement level resulted in less differences according to the qualifying criteria (differences ≤ .10) than conditions with lower agreement levels. Full results of mean coefficient values across the 35 3x3 agreement matrices data conditions are provided in Table 4.3.

For the 3x3 data conditions, the mean value of Lambda-1 was similar to *S*, AC1, and Krippendorff's Alpha across all 35 data conditions. Lambda-1 differed from Lambda-2 across six out of 35 data conditions. Lambda-1 differed from the Kappa coefficient in 20 out of 35 scenarios. Lambda-1 and Kappa met the criteria when the number of ratings combinations varied across all three agreement cells (data conditions 21 – 35). Overall, Lambda-2 produced mean values most similar to Krippendorff's Alpha in the 3x3 data conditions in this study. Lambda-2 fell outside of the range with this coefficient in the results from data conditions 5 and 20 (both conditions with 75% agreement). There were six situations where Lambda-2 was outside of the acceptable criteria when compared to Lambda-1 and *S*. Lambda-2 had the highest number of differences with the mean values for Kappa and AC1, occurring in eight out of 35 possible occasions.

Table 4.3

3x3 Mean Coefficient Values

Data Condition	n	Lambda-1	Lambda-2	Kappa	*S*	AC1	Kripp.'s Alpha	Ratio
1	96	0.92	0.87	0.80	0.93	0.94	0.85	3
2	91	0.85	0.76	0.67	0.85	0.89	0.76	5
3	86	0.77	0.65	0.57	0.78	0.83	0.69	8
4	81	0.69	0.54	0.48	0.70	0.77	0.64	7
5	76	0.61	0.44	0.40	0.63	0.71	0.59	8
6	96	0.93	0.89	0.80	0.93	0.94	0.85	3
7	91	0.85	0.81	0.67	0.85	0.89	0.76	5
8	86	0.78	0.75	0.57	0.78	0.83	0.69	6
9	81	0.72	0.69	0.48	0.70	0.77	0.64	6
10	76	0.65	0.64	0.40	0.63	0.71	0.59	6
11	96	0.93	0.89	0.80	0.93	0.94	0.85	3
12	91	0.85	0.81	0.67	0.85	0.88	0.76	5
13	86	0.78	0.75	0.57	0.78	0.82	0.69	6
14	81	0.72	0.69	0.48	0.70	0.76	0.64	6
15	76	0.65	0.64	0.40	0.63	0.70	0.59	6
16	96	0.93	0.87	0.80	0.93	0.94	0.95	4
17	91	0.85	0.76	0.67	0.85	0.88	0.76	5
18	86	0.78	0.65	0.57	0.78	0.82	0.69	8
19	81	0.70	0.55	0.48	0.70	0.76	0.64	8
20	76	0.63	0.45	0.40	0.63	0.70	0.59	9
21	4656	0.93	0.92	0.89	0.93	0.94	0.90	0
22	4186	0.85	0.84	0.80	0.85	0.87	0.82	0
23	3741	0.78	0.76	0.71	0.78	0.81	0.76	0
24	3321	0.70	0.68	0.63	0.70	0.74	0.70	1
25	2926	0.63	0.61	0.55	0.63	0.68	0.65	1
26	4656	0.93	0.92	0.89	0.93	0.94	0.90	0
27	4186	0.85	0.84	0.80	0.85	0.87	0.82	0
28	3741	0.78	0.76	0.71	0.78	0.81	0.76	0
29	3321	0.71	0.68	0.63	0.70	0.75	0.70	1
30	2926	0.64	0.61	0.55	0.63	0.68	0.65	1
31	4656	0.93	0.92	0.89	0.93	0.94	0.90	0
32	4186	0.85	0.84	0.80	0.85	0.87	0.82	0
33	3741	0.78	0.76	0.71	0.78	0.81	0.76	0
34	3321	0.70	0.68	0.63	0.70	0.74	0.70	1
35	2926	0.63	0.60	0.55	0.63	0.68	0.65	1

Note. Ratio refers to the number of pairs of coefficients falling outside of the acceptable range of ≤ 0.10 out of a possible 15 coefficient pairs per data condition. n = combinations of ratings generated under current data condition.

Overall, Kappa accounted for the highest number of differences between pairs of coefficients. Kappa fell out of acceptable range with AC1 across 26 data conditions, 20 out of 35 conditions when compared to Lambda-1 and *S*, 12 out of 35 conditions when compared to Krippendorff's Alpha, and eight out of 35 conditions in comparison to Lambda-2. All possible pairs of Lambda-1, *S*, and AC1 were within the criteria across all 35 data conditions. Table 4.4 provides an overview of the counts of pairs outside the acceptable range of ≤ .10.

Table 4.4

Out of Range Values Between Coefficient Pairs (3x3)

	1	2	3	4	5	6
1 - Lambda-1	-	6	20	0	0	0
2 - Lambda-2	6	-	8	6	8	2
3 - Kappa	20	8	-	20	26	12
4 - *S*	0	6	20	-	0	0
5 - AC1	0	8	26	0	-	16
6 - Kripp.'s Alpha	0	2	12	0	16	-

4x4 Coefficient Comparisons

To maintain uniformity across chapters, data conditions were assigned the same number given in the previous chapter. The final group of 45 data conditions span across conditions 1 – 30 and 51 – 65 in the results analysis. Similar to results from the two-point and three-point agreement data, higher coefficient values were produced where the data conditions had the highest specified values of agreement across 4x4 data conditions. Data conditions 51 – 65 produced the greatest number of possible rating combinations since the agreement was set to vary across all four agreement cells (*a*, *f*, *k*, and *p*) in these scenarios. The total number of ratings provided under these conditions ranged from 9,152 to 52,976 across agreement percentages of 75% to 95%. The number of ratings provided in data conditions 1 – 30 ranged from 76 to 96 ratings. Once again, the increase in ratings combinations led to less variation between the mean coefficient values. For data conditions 51 – 65 all mean coefficient values were within the

acceptable range of differences ≤ .10. For further explanation about individual 4x4 data conditions see Table 3.3.

Table 4.5

4x4 Mean Coefficient Values

Data Condition	n	Lambda-1	Lambda-2	Kappa	*S*	AC1	Kripp.'s Alpha	Ratio
1	96	0.93	0.95	0.80	0.93	0.94	0.85	4
2	91	0.87	0.89	0.67	0.87	0.89	0.76	8
3	86	0.80	0.84	0.57	0.80	0.83	0.69	9
4	81	0.74	0.78	0.48	0.73	0.77	0.64	8
5	76	0.67	0.72	0.40	0.67	0.71	0.59	7
6	96	0.93	0.95	0.80	0.93	0.94	0.85	4
7	91	0.87	0.90	0.67	0.87	0.89	0.76	8
8	86	0.81	0.84	0.57	0.80	0.83	0.69	9
9	81	0.74	0.79	0.48	0.73	0.77	0.64	8
10	76	0.68	0.74	0.40	0.67	0.71	0.59	7
11	96	0.93	0.91	0.80	0.93	0.94	0.85	4
12	91	0.87	0.83	0.67	0.87	0.89	0.76	7
13	86	0.80	0.75	0.57	0.80	0.83	0.69	8
14	81	0.73	0.68	0.48	0.73	0.77	0.64	6
15	76	0.67	0.62	0.40	0.67	0.71	0.59	6
16	96	0.93	0.92	0.83	0.93	0.94	0.85	1
17	91	0.87	0.84	0.72	0.87	0.88	0.77	5
18	86	0.80	0.75	0.61	0.80	0.83	0.69	7
19	81	0.73	0.68	0.61	0.73	0.76	0.69	3
20	76	0.67	0.63	0.49	0.67	0.70	0.61	5
21	96	0.93	0.91	0.80	0.93	0.94	0.85	4
22	91	0.87	0.81	0.67	0.87	0.89	0.76	7
23	86	0.80	0.72	0.57	0.80	0.83	0.69	9
24	81	0.73	0.63	0.48	0.73	0.77	0.64	7
25	76	0.67	0.55	0.40	0.67	0.71	0.59	9
26	96	0.93	0.92	0.83	0.93	0.94	0.85	1
27	91	0.87	0.83	0.72	0.87	0.88	0.77	5
28	86	0.80	0.74	0.61	0.80	0.83	0.69	6
29	81	0.74	0.67	0.61	0.73	0.76	0.69	3
30	76	0.68	0.60	0.49	0.67	0.71	0.61	6
51	42608	0.93	0.93	0.93	0.93	0.94	0.93	0
52	39480	0.87	0.87	0.85	0.87	0.87	0.87	0
53	52976	0.80	0.80	0.78	0.80	0.80	0.80	0
54	30014	0.73	0.74	0.71	0.73	0.74	0.75	0
55	24860	0.67	0.67	0.65	0.67	0.67	0.70	0
56	12557	0.93	0.94	0.93	0.93	0.94	0.93	0
57	12848	0.87	0.87	0.85	0.87	0.87	0.86	0
58	52976	0.80	0.80	0.79	0.80	0.81	0.80	0
59	9985	0.74	0.75	0.72	0.73	0.74	0.75	0
60	9152	0.68	0.69	0.65	0.67	0.68	0.69	0
61	26557	0.93	0.93	0.93	0.93	0.94	0.93	0
62	23994	0.87	0.87	0.85	0.87	0.87	0.86	0
63	52976	0.80	0.80	0.78	0.80	0.80	0.80	0
64	17679	0.74	0.74	0.71	0.73	0.74	0.75	0
65	14022	0.67	0.67	0.64	0.67	0.67	0.69	0

Note. Ratio refers to the number of pairs of coefficients falling outside of the acceptable range of ≤ 0.10 out of a possible 15 coefficient pairs per data condition. n = combinations of ratings generated under current data condition.

Data conditions 1 – 30 varied agreement levels across two out of four agreement cells. These limited number of combinations produced a greater amount of variation in the mean coefficient values under each data condition as compared to data conditions 51 – 65. Lower agreement levels produced lower coefficient values in each cluster of five data conditions. Data conditions with a higher specified agreement level resulted in less differences according to the qualifying criteria (differences ≤ .10) than conditions with lower agreement levels. Data conditions 16 – 20 and 26 – 30 distributed disagreement across three cells, while 1 – 15 and 21 – 25 fixed disagreement in a single cell. Distributing disagreement across multiple cells led to more consistent coefficient values in comparison to distributions fixing disagreement to one cell location. Data conditions 16 and 26 found one difference outside of the acceptable range. Both of the out of range differences were between Kappa and AC1 at a value of 0.11. Full results of mean coefficient values across the 45 4x4 agreement matrices data conditions are provided in Table 4.5.

Table 4.6

Out of Range Values Between Coefficient Pairs (4x4)

	1	2	3	4	5	6
1 - Lambda-1	-	1	30	0	0	10
2 - Lambda-2	1	-	26	1	4	8
3 - Kappa	30	26	-	30	30	14
4 - *S*	0	1	30	-	0	10
5 - AC1	0	4	30	0	-	20
6 - Kripp.'s Alpha	10	8	14	10	20	-

Overall, Kappa accounted for the highest number of differences between pairs of coefficients, differing 130 out of 225 pairs. Kappa fell out of acceptable range with AC1, Lambda-1, and *S* across 30 data conditions. Kappa was out of range in 26 out of 45 conditions when compared to Lambda-2, and 14 out of 45 conditions in comparison to Krippendorff's Alpha. All possible pairs of Lambda-1, *S*, and AC1 were within the criteria across all 45 data

conditions. Lambda-2 was out of range with Lambda-1 and *S* under one data condition (#25). Table 4.6 provides an overview of the counts of pairs outside the acceptable range of ≤ .10.

Results: Research Question 2

Classification Consistency of Coefficient Values

The first research question addressed the similarity of coefficients according to the mean value for each coefficient under each data conditions. Similarity of means does not guarantee similarity of distributions. The second research question addressed the classification consistency for each coefficient across each data condition. There have been three classifications primarily used for interpreting the magnitude of the Kappa coefficient (Altman, 1991; Fleiss, 1981; Landis & Koch, 1977). Researchers acknowledged that the benchmark agreement categories are arbitrary, however they were developed as helpful guidelines for making interpretations of chance-corrected agreement coefficients (Ludbrook, 2002; Oleckno, 2008). In the current study, Landis and Koch's (1977) recommendations were used to classify coefficient values produced for each data condition. Landis and Koch (1977) considered coefficients values ranging from 0.81 to 1.00 as "almost perfect" agreement, 0.61 to 0.80 as "substantial" agreement, 0.41 to 0.60 as "moderate" agreement, 0.21 to 0.40 as "fair" agreement, 0.00 to 0.20 as "slight" agreement, and values less than 0.00 as no agreement.

Table 4.7

Count Across Benchmark Agreement Levels for Data Condition 1 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	96
Lambda-2	0	0	0	4	16	76
Kappa	0	2	2	6	16	70
S	0	0	0	0	0	96
AC1	0	0	0	0	0	96
Kripp. Alpha	0	0	0	4	14	78

Note. Benchmark agreement categories from Landis & Koch (1977).

2x2 Classification Consistency

Lambda-1 produced coefficient values within the same benchmark category in 11 out of 15 data conditions for the 2x2 agreement matrices. In the other four data conditions the coefficient values were distributed across two benchmark categories. Lambda-1 was the second most consistent coefficient according to benchmark agreement distribution. Since *S* produces a constant value for data conditions according to rating scale size it was the most consistently classified coefficient. Across all 15 data conditions for the 2x2 agreement matrices *S* produced values in a single category in each of the 15 conditions. For the conditions with agreement at 85% or higher all classifications were in the "almost perfect" category and the conditions with agreement at 75% and 80% resulted in coefficient values in the "substantial" agreement category. For example, data condition 1 produced 96 agreement distributions according to the specifications. All 96 distributions resulted in coefficient values ranging from 0.81 to 1.00 for Lambda-1, *S*, and AC1. Tables 4.7 and 4.8 and Figure 4.2 show full results for data condition 1. For further information about individual 2x2 data conditions see Table 3.1.

Table 4.8

Percentage of Values within Benchmark Levels for Data Condition 1 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	4.2%	16.7%	79.2%
Kappa	0.0%	2.1%	2.1%	6.3%	16.7%	72.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	4.2%	14.6%	81.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

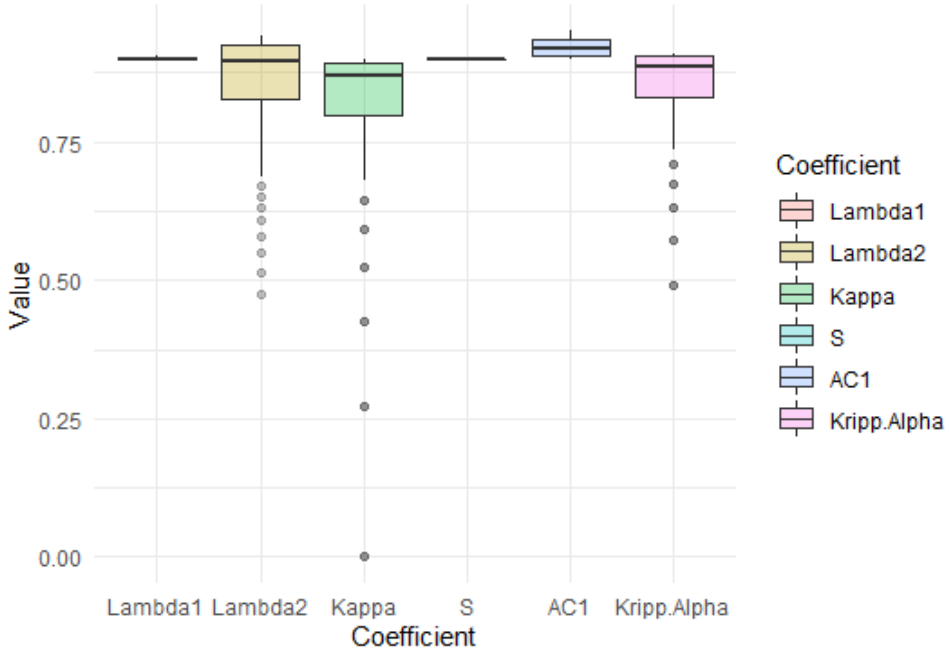


Figure 4.2. Coefficient Values Under Data Condition 1 (2x2).

Lambda-2 distributed coefficients across two to six categories across the 15 data conditions, making it one of the least consistent coefficients under the 2x2 agreement conditions. Kappa had values spread across four to six benchmark agreement categories depending on the data condition. Results from data condition 14 are displayed in Tables 4.9 and 4.10 and Figure 4.3. Data condition 14 was one of four conditions that resulted in Lambda-2 being more variant than Kappa. AC1 distributions fell across two benchmark categories in 12 out of 15 data conditions. Krippendorff's Alpha distributed values across two categories under eight conditions and across three categories in seven conditions. Distributions for each data condition can be found in Appendix D.

Table 4.9

Count Across Benchmark Agreement Levels for Data Condition 14 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	81	0	0
Lambda-2	4	6	10	21	40	0
Kappa	4	6	14	57	0	0
S	0	0	0	0	81	0
AC1	0	0	0	1	80	0
Kripp. Alpha	0	0	0	20	61	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table 4.10

Percentage of Values within Benchmark Levels for Data Condition 14 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
Lambda-2	4.9%	7.4%	12.3%	25.9%	49.4%	0.0%
Kappa	4.9%	7.4%	17.3%	70.4%	0.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	1.2%	98.8%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	24.7%	75.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

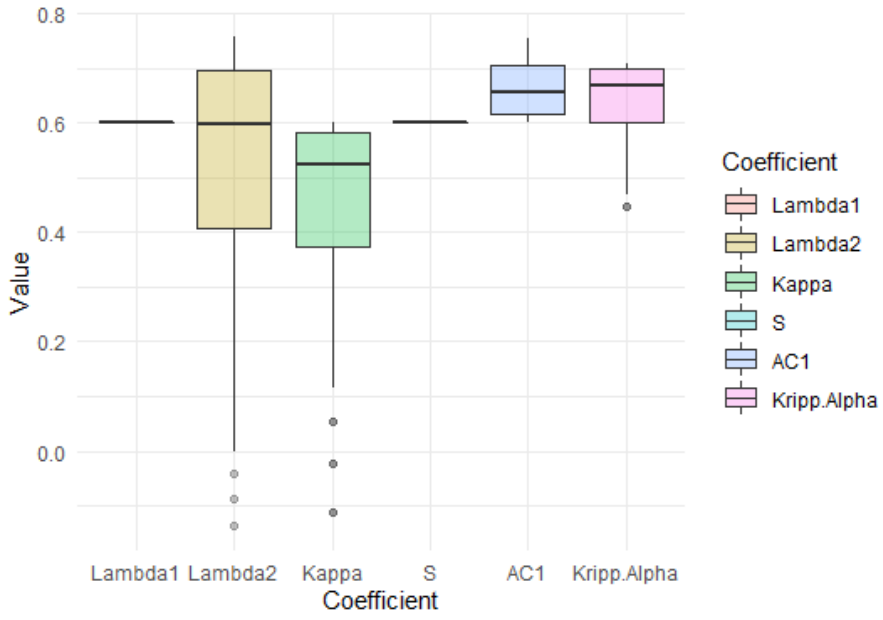


Figure 4.3. Coefficient Values Under Data Condition 14 (2x2).

3x3 Classification Consistency

All coefficients produced for the *S* coefficient fell in the same benchmark category across all 35 data conditions. *S* produced values with the highest agreement classification in 28 out of 35 data conditions. Lambda-1 classified all coefficients in the same benchmark category in 33 out of 35 data conditions. In the other two data conditions (5 and 25) Lambda-1 values were classified across two categories. In data condition 5, 65.8% of Lambda-1 coefficient values were in the "substantial" agreement category and the remaining 34.2% in the "moderate" category. Data condition 25 resulted in 98.3% of coefficient values in the "substantial" agreement category and 1.7% in the "moderate" category. In both data conditions 5 and 25, agreement was set at 75% and disagreement at 25%. The two conditions differed in the number of cells agreement (2 and 3 cells) and disagreement (1 and 2 cells) could vary.

Table 4.11

Count Across Benchmark Agreement Levels for Data Condition 23 (3x3)

	No Agreement < 0.00	Slight 0.00 - 0.20	Fair 0.21 - 0.40	Moderate 0.41 - 0.60	Substantial 0.61 - 0.80	Almost Perfect 0.81 - 1.00
Lambda-1	0	0	0	0	3741	0
Lambda-2	0	0	0	115	2509	1117
Kappa	1	5	34	248	3453	0
S	0	0	0	0	3741	0
AC1	0	0	0	0	493	3248
Kripp. Alpha	0	0	0	69	3396	276

Note. Benchmark agreement categories from Landis & Koch (1977).

AC1 had at least a share of the highest agreement classification category in all 35 data conditions. In seven data conditions (3, 8, 13, 18, 23, 28, and 33) AC1 stood alone as producing the highest benchmark agreement level, while in the other 28 data conditions it shared this title with Lambda-1 or *S*. In all seven conditions where AC1 had the highest benchmark agreement level produced the agreement was set at 85% and disagreement at 15%. Tables 4.11 and 4.12 and Figure 4.4 show the full results for data condition 23 (3x3).

Table 4.12

Percentage of Values within Benchmark Levels for Data Condition 23 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	3.1%	67.1%	29.9%
Kappa	0.0%	0.1%	0.9%	6.6%	92.3%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	13.2%	86.8%
Kripp. Alpha	0.0%	0.0%	0.0%	1.8%	90.8%	7.4%

Note. Benchmark agreement categories from Landis & Koch (1977).

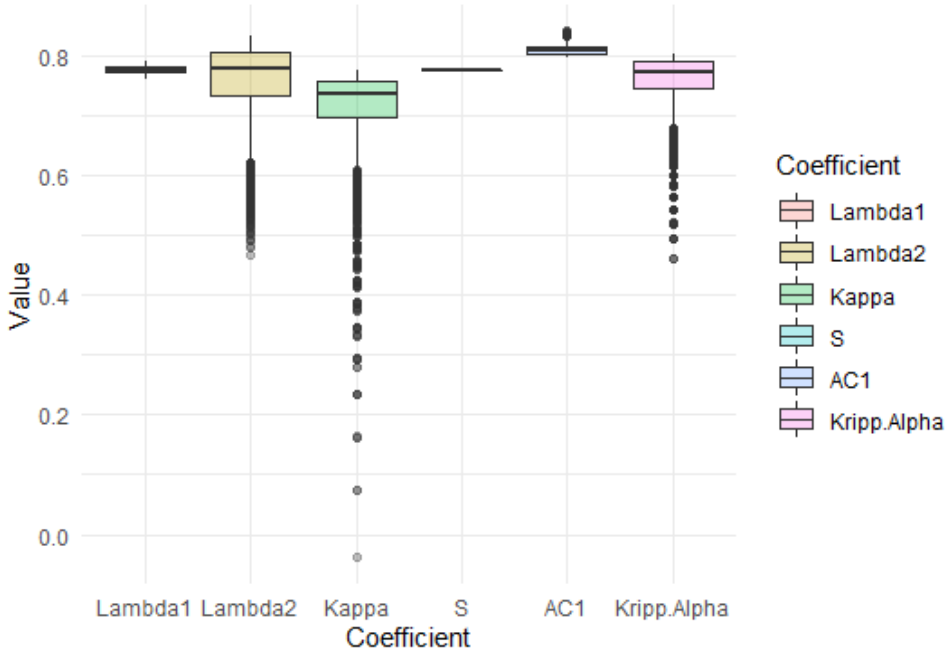


Figure 4.4. Coefficient Values Under Data Condition 23 (3x3).

Overall, Lambda-2 had the most similar category distributions as Krippendorff's Alpha across the 3x3 agreement matrices. Lambda-2 was distributed across two categories in 15 conditions, three categories in 19 conditions, and four categories in one data condition. Krippendorff's Alpha was spread across two benchmark agreement levels in 18 data conditions and across three benchmark agreement levels in 17 data conditions. In data condition 5, Lambda-2 produced the lowest benchmark agreement level results out of the six coefficients. Under data condition 5 agreement was set at 95% across two cells and disagreement was set at 5% in a

single cell. Out of the 76 agreement combinations used in condition 5, Lambda-2 was under 0.61 (below "substantial" benchmark) in all situations. Krippendorff's Alpha was the closest to Lambda-2 in data condition 5, with 32 out of 76 coefficient values under 0.61. Kappa produced coefficient values falling in the lowest benchmark category in the other 34 data conditions. Also, Kappa was the most variant coefficient across all 35 conditions. Kappa produced results across at least four different benchmark agreement levels across all 35 conditions. When agreement was set at 90%, Kappa always produced coefficient values spanning all six benchmark categories. Distributions for each data condition can be found in Appendix E. For a further explanation about individual 3x3 data conditions see Table 3.2.

4x4 Classification Consistency

Lambda-1 produced coefficient values within a single benchmark category in 43 out of 45 data conditions for the 4x4 agreement matrices. In the other two data conditions (3 and 53) the coefficient values were distributed across two benchmark categories, these two conditions both set agreement at 85%. Lambda-1 was the third most consistent coefficient according to benchmark agreement distribution. *S* was the most consistently classified coefficient. Across all 45 data conditions for the 4x4 agreement matrices *S* produced values in a single category. AC1 produced coefficient values within a single benchmark category in 44 out of 45 data conditions. In data condition 53 AC1 values were categorized across two benchmark levels. Data condition 53 had agreement set at 85% and distributed across all four agreement cells. Disagreement was set at 15% and distributed equally across all six strict cells under data condition 53. Tables 4.13 and 4.14 and Figure 4.5 show full results for data condition 53.

Table 4.13

Count Across Benchmark Agreement Levels for Data Condition 53 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	36864	16112
Lambda-2	0	0	0	0	25121	27855
Kappa	0	0	0	0	52976	0
S	0	0	0	0	52976	0
AC1	0	0	0	0	2893	50083
Kripp. Alpha	0	0	0	0	11940	41036

Note. Benchmark agreement categories from Landis & Koch (1977).

Table 4.14

Percentage of Values within Benchmark Levels for Data Condition 53 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	69.6%	30.4%
Lambda-2	0.0%	0.0%	0.0%	0.0%	47.4%	52.6%
Kappa	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	5.5%	94.5%
Kripp. Alpha	0.0%	0.0%	0.0%	0.0%	22.5%	77.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

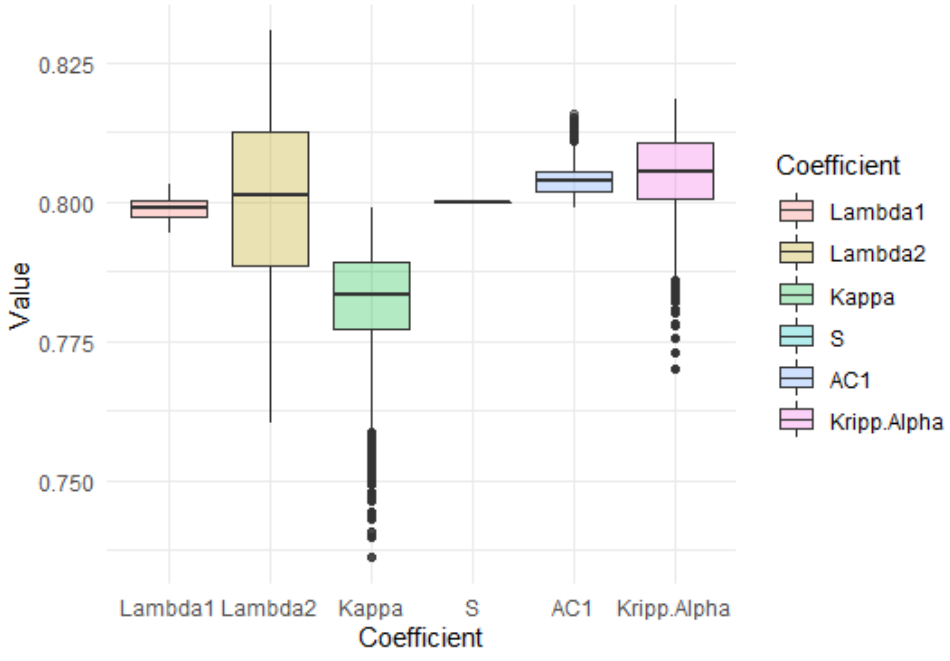


Figure 4.5. Coefficient Values Under Data Condition 53 (4x4).

Lambda-2 distributed coefficients across one benchmark category in 31 conditions and across two categories in the other 14 out of 45 data conditions, making it one of the more consistent coefficients under the 4x4 agreement conditions. Kappa had values spread across one to six benchmark agreement categories depending on the data condition. Kappa produced consistently classified results across data conditions 51 to 65. These conditions varied agreement across all four cell locations and disagreement in six to 12 cells. In 16 out of 45 data conditions, Kappa produced coefficients across five benchmark categories. Results from data condition 14 are displayed in Tables 4.9 and 4.10. Krippendorff's Alpha distributed values across one category in 12 data conditions, two categories under 21 conditions, and across three categories in 12 conditions. Distributions for each data condition can be found in Appendix F. For further descriptions about individual 4x4 data conditions see Table 3.3.

Results: Research Question 3

Correlation of Coefficient Values

In order to determine the extent to which coefficients were measuring agreement in similar ways and if they were producing similar values, correlations between coefficients within each data condition were analyzed. The *S* coefficient was removed entirely from analyzing this research question since it produces a constant value and correlations are indeterminable when variance does not exist. Also, under a few circumstances Lambda-1 produced a constant value and is removed from the analysis in these corresponding data conditions. The conditions where this occurred will be noted under each applicable agreement matrix.

2x2 Coefficient Correlations

Correlations between the coefficients ranged from positive to negative, no relationship to perfectly linear, and everything in between. Lambda-1 was not included in the correlation

analysis under data conditions 11 – 15 since Lambda-1 produced a constant value across these conditions. The location of disagreement had an impact on the direction of the correlation between Lambda-1 and Lambda-2 values across data conditions 1 – 10. In data conditions 1 – 5 the disagreement was located in the strict cell, while in conditions 6 – 10 the disagreement was located in the lenient cell. Under the first five data conditions Lambda-1 and Lambda-2 had a strong negative correlation (-0.84 to -0.83). In data conditions 6 – 10 the correlations were strong and positive between Lambda-1 and Lambda-2 (0.90 to 0.98). The relationship between Lambda-2 and AC1 relationship had a similar change in direction across different clusters of data conditions. Under data conditions 1 – 7 (-0.90 to -0.05) and 11 – 15 (-0.41 to -0.33) Lambda-2 was negatively correlated to AC1, while in conditions 8 – 10 (0.17 to 0.57) the relationship was positive.

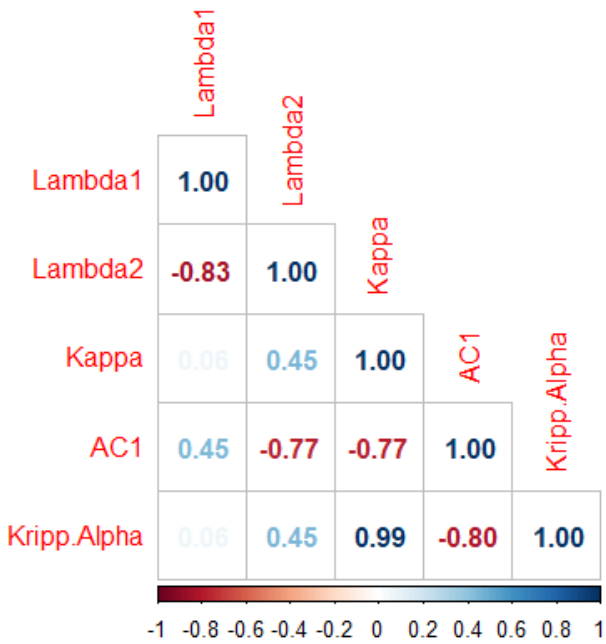


Figure 4.6. Correlation results under data condition 3 (2x2).

There were differences in the magnitude of correlations between Lambda-2 and AC1 across clusters of data conditions. The strength of correlation ranged from moderate to strong in conditions 1 – 5, weak to moderate in conditions 6 – 10, and weak to moderate in conditions 11 –

15. Figures 4.6 – 4.8 display the direction and magnitude of correlations from selected data conditions. Red numbers indicate a negative correlation, while numbers in blue indicate a positive correlation. The darkness of the font's color shade helps indicate the magnitude of the correlation. The darker the color of the font the higher the correlation and the lighter the color of the font the weaker the correlation value.

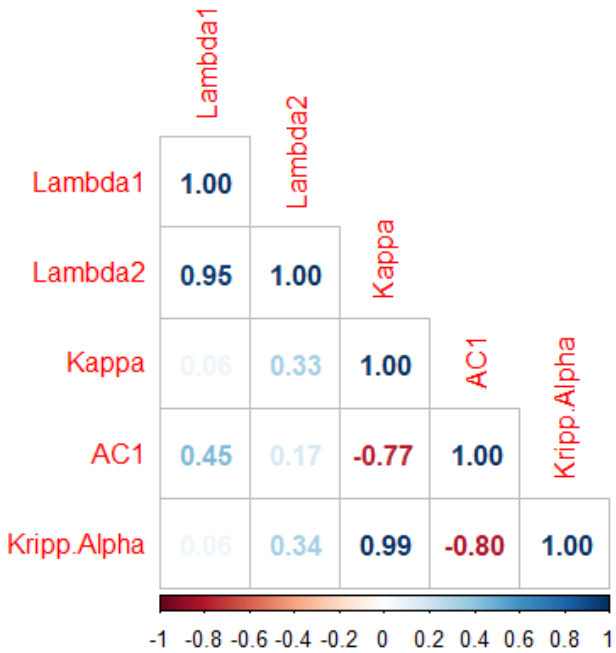


Figure 4.7. Correlation results under data condition 8 (2x2).

The relationships between Kappa and AC1, AC1 and Krippendorff's Alpha, and Kappa and Krippendorff's Alpha remained consistent across the 15 data conditions for the 2x2 agreement matrices. AC1 produced strong negative correlations to both Kappa and Krippendorff's Alpha, indicating that AC1 tends to measure chance-corrected agreement differently from these two coefficients. Kappa and Krippendorff's Alpha produced near perfect, positive correlations, indicating these two coefficients closely measure agreement. To view all correlation tables and figures for 2x2 conditions see Appendix G. For further information about individual 2x2 data conditions see Table 3.1.

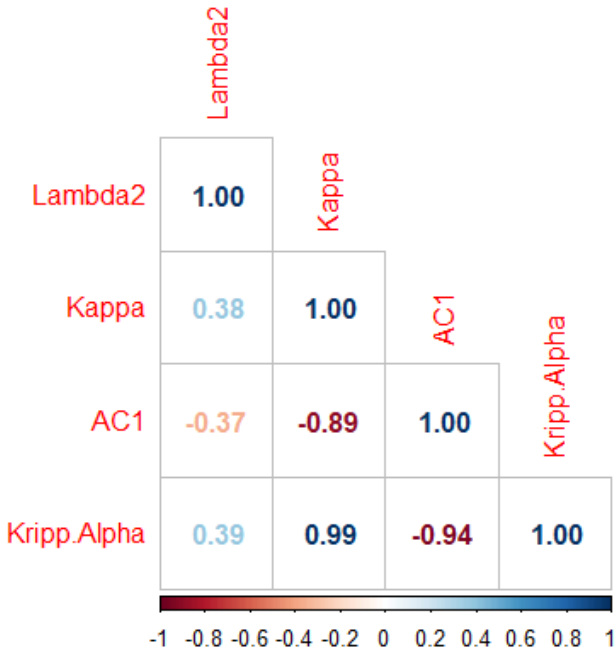


Figure 4.8. Correlation results under data condition 13 (2x2).

3x3 Coefficient Correlations

Correlations between the coefficients produced different values in many cases, however some coefficients similarly measured agreement. Lambda-1 was not included in the correlation analysis under data conditions 16 – 20 since Lambda-1 produced a constant value across these conditions. Lambda-1 and Lambda-2 had positive correlations across all conditions except conditions 1 – 5. In data conditions 1 – 5 there was a strong, negative correlation between Lambda-1 and Lambda-2 (-0.98 to -0.89). There was a strong, positive correlation for this coefficient pair across conditions 11 – 20 (0.91 to 0.99). Data conditions 1 – 20 distributed ratings across three out of nine possible cell locations in the agreement matrix. Data conditions 21 – 35 varied agreement across all three agreement locations and distributed disagreement in two to four cell locations. These conditions produced moderate and positive correlation values. Lambda-1 and AC1 produced positive correlation values in data conditions 1 – 15 and 22 – 30 ranging from weak to strong. The highest correlation values within each cluster of data conditions (cluster refers to every five data conditions with the same agreement and

disagreement specifications) occurred in the condition with the lowest agreement level, see Table 4.15 for results from data condition 5. As agreement level decreased, the correlation between Lambda-1 and AC1 increased. Lambda-1 had no relationship with Kappa or Krippendorff's Alpha.

Table 4.15

Correlation Matrix for Data Condition 5 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	-0.98***	-			
Kappa	0.03	0.17	-		
AC1	0.77***	-0.88***	-0.60***	-	
Kripp.Alpha	0.03	0.17	1.00***	-0.61***	-

Note. *** $p < .001$.

Lambda-2 and Kappa had moderate positive relationships across all conditions. Under conditions 1 – 20 the correlation value decreased as agreement decreased, while under conditions 21 – 35 the correlation value increased as agreement decreased. Again, data conditions 1 – 20 distributed agreement and disagreement across three out of nine cells in the agreement matrix. While data conditions 21 – 35 distributed agreement and disagreement across five to seven out of nine possible cells. Across data conditions 1 – 5 and 16 – 20, Lambda-2 and AC1 produced moderate to strong and negative correlation values. Under data conditions 6 – 15 the correlations between Lambda-2 and AC1 began as weak negative in the 95% agreement conditions to strong positive correlations as agreement decreased to 75% agreement. All correlation values between Lambda-2 and AC1 under conditions 21 – 35 were weak and negative levels. Lambda-2 had moderate positive correlation levels with Krippendorff's Alpha across conditions with 95% agreement, as agreement decreased the correlations approached 0.0. Figures 4.9 and 4.10 help illustrate changes within groups of data conditions as agreement decreased and disagreement increased.

Kappa and AC1 produced strong, negative correlation values across all 35 data conditions, ranging from -0.92 to -0.60. Likewise, AC1 and Krippendorff's Alpha had strong, negative correlation values across all 35 data conditions spanning the same range (-0.92 to -0.60). Krippendorff's Alpha and Kappa produced near perfect and strong correlation values across the 35 conditions (0.94 to 1.00), meaning that these two coefficients measured agreement in similar ways. To view all correlation tables and figures for 3x3 conditions see Appendix H. For further explanation about individual 3x3 data conditions see Table 3.2.

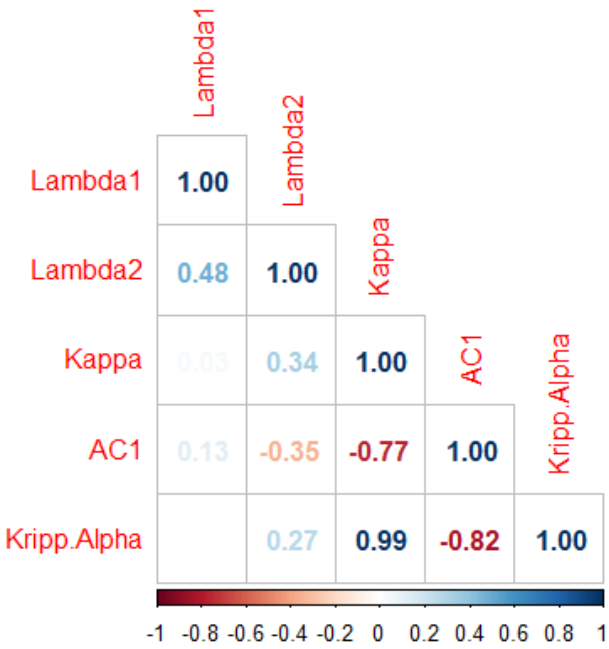


Figure 4.9. Correlation results under data condition 26 (3x3).

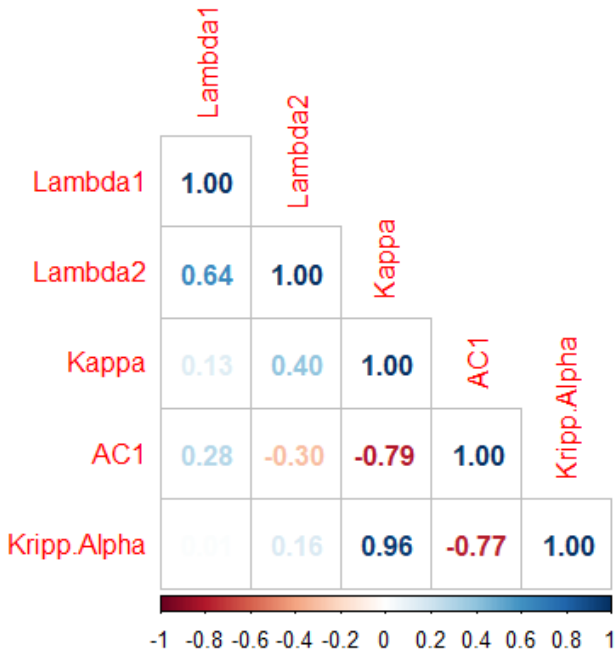


Figure 4.10. Correlation results under data condition 29 (3x3).

4x4 Coefficient Correlations

Lambda-1 was not included in the correlation analysis under data conditions 11 – 30 since Lambda-1 produced a constant value across these conditions. Lambda-1 had near perfect, positive correlations with Lambda-2 in data conditions 1 – 10. Under conditions 56 – 60 Lambda-1 and Lambda-2 had moderate to strong positive correlations (0.53 to 0.88). All correlations between Lambda-1 and Lambda-2 were positive. Lambda-1 had no relationship with Kappa and Krippendorff's Alpha in conditions 1 – 10. Most other correlations between Lambda-1 and these two coefficients were weak and positive. Lambda-1 and AC1 had moderate correlations across conditions 1 – 10 and 51 – 55. Tables 4.16 and 4.17 illustrate the difference between correlation results for Lambda-1 when agreement and disagreement varied across a small number of cells (data condition 10) and a larger number of cells (data condition 55).

Table 4.16

Correlation Matrix for Data Condition 10 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	1.00***	-			
Kappa	0.02	0.00	-		
AC1	0.78***	0.79***	-0.60***	-	
Kripp.Alpha	0.02	0.00	1.00***	-0.61***	-

Note. *** $p < .001$.

Table 4.17

Correlation Matrix for Data Condition 55 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.22***	-			
Kappa	-0.11***	0.06***	-		
AC1	0.61***	0.27***	-0.82***	-	
Kripp.Alpha	-0.24***	0.11***	0.98***	-0.87***	-

Note. *** $p < .001$.

Lambda-2 had no relationship with Kappa or Krippendorff's Alpha across conditions 1 – 10. Data conditions 16 and 17 were the only correlations that were weak for Lambda-2's pairings with Kappa and Krippendorff's Alpha. In these two unique cases the correlation values were strong and negative (-0.72 and -0.79). The correlation values between Lambda-2 and AC1 had greater variation across data conditions 1 – 25 than across 26 – 30 and 51 – 65. The first 25 conditions produced correlation values ranging from strong positive (under data conditions 15-17) to strong negative (under data conditions 23 – 25). The correlation values across the remaining conditions between Lambda-2 and AC1 were weak negative and weak positive levels.

AC1's pairings with Kappa and Krippendorff's Alpha had matching correlation values. Under data conditions 1 – 15 and 21 – 25 the correlation values were strong and negative. Under conditions 16 – 20, 26 – 30, and 51 – 65 correlation values between AC1's pairings with Kappa and Krippendorff's Alpha were near perfect and negative. Kappa and Krippendorff's Alpha had near perfect, positive correlation values across all data conditions. Figure 4.11 illustrates the

correlation patterns between Kappa, AC1, and Krippendorff's Alpha. To view all correlation tables and figures for the 4x4 data conditions see Appendix I. For further clarification about individual 4x4 data conditions see Table 3.3.

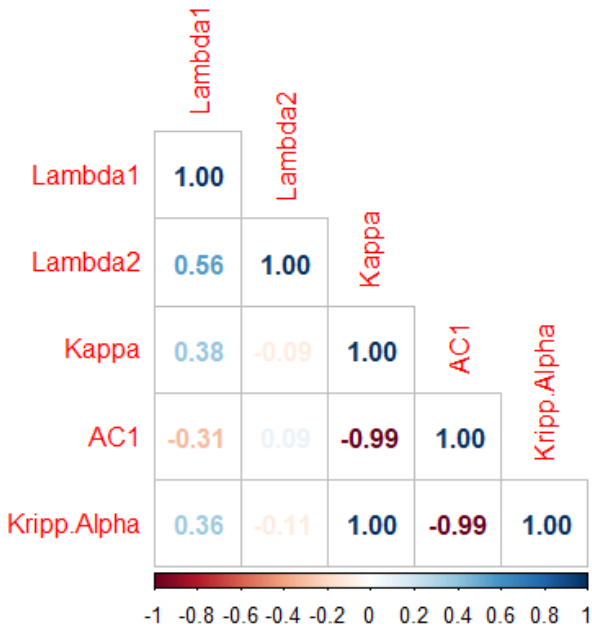


Figure 4.11. Correlation results under data condition 56 (4x4).

Summary

Overall, the coefficients produced the most similar values when the agreement was at the higher levels. Also, coefficient values increased and were similar when conditions distributed agreement and disagreement across more cells. Under the 2x2 and 3x3 data conditions, Lambda-1 had similar mean values to *S*, AC1, and Krippendorff's Alpha in all scenarios. Across all of the 4x4 agreement conditions Lambda-1 had similar mean values with *S* and AC1, while Lambda-2 only had one difference out of 45 possible conditions with Lambda-1 and *S*. Kappa differed the most from other coefficients across each size rating scale. When agreement was highest (95%) all coefficients produced their highest values. As agreement decreased, the coefficient values decreased and fell out of acceptable range (within .10 of other coefficient values; Raadt et al., 2021) more frequently. The location of disagreement was most impactful for Lambda-2, this

occurred due to the extreme population proportion values applied in this study. As disagreement was distributed more evenly across the cell locations, coefficients produced more similar mean values. Kappa performed like the other coefficients when agreement and disagreement was distributed across more possible cells. For example, under data conditions 21 – 35 for the 3x3 agreement matrix the location of disagreement was distributed across two to four cell locations and Kappa was most like other coefficients in these scenarios. This was also true in the 4x4 conditions varying agreement and disagreement across more cells.

While the mean value of coefficients within each data condition was usually similar, the coefficients produced different amounts of variability when looking across all values within each condition. *S* was the most consistent coefficient across all rating scale sizes, as it produced a constant value. When using Landis and Koch's (1977) classification categories, Lambda-1 and AC1 were the most consistent coefficients. Each of these two coefficients produced values across one or two benchmark categories under all data conditions. As the rating scales grew in size, Lambda-2 became more consistent. In the 2x2 conditions Lambda-2 values stretched across five out of six benchmark classifications, while under the 3x3 conditions values spanned across three categories, and just two categories under the 4x4 conditions. Kappa was the least consistent coefficient across all rating scale sizes.

Pearson correlation coefficients were calculated to understand the relationship between coefficients. The correlation coefficient provided insight into whether two coefficients were measuring agreement in similar ways. *S* was not included in this part of the analysis since it produced a constant value across all conditions. Lambda-1 was removed from correlation analysis under data conditions 16 to 20 (3x3) and 11 to 30 (4x4) since it produced a constant under these specific conditions. Kappa and Krippendorff's Alpha had almost perfect, positive

correlations across all data conditions. Lambda-1 was most correlated with AC1 when agreement was at 75% on the three-point scale and had high correlation values with Lambda-2 under certain data conditions. AC1 produced strong, negative correlations with Kappa and Krippendorff's Alpha across all data conditions.

CHAPTER V: DISCUSSION AND IMPLICATIONS

This dissertation investigated the usefulness and application of two recently developed chance-corrected agreement coefficients designed for use with rater-mediated assessment, Lambda-1 and Lambda-2 (Lambert et al., 2021). Rater-mediated assessments are complex and heavily reliant on the consistency and accuracy of ratings to provide valid information about scores produced from these assessments. This study combined tenets of rater-mediated assessment and interrater reliability theories. Also, the study focused on one specific aspect of rater-mediated assessments, the reliability measures, to evaluate the viability of alternative reliability measures.

This study serves as a next step in understanding the performance and application of Lambda-1 and Lambda-2. These coefficients along with commonly studied IRR coefficients were calculated under prespecified data conditions. These conditions varied the amount and location of agreement and disagreement between a hypothetical rater and a correct rating. Additionally, a comparison of the performance of the coefficients was examined more in depth in this dissertation.

Research related to rater-mediated assessment theory and pertinent topics linked to IRR were examined to identify the applications and concerns with existing measures of agreement involving raters (see Chapter I and Chapter II). Next, selected chance-corrected agreement coefficients were reviewed to provide background on the development, uses, and knowledge of existing coefficients appropriate for use in the context of this study (Chapter III). Methods for evaluating the performance of the coefficients of interest were considered and applied across 135 planned data conditions for two-, three-, and four-point agreement matrices (Chapter III and

Chapter IV). This chapter discusses conclusions based on the study's research questions, limitations, implications, and suggestions for future research.

Conclusions

Findings According to Research Question 1: Under what conditions do Cohen's Kappa, *S*, Krippendorff's Alpha, Gwet's AC1, Lambda-1, and Lambda-2 produce similar values?

The performance of Lambda-1 and Lambda-2 and four other IRR coefficients were examined under prespecified data conditions. The other coefficients were Kappa, *S*, AC1, and Krippendorff's Alpha. In the first research question the mean value of coefficients under the same prespecified conditions were compared to see if the average coefficient values were similar. A criterion of having a mean value within 0.10 of other coefficients was used as a similarity point (Raadt et al., 2021). In data conditions using fewer cells to distribute agreement and disagreement and when agreement was lower more differences between coefficients were found.

Results indicated that Lambda-1 produced values similar to other established chance-corrected agreement coefficients. Also, the results showed that Lambda-1 was not subject to known paradoxes associated with Kappa. Lambda-1 remained stable when agreement was very high and was not impacted by the number of cells used to distribute agreement and disagreement. Overall, Kappa was the least similar coefficient to the others in the study and produced the lowest coefficient values across conditions. This study confirmed findings that Kappa is unstable under conditions with moderately high to high agreement and when raters accurately use a limited number of rating scale points (Cicchetti & Feinstein, 1990; Gwet, 2008; Holcomb et al., 2022; Lambert et al., 2021; Xie, 2013).

Under the factors that were systematically varied in the data conditions in this study, Lambda-2 performed more like the other coefficients as the rating scale grew. The performance of Lambda-2 is dependent upon the population proportion values applied in the study. The current study applied values from actual teacher evaluation data from one state's teacher evaluation process. Lambda-2 is most interpretable when the coefficient is calculated under the rating scale size that matches the actual instrument used in practice. The other coefficients calculated in this study performed consistently across rating scale sizes. The *S* coefficient can produce a somewhat artificially inflated reliability value in situations where categories are not used by the raters (Warrens, 2012). For this reason, *S* can be best applied in situations where a rater is using all points on a rating scale. Kappa and Krippendorff's Alpha produce more consistent and stable values when all points on the rating scale are used and are selected more evenly.

Findings According to Research Question 2: To what extent are the IRR coefficients, calculated using Cohen's Kappa, S, Krippendorff's Alpha, Gwet's AC1, Lambda-1, and Lambda-2, placed in the same classification categories according to well-known taxonomies (i.e., almost perfect, substantial, moderate, fair, slight, and no agreement) under the same levels of agreement?

While the first question assessed the mean value of the six coefficients across data conditions, the second research question dealt with the classification consistency of reliability values according to historically referenced benchmark levels (Landis & Koch, 1977). Lambda-1 was classified across categories in consistent and predictable ways. When the specified level of agreement was highest, Lambda-1 was consistently classified in just one or two categories. As agreement decreased, Lambda-1 was categorized in two or three categories. The results

demonstrate that Lambda-1 is a stable measure of IRR and performed similarly to AC1 and *S* in terms of classification consistency. Lambda-2 applied population proportion values from a four-point scale and was most consistent in data conditions from the 4x4 agreement matrix, as Lambda-2 did under research question 1.

Across all sizes of agreement matrices used in this study, Kappa was the most variable coefficient. This became more apparent as the fixed agreement level increased up to 95% and category prevalence decreased down to locating disagreement in a single cell. This confirmed previous findings on the classification consistency of Kappa (Holcomb et al., 2022). This is important to point out because the other coefficients were much more consistent in producing the same benchmark agreement level under the same data condition no matter the location of agreement or disagreement. Misclassification of the reliability levels due to a statistical flaw could lead to the misrepresentation of the reliability of scores.

Findings According to Research Question 3: To what extent do Cohen's Kappa, *S*, Krippendorff's Alpha, Gwet's AC1, Lambda-1, and Lambda-2 measure agreement in similar ways?

The final research question aimed to understand whether the agreement coefficients were producing values within data conditions in a similar way as other coefficients. Lambda-1 was removed from certain scenarios when it produced a constant value. The *S* coefficient was left out of this portion of the analysis entirely since it produces a constant based on the number of categories on the rating scale. Some coefficients were strongly correlated to others across all conditions and agreement matrices, while others were dependent upon certain conditions. The magnitude and direction of the correlations for Lambda-1 and Lambda-2 to other coefficients varied more according to the location and amount of agreement and disagreement. However,

other coefficients were not as impacted by these features of the data conditions. Lambda-1 was more correlated with AC1 as agreement was reduced to 75%, this was the case across two-, three-, and four-point data conditions. As the rating scale size increased, the correlations between Lambda-1 and Kappa and Lambda-1 and Krippendorff's Alpha approached 0.0. Under the four-point data conditions with greater use of cells and distribution of agreement and disagreement Lambda-1 had a negative, moderate correlation with Kappa and Krippendorff's Alpha. As the size of the rating scale increased, Lambda-2 was less correlated with the other coefficients. Under the 2x2 agreement matrix conditions, Lambda-2 was more correlated with Krippendorff's Alpha. In the 3x3 and 4x4 data conditions, Lambda-2 was more correlated with AC1 under certain conditions. When agreement decreased Lambda-2 performed similarly to AC1. As the distribution of disagreement and agreement occurred across more cell locations, the correlations between Lambda-2 and other coefficients were similar no matter the amount of agreement or disagreement.

Krippendorff's Alpha and Kappa were strongly correlated across all data conditions even though Kappa typically produced much lower coefficient values. The variation of Krippendorff's Alpha calculated in this study, seems to be a less extreme version of Kappa as it does not overcorrect for chance-agreement as much as Kappa. To further illustrate the similarity of the performance of these two coefficients Figures 5.1 and 5.2 display the behavior of each coefficient under certain data conditions. As you can see in both figures, Krippendorff's Alpha and Kappa follow very similar patterns when analyzing coefficient values produced using the prevalence index values. AC1 produced strong negative correlations with both Krippendorff's Alpha and Kappa across data conditions. This supports previous findings that AC1 corrects for issues with the Kappa coefficient (Xie, 2013).

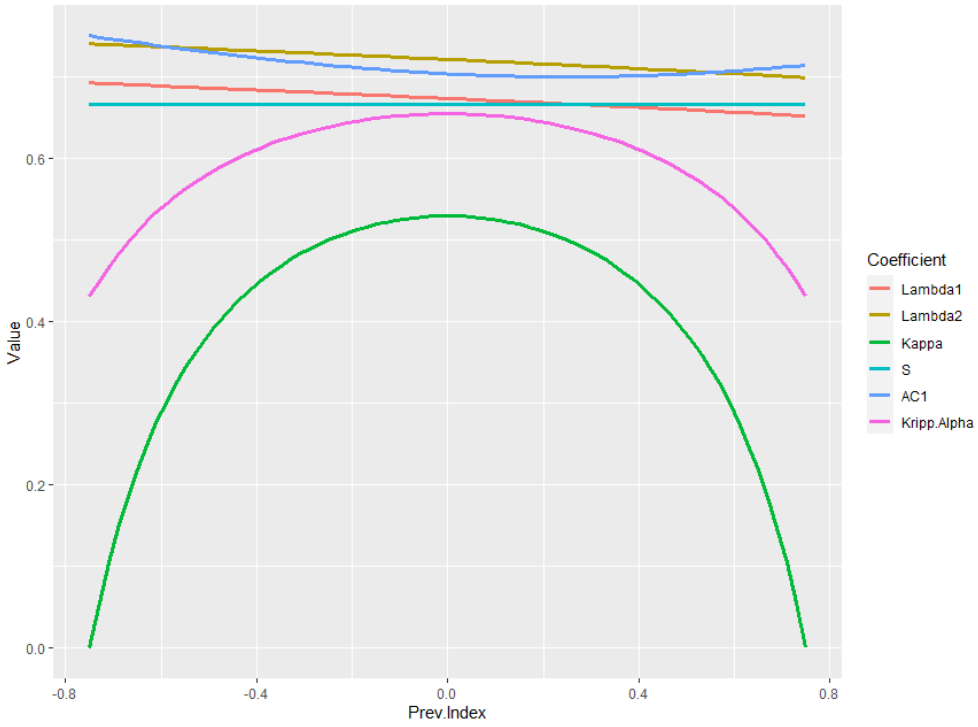


Figure 5.1. Line graph displaying coefficient performance under data condition 5 (4x4).

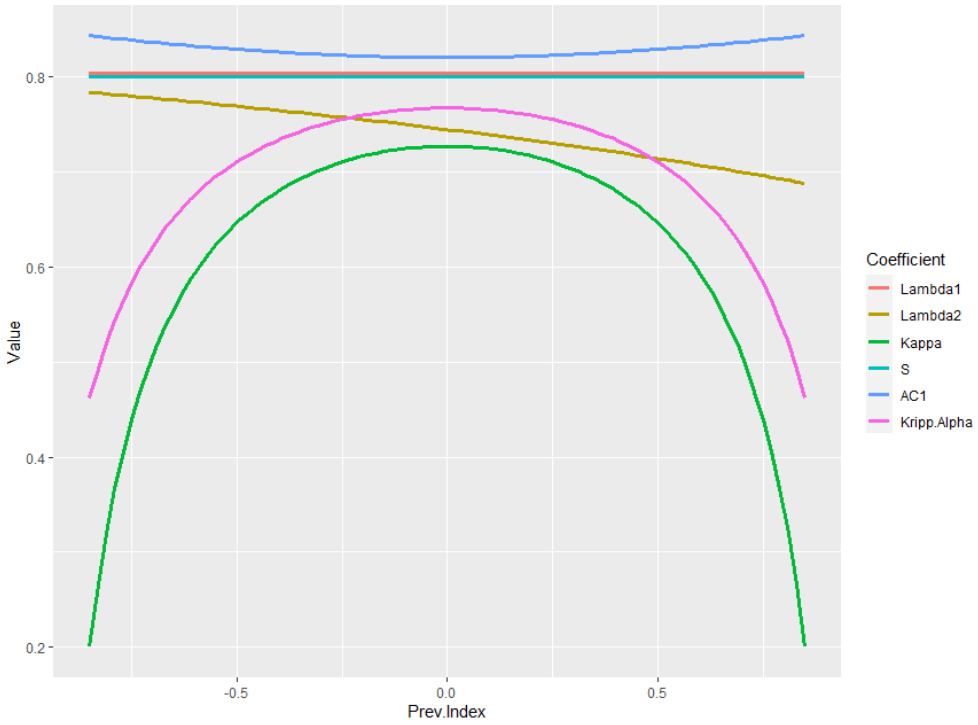


Figure 5.2. Line graph displaying coefficient performance under data condition 28 (4x4).

Limitations

There are various limitations to consider when making inferences based on this study. This study was not intended or designed to evaluate all possible data arrangements or rating scale sizes to calculate chance-corrected agreement coefficients. Conditions tested in this study systematically varied agreement matrices to explore a pattern of ratings to calculate coefficients under selected conditions. This research followed a design similar to a previous study on alternative chance-corrected agreement coefficients (Xie, 2013).

Likewise, all chance-corrected agreement coefficients and reliability measures that could be used for similar purposes were not exhausted in this study. This study targeted coefficients that have been reported and recommended for use in rater-mediated assessment situations. Also, conditions known to be problematic for the most widely used chance-corrected agreement coefficient, Kappa, were of interest in the current study to confirm and extend previous findings on alternatives to this coefficient.

This study examined agreement between a single rater and a set of correct ratings. In certain situations, this may be appropriate. However, in some applications it may be more common for two or more raters to evaluate the same performance. In these situations, other variations of the Kappa coefficient and other multiple rater reliability measures would be more appropriate to report. The current versions of Lambda-1 and Lambda-2 are applicable under conditions comparing the agreement between two raters or a single rater and a set of correct ratings. A different version of Lambda could be developed to compare the agreement between a group of raters.

The criterion value (0.10) used to evaluate research question one has been implemented in a single study and was arbitrarily selected (Raadt et al., 2021). Research question two looked

at variability according to recommended guidelines to reach a conclusion about the level of reliability. The recommended classification guidelines adopted in this study had category thresholds of 0.20 benchmark each "level" (Landis & Koch, 1977). As suggested by Raadt and colleagues (2021), other cutoff values could be considered in future studies comparing reliability coefficient values. Landis and Koch's (1977) classifications were guidelines designed to be used with Kappa. Other recommended guidelines provided by Fleiss (1981) and Altman (1991) were also designed based on Kappa coefficient values. It is important to point out that these guidelines were not developed with other IRR coefficients in consideration.

The Lambda-2 coefficient was designed to be implemented in rater-mediated assessment situations in which the population proportion of ratings is known. For the purposes of this study, the population proportion of ratings from one state's four-point teacher evaluation rubric were applied to calculate Lambda-2 coefficient values. It is quite likely that results would differ as population proportions were more or less evenly distributed across rating scales of different sizes. As the population proportion values would get closer to an even distribution across rating scale points Lambda-2 would perform more similarly to Lambda-1. For instance, on a four-point scale if population proportions of 0.25 were applied for each of the four categories, Lambda-2 would be equivalent to Lambda-1 (Lambert et al., 2021).

Implications for Policy and Practice

The initial purpose for the development and use of chance-corrected agreement coefficients is to provide an IRR measure beyond simple percentage agreement (Cohen, 1960). Chance-correct agreement coefficient have limitations. It is best practice to compute and report multiple measures of agreement whenever possible and especially in instances where results could be used for high-stakes purposes (AERA et al. 2014; Graham et al., 2012). IRR needs to be

addressed when states are designing and implementing teacher evaluation systems and formative assessment processes involving raters. This can occur through a training certification process leading to IRR certification. Another option could be to require multiple raters conduct co-observations over a certain number of observations or ratings situations. For instance, if a teacher is observed five times a year there could be at least one observation that is conducted by multiple raters. This is something that could be implemented in the short-term. The majority of teacher evaluation processes and formative assessment processes involve a single rater. This makes it impossible to compute or report a reliability measure.

While Kappa may be the most widely known and calculated coefficient across IRR studies (Xie, 2013), this study illustrates that all other coefficients were more consistent and stable under the conditions examined in this study. It is difficult to understand why Kappa is considered the standard IRR coefficient beyond reporting percentage agreement given the major flaws with this coefficient (Gwet, 2008; Xie, 2013). Other coefficients may not be readily available across statistical software packages, however other coefficients can easily be programmed and calculated in certain software by the researcher.

The majority of the coefficients analyzed in this study are not pre-programmed in software, however the coefficients can be easily computed using R once the distribution of ratings is produced or provided. Kappa has three different benchmark classifications to help interpret reliability levels (Altman, 1991; Fleiss, 1981; Landis & Koch, 1977). While the classifications were not designed specifically for any other coefficient, they have been used to interpret coefficient values in other studies (Holcomb et al., 2022; Xie, 2013). In addition to being easily computed and interpretable, an IRR coefficient should be accurate and stable across various rating scale sizes, agreement levels, and ratings distributions. In the current study,

Lambda-1, *S*, and AC1 demonstrated these features. Finally, an IRR coefficient should be theory driven. Several coefficients used in this study were developed to overcome known limitations with the Kappa coefficient. Lambda-1 and Lambda-2 were developed according to rater-mediated assessment theory.

Lambda-1 and Lambda-2 are applicable across any rater-mediated assessment situation using an ordinal scale (Lambert et al., 2021). For example, these coefficients can be utilized to assess the reliability of scores from assessments made up of constructed response items or essays, as well as performance assessments involving a rater. Kappa is a coefficient that overcorrects, which is problematic because it makes it appear that raters are much poorer in their performance than they really are in practice (Holcomb et al., 2022; Jimenez & Zepeda, 2020; Wongpakaram et al., 2013).

The conditions where Kappa seems to dramatically overcorrect are not strange, esoteric conditions that do not happen in the field. In fact, they are conditions that often occur in practice when calculated in recent studies using empirical teacher evaluation data from professionally trained observers of early childhood educators through high school educators in multiple states (Holcomb et al., 2022; Jimenez & Zepeda, 2020). A misallocation of resources could have occurred had the Kappa coefficient been the sole reliability measure used in these studies. For example, the types of training and support that was needed for individual evaluators who are making rater judgments could have been completely misjudged. Whereas other, more stable, and consistent measures were available to indicate that the majority of raters were providing reliable evaluation scores for teachers (Lambert et al., 2021).

Ultimately, the issue with Kappa is theoretical, not just mathematical. Kappa was developed under the assumption that raters either guess when uncertain or the rater just happens

to agree with a standard rating because they are guessing (Cohen, 1960). However, this is not the case in rater-mediated assessments. Rater-mediated assessments operate on the expertise of raters. Lambda-1 and Lambda-2 are based on rater-mediated assessment theory (Lambert et al., 2021) and provide more solid ground than the Kappa coefficient and other alternatives that were designed to correct its statistical flaws. It is recommended that multiple measures of agreement are reported along with rater-mediated assessments, especially when the results could inform high-stakes decisions (AERA et al., 2014). The coefficients used in the current study are recommended for use with rating scales with five or fewer rating points (Graham et al., 2012). In any assessment involving human raters, multiple measures of reliability and agreement should be reported. Based on results from this study, which involved conditions with high agreement and various combinations of rating scale use, Lambda-1 and AC1 can be used with confidence. The use and application of Lambda-2 would be predicated on whether reliable estimates of the population proportion parameters were accessible. Jimenez & Zepeda (2020) recommended AC1 be used in place of Kappa when reporting the reliability of teacher evaluation scores. The current study confirmed this recommendation. The versions of Kappa and Krippendorff's Alpha evaluated in the current study should not be paired together as the sole measures of reliability. Caution should be taken with calculating and interpreting both of these coefficients and the *S* coefficient when ratings typically fall within a small range of possible categories, such as teacher evaluation instruments (Warrens, 2012; Weisberg et al., 2009). It is recommended if one of these coefficients is reported, the limitations of the coefficient are clearly stated, and the coefficient is paired with at least one other alternative measure of IRR.

Logistically, a single coefficient cannot overcome the barriers to implementing a high-quality IRR training and certification process. However, assurances should be made that raters

are providing valid and reliable ratings in rater-mediated assessment situations. It is important to have measures in place whether the rater-mediated assessment is an assessment involving essays, short-answer items, a developmental assessment, or a teacher evaluation instrument. All of these assessments rely on the expertise and training of raters to make quality and accurate scoring placements. Reporting reliability measures through common observations or during a rater training procedure should be required steps in educational assessments and evaluation systems. Similar to suggestions from other studies applying these coefficients in rater-mediated situations, the reliability of scores from these instruments should not be bypassed (Zepeda & Jimenez, 2019). Investigation of these coefficients through the current study and in the literature demonstrate the need for the development of policies that emphasize and require the assessment of the reliability of a measure while it is in use.

Since the adoption of RTT legislature, accountability measures have been put in place in teacher evaluation systems across the country. In many states, value-added measures and student achievement results on standardized tests are tied to teacher performance evaluations, compensation, and job retention (Bleiberg et al., 2021; Kraft & Gilmour, 2017; Murphy et al., 2013; Rodriguez & Hunter, 2021). Students, teachers, and administrators need to have confidence that scores from rater-mediated assessments are reliable, valid, and fair. Careful attention should be given to resources allocated to implement rater-mediated assessments. As states and school districts continue to dedicate funding and time to initiatives involving rater-mediated assessments, it is important that practices are in place to assure these resources are being put to best use. Again, a single measure of reliability cannot solve these problems. However, having a process in place that allows educational systems to assess the reliability of rater-mediated assessments is invaluable.

Directions for Further Research

This dissertation had clearly defined aims and objectives for utilizing a data generating mechanism according to specified conditions. A more traditional simulation study could be conducted to evaluate the performance of these chance-corrected agreement coefficients in situations with less restrictions on data condition parameters and fixed conditions (Burton et al., 2006). For example, future studies could investigate how well these indices estimate the true extent to which raters with varying characteristics agree with correct ratings. While the simulated data should capture real-world applications to provide meaningfulness to the results, the specifications can allow for more variation in condition choices and the inclusion of sampling from a universe of all possible rater behaviors. Capturing raters' reliability levels across a wider range of ratings, agreement conditions, and calculating the corresponding coefficient values across these variables is an area of research worth investigating for Lambda-1, Lambda-2, and the more researched IRR coefficients.

It is recommended that more investigation occurs with Lambda-2 in regard to evaluating the performance of the coefficient under the same conditions in the current study under varied population proportion rating allocations. The population proportion of ratings from NCTEP data were used as in example in this study. Other hypothetical or actual population proportions should be tested to see how Lambda-2 performs in comparison to other chance-corrected agreement coefficients.

REFERENCES

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis, 39*(1), 54-76. https://doi.org/10.3102/0162373716663646

Almy, S. (2011). *Fair to everyone: Building the balanced teacher evaluations that educators and students deserve.* Education Trust.

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.

Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System (EVAAS). *Educational Researcher, 37*(2), 65-75. https://doi.org/10.3102/0013189X08316420

Andrade, H. G. (2000). Using rubrics to promote thinking and learning. *Educational Leadership, 57*(5), 13-18.

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

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25.

Bitler, M., Corcoran, S. P., Domina, T., & Penner, E. K. (2021). Teacher effects on student achievement and height: A cautionary tale. *Journal of Research on Educational Effectiveness, 14*(4), 900-924. https://doi.org/10.1080/19345747.2021.1917025

Bleiberg, J., Brunner, E., Harbatkin, E., Kraft, M. A., & Springer, M. (2021). The effect of teacher evaluation on achievement and attainment: Evidence from statewide reforms. (EdWorkingPaper: 21-496). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/blak-r251

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education, 3,* 1-12. https://doi.org/10.3389/feduc.2018.00022

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review, 67*(3), 343-368. http://dx.doi.org/10.1080/00131911.2014.929565

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

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine, 25*, 4279 – 4292. https://doi.org/10.1002/sim.2673

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

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement, 75*(2), 311-337. https://doi.org/10.1177/0013164414539163

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*(3), 529-542. https://doi.org/10.1016/j.ecresq.2011.12.006

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., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher, 45*(6), 378-387. https://doi.org/10.3102/0013189X16659442

Darling-Hammond, L. (2017). *Developing and measuring higher order skills: Models for state performance assessment systems*. Learning Policy Institute and Council of Chief State School Officers.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8-15.

Daro, V., & Wei, R. C. (2015). How can teachers learn deeply? By scoring student assessments. *EdWeek*. Retrieved from: https://www.edweek.org/teaching-learning/opinion-how-can-teachers-learn-deeply-by-scoring-student-assessments/2015/06

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management, 34*(2), 267-297. https://doi.org/10.1002/pam.21818

Dowle, M. (2021). Package 'data.table'. R package version 1.14.2. https://cran.r-project.org/web/packages/data.table/index.html

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

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for all students: Development, implementation, and analysis* (pp. 261-287). Erlbaum.

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

Engelhard, G., Wang, J., & Wind, 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.

Engelhard, G., & Wind, S. (2019). Introduction to the special issue on rater-mediated assessments. *Journal of Educational Measurement, 56*(3), 475-477. https://doi.org/10.1111/jedm.12221

Firestone, W. A. (2014). Teacher evaluation policy and conflicting theories of motivation. *Educational Researcher, 43*(2), 100-107. https://doi.org/10.3102/0013189X14521864

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* (2nd ed.). John Wiley.

Garrett, R., & Steinberg, M. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis, 37*, 224-242.

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Washington D.C.: Center for Educator Compensation Reform, U.S. Department of Education.

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. https://doi.org/10.1177/0013164415596420

Haertel, E. H. (2013). *Reliability and validity of inferences about teachers based on student test scores.* Education Testing Service.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Heritage, M. (2013). *Formative assessment in practice: A process of inquiry and action.* Harvard Education Press.

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (2014). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record, 116*(1), 1-28.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56-64. https://doi.org/10.3102/0013189X12437203

Hill, H. C., & Herlihy, C. (2011). Prioritizing teaching quality in a new system of teacher evaluation. *Education Outlook*. Retrieved from http://www.aei.org/outlook/101089

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831. https://doi/org/10.3102/0002831210387916

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel* (Measures of Effective Teaching Project research paper). Bill and Melinda Gates Foundation.

Hogarth, R. (1987). *Judgement and choice: The psychology of decisions* (2nd ed.). John Wiley & Sons.

Holcomb, T. S., Lambert, R. G., & Bottoms, B. L. (in press). Reliability evidence for the North Carolina Teacher Evaluation Process using a variety of indicators of inter-rater agreement. *Journal of Educational Supervision*.

Jimenez, A. M., & Zepeda, S. J. (2020). A comparison of Gwet's AC1 and Kappa when calculating inter-rater reliability coefficients in a teacher evaluation context. *Journal of Education Human Resources, 38*(3), 290-300. https://doi.org/10.3138/jehr-2019-0001

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. Guilford Press.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*, 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Kendall, M. G. (1948). *Rank correlation methods.* Charles Griffin & Company.

Kim, J. (2021). Package 'kim'. R package version 0.4.21. https://github.com/jinkim3/kim

Kimball, S. M., & Milanowski, A. (2009). Examining teacher evaluation validity and leadership decision-making within a standards-based evaluation system. *Educational Administration Quarterly, 45*(1), 34-70. https://doi.org/10.1177/0013161X08327549

Koslow, R. (2018). Exact and adjacent inter-rater agreement associated with peer review of teaching. *Assessment Update, 30*(2), 1-3. https://doi.org/10.1002/au.30125

Kraft, M. A., & Gilmour, A. F. (2016). Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly, 52*(5), 711-753. https://doi.org/10.1177/0013161X16653445

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting *The Widget Effect*: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher, 46*(5), 234-249. https://doi.org/10.3102/0013189X17718797

Krippendorff, K. (1970). Estimating the reliability, systematic error, and random error of interval data. *Educational and Psychological Measurement, 30*(1), 61-70. https://doi.org/10.1177/001316447003000105

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology.* Sage.

Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. Retrieved from https://repository.upenn.edu/asc_papers/43

Lambert, R. G., Holcomb, T. S., & Bottoms, B. L. (2021). *Examining inter-rater reliability of evaluators judging teacher performance: Proposing an alternative to Cohen's Kappa* [Paper presentation]. National Council on Measurement in Education (NCME) 2021 Annual Conference, Virtual.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174. https://doi.org/10.2307/2529310

Lane, S. (2019). Modeling rater response processes in evaluating score meaning. *Journal of Educational Measurement, 56*(3), 653-663. https://doi.org/10.1111/jedm.12229

Lane, S., & Stone, C. (2006). Performance assessment. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387-431). American Council on Education and Praeger.

Lewis, J. M., Reid, D. B., Bell, C. A., Jones, N., & Qi, Y. (2020). The mantle of agency: Principals' use of teacher evaluation policy. *Leadership and Policy in Schools*, (pre-print), 1-17. https://doi.org/10.1080/15700763.2020.1770802

Ludbrook, J. (2002). Statistical techniques for comparing measures and methods of measurement: A critical review. *Cleinical and Experimental Pharmacology and Physiology, 29*(7), 527-536.

Lunenberg, F. C. (2012). Performance appraisal: Methods and rating errors. *International Journal of Scholarly Academic Intellectual Diversity, 14*(1), 1-9.

Maier, A., Adams, J., Burns, D., Kaul, M., Saunders, M., & Thompson, C. (2020). *Using performance assessments to support student learning: How district initiatives can make a difference*. Learning Policy Institute.

McClarty, K. L., Way, W. D., Porter, A. C., Beimers, J. N, & Miles, J. A. (2013) Evidence-based standard setting: Establishing a validity framework for cut scores. *Educational Researcher, 42*(2), 78-88. https://doi.org/10.3102/0013189X12470855

McHugh, M. L. (2012). Interrater reliability: The Kappa statistic. *Biochemia Medica, 22*(3), 276-282.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*, 2074-2102. https://doi.org/10.1002/stm.8086

Moskal, B. M. (2000). Scoring rubrics: What, when and how? *Practical Assessment, Research & Evaluation, 7*(3), 1-6. https://doi.org/10.7275/a5vq-7q66

Murphy, J., Hallinger, P., & Heck, R. H. (2013) Leading via teacher evaluation: The case of the missing clothes? *Educational Researcher, 42*(6), 349-354. https://doi.org/10.3102/0013189X13499625

Myford, C. M. (2012). Rater cognition research: Some possible directions for the future. *Educational Measurement: Issues and Practice, 31*(3), 48-49. https://doi.org/10.1111/j.1745-3992.2012.00243.x

National Association for the Education of Young Children & National Association of Early Childhood Specialists in State Departments of Education. (2003). Early childhood

curriculum, assessment, and program evaluation: Building an effective, accountable system in programs for children birth through age eight. A Joint Position Statement of the NAEYC and the NAECS/SDE. Washington, D.C.: Author.

National Council on Measurement in Education (2018). *Position Statement on K-12 Classroom Assessment.* A Position Statement of the NCME. Washington, D.C.: Author.

Nijveldt, M., Beijaard, D., Brekelmans, M., Wubbels, T., & Verloop, N. (2009). Assessors' perceptions of their judgement processes: Successful strategies and threats underlying valid assessment of student teachers. *Studies in Educational Evaluation, 35*(1), 29-36.

Oleckno, W. (2008). *Epidemiology: Concepts and Methods.* Waveland Press, Inc.

Papay, J. P. (2010). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193. https://doi.org/10.3102/0002831210362589

Peterson, C. H., Schulz, E. M., & Engelhard, G. (2011). Reliability and validity of bookmark-based methods for standard-setting: Comparisons of Angoff-based methods in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 30*(2), 3-14. https://doi.org/10.1111/j.1745-3992.2011.00200.x

Qi, Y., Bell, C. A., Jones, N. D., Lewis, J. M., Witherspoon, M. W., & Redash, A. (2018). *Administrators' uses of teacher observation protocol in different rating contexts* (Research Report No. RR-18-18). Educational Testing Service. https://doi.org/10.1002/ets2.12205

Raadt, A., Warrens, M. J., Bokser, R. J., & Kiers, H. A. L. (2021). A comparison of reliability coefficients for ordinal rating scales. *Journal of Classification, 38*, 519-543. https://doi.org/10.1007/s00357-021-09386-5

Rockoff, J., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review, 100*, 261-266.

Rodriguez, L. A., & Hunter, S. B. (2021). Making do: Why do administrators retain low-performing teachers? *Educational Researcher, 50*(9), 673-676. https://doi.org/10.3102/0013189X211039450

Sartain, L., Stoelinga, S., & Krone, E. (2010). *Rethinking teacher evaluation: Findings from the First Year of the Excellence in Teaching Project in Chicago Public Schools*. University of Chicago Consortium on School Research.

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

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Steinberg, M., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis, 38*(2), 293-317. https://doi.org/10.3102/0162373715616249

Steinberg, M., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy, 10*, 535-572.

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29-49). Sage.

Stevens, M. R., Lyles, W., & Berke, P. R. (2014). Measuring and reporting intercoder reliability in plan quality evaluation research. *Journal of Planning Education and Research, 34*(1), 77-93. https://doi.org/10.1177/0739456X13513614

Suto, I. (2012). A critical review of some qualitative research methods used to explore rater cogntion. *Educational Measurement: Issues and Practice, 31*(3), 21-30. https://doi.org/10.1111/j.1745-3992.2012.00240.x

Trumball, E., & Lash, A. (2013). *Understanding formative assessment: Insights from learning theory and measurement theory*. WestEd.

Uebersax, J. (2002). Kappa coefficients: A critical appraisal. Retrieved from https://www.john-uebersax.com/stat/kappa.htm

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

Walsh, P., Thornton, J., Asato, J., Walker, N., McCory, G., Baal, Jo., Baal, Je., Mendoza, N., & Banimahd, F. (2014). Approaches to describing inter-rater reliability of the overall clinical appearance of febrile infants and toddlers in the emergency department. *PeerJ, 2*, 1-19. https://doi.org/10.7717/peerj.651

Wang, J., Engelhard, G., Raczynski, K., Song, T., & Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing, 33*, 36-47. https://doi.org/10.106/j.asw.2017.03.003

Warrens, M. J. (2012). The effect of combining categories on Bennett, Alpert, and Goldstein's *S*. *Statistical Methodology, 9*, 341-352. https://doi.org/10.1016/j.stamet.2011.09.001

Warrens, M. J. (2014). Power weighted versions of Bennett, Alpert, and Goldstein's *S*. *Journal of Mathematics, 2014*, 1-9. http://dx.doi.org/10.1155/2014/231909

Wei, T., & Simko, V. (2021). Package ‘corrplot’. R package version 0.92. https://github.com/taiyun/corrplot

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New Teacher Project.

Whitehurst, G., Chingos, M., & Lindquist, K. (2014). *Evaluating teachers with classroom observations: Lessons learned in four districts*. Brown Center on Education Policy at Brookings.

Wickham, H. (2021). Package ‘tidyverse’. R package version 1.3.1. https://tidyverse.tidyverse.org/

Wickham, H. (2022). Package ‘dplyr’. R package version 1.0.8. https://dplyr.tidyverse.org/

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D. (2021). Package ‘ggplot2’. R package version 3.3.5. https://cran.r-project.org/web/packages/ggplot2/index.html

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Taylor & Francis Group.

Wind, S. A. (2019). Nonparametric evidence of validity, reliability, and fairness for rater-mediated assessments: An illustration using Mokken scale analysis. *Journal of Educational Measurement, 56*(3), 478-504. https://doi.org/10.1111/jedm.12222

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 raters*. In proceedings of the Federal Committee on Statistical Methodology Research Conference, The Council of Professional Associations on Federal Statistics, Washington, D.C.

Zepeda, S. J., & Jimenez, A. M. (2019). Teacher evaluation and reliability: Additional insights gathered from inter-rater reliability analyses. *Journal of Educational Supervision, 2*(2), 11-26. https://doi.org/10.31405/jes.2.2.2

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

APPENDIX A

AGREEMENT MATRICES WITH PROBABILITY OF A CORRECT GUESS CALCULATIONS (p_e)

Agreement matrices in Appendix A were adapted from Lambert et al., 2021 for the current study.

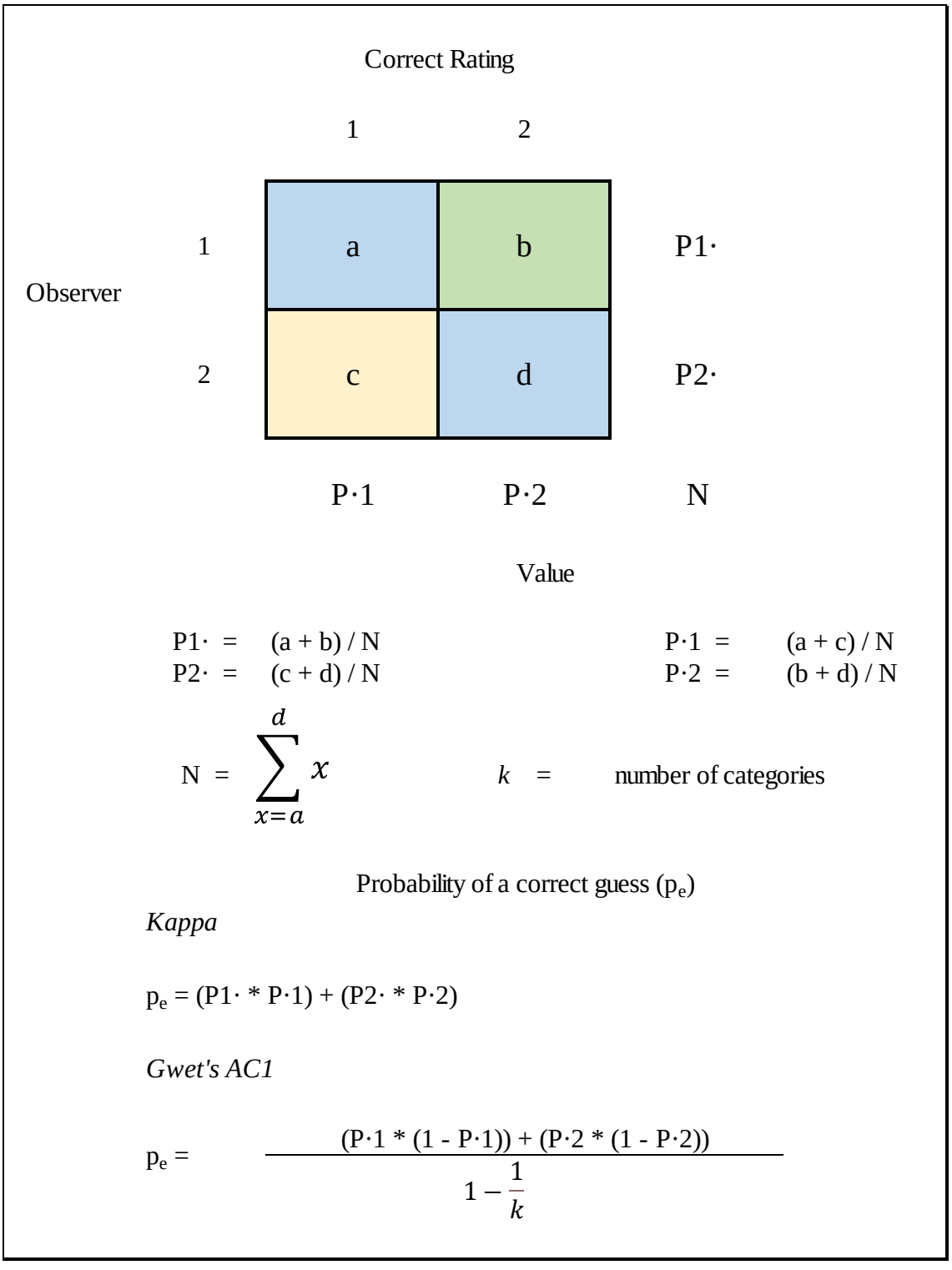


Figure A1. 2x2 agreement matrix probability distribution with Kappa and Gwet's AC1 probability of a correct guess (p_e) calculations.

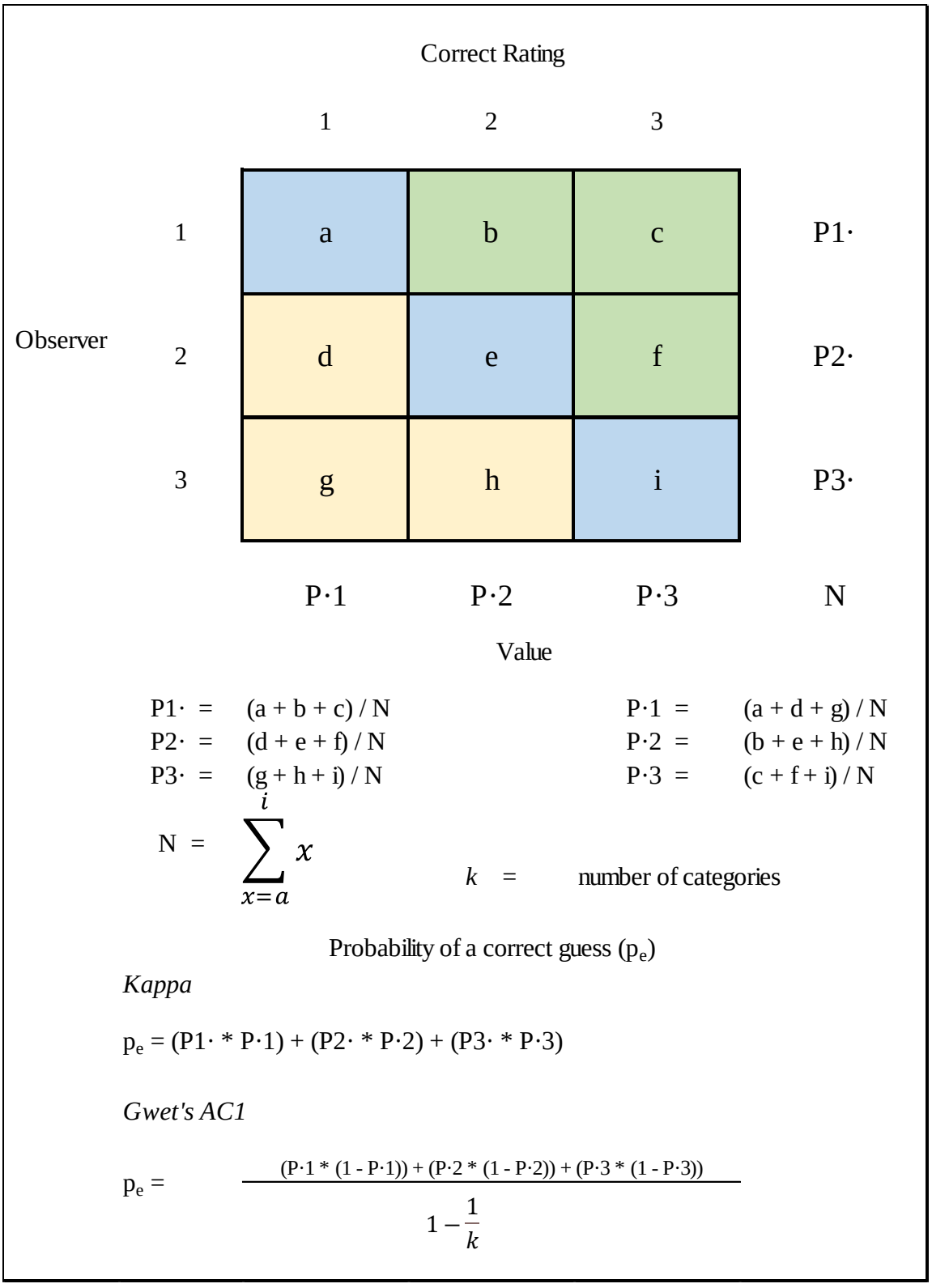


Figure A2. 3x3 agreement matrix probability distribution with Kappa and Gwet's AC1 probability of a correct guess (p_e) calculations.

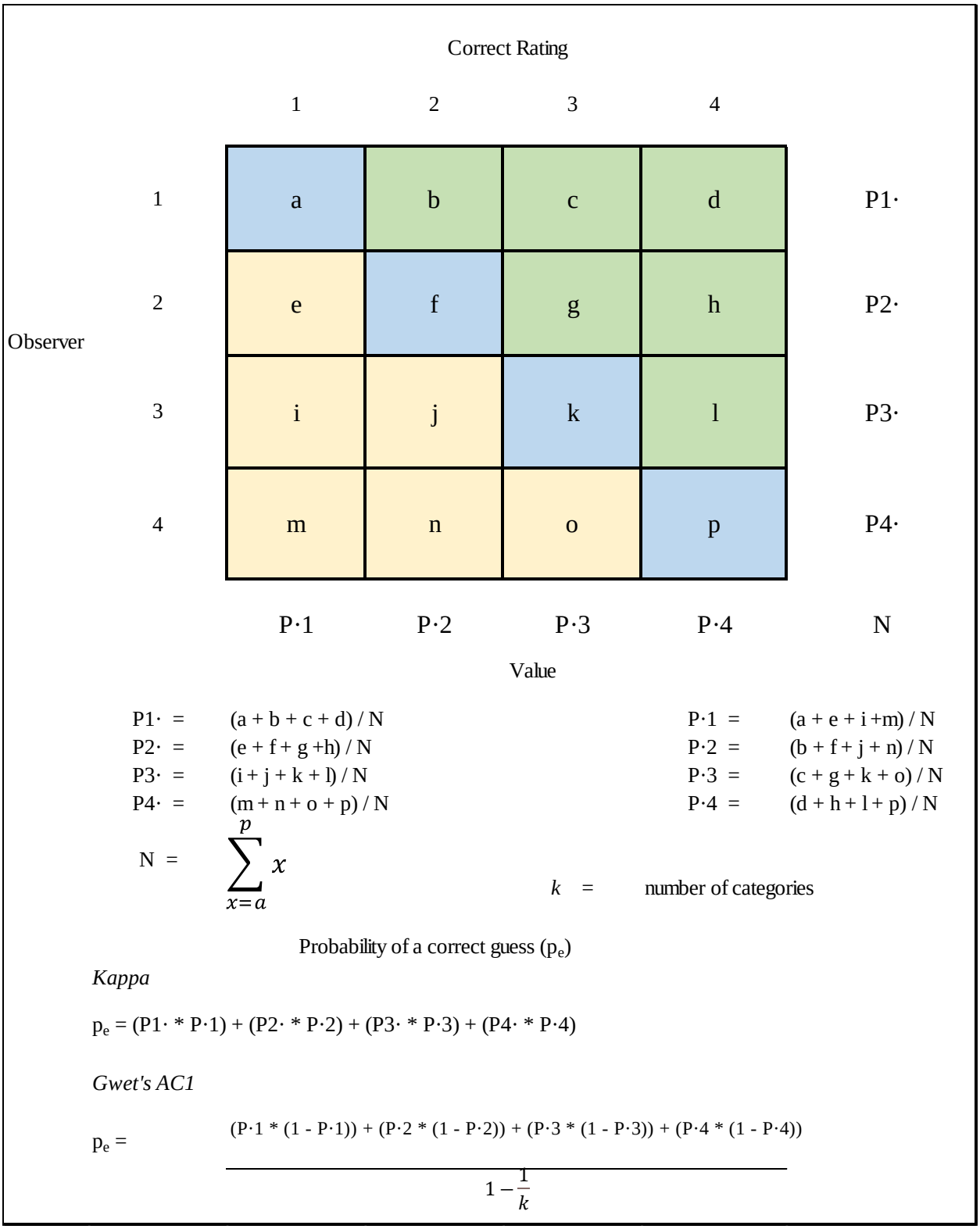


Figure A3. 4x4 agreement matrix probability distribution with Kappa and Gwet's AC1 probability of a correct guess (p_e) calculations.

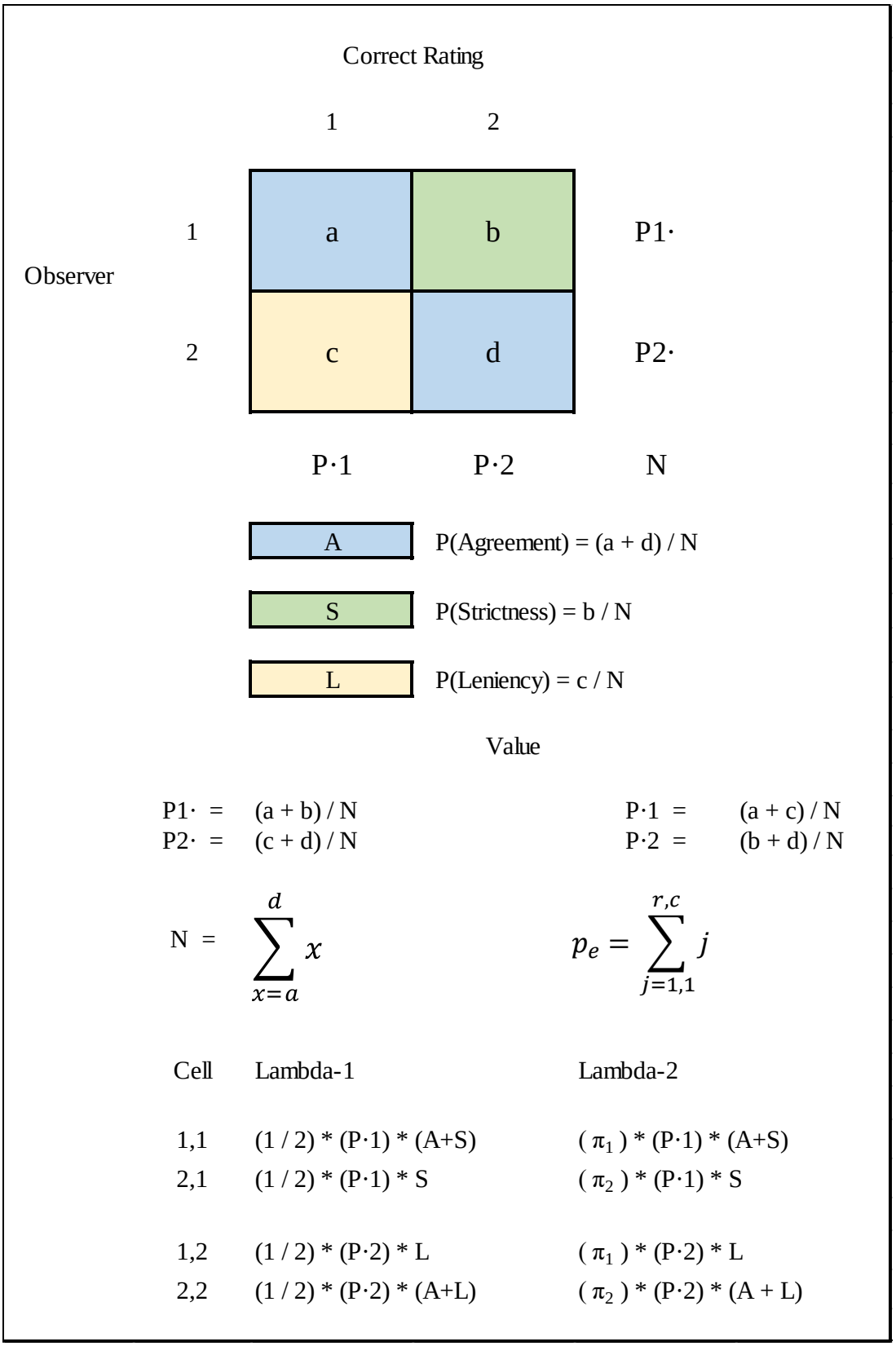


Figure A4. 2x2 agreement matrix probability distribution and Lambda-1 and -2 calculations.

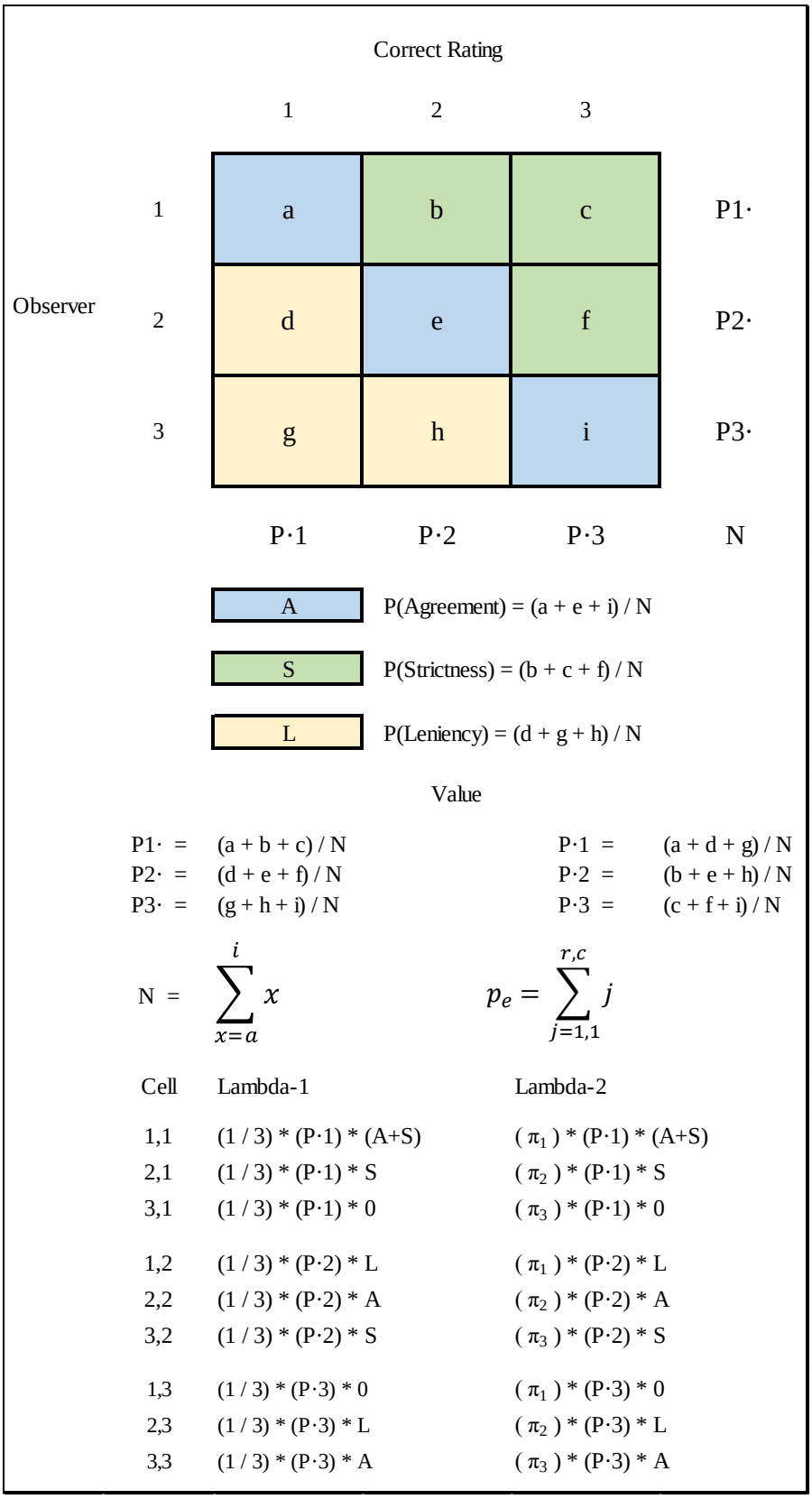


Figure A5. 3x3 agreement matrix probability distribution and Lambda-1 and -2 calculations.

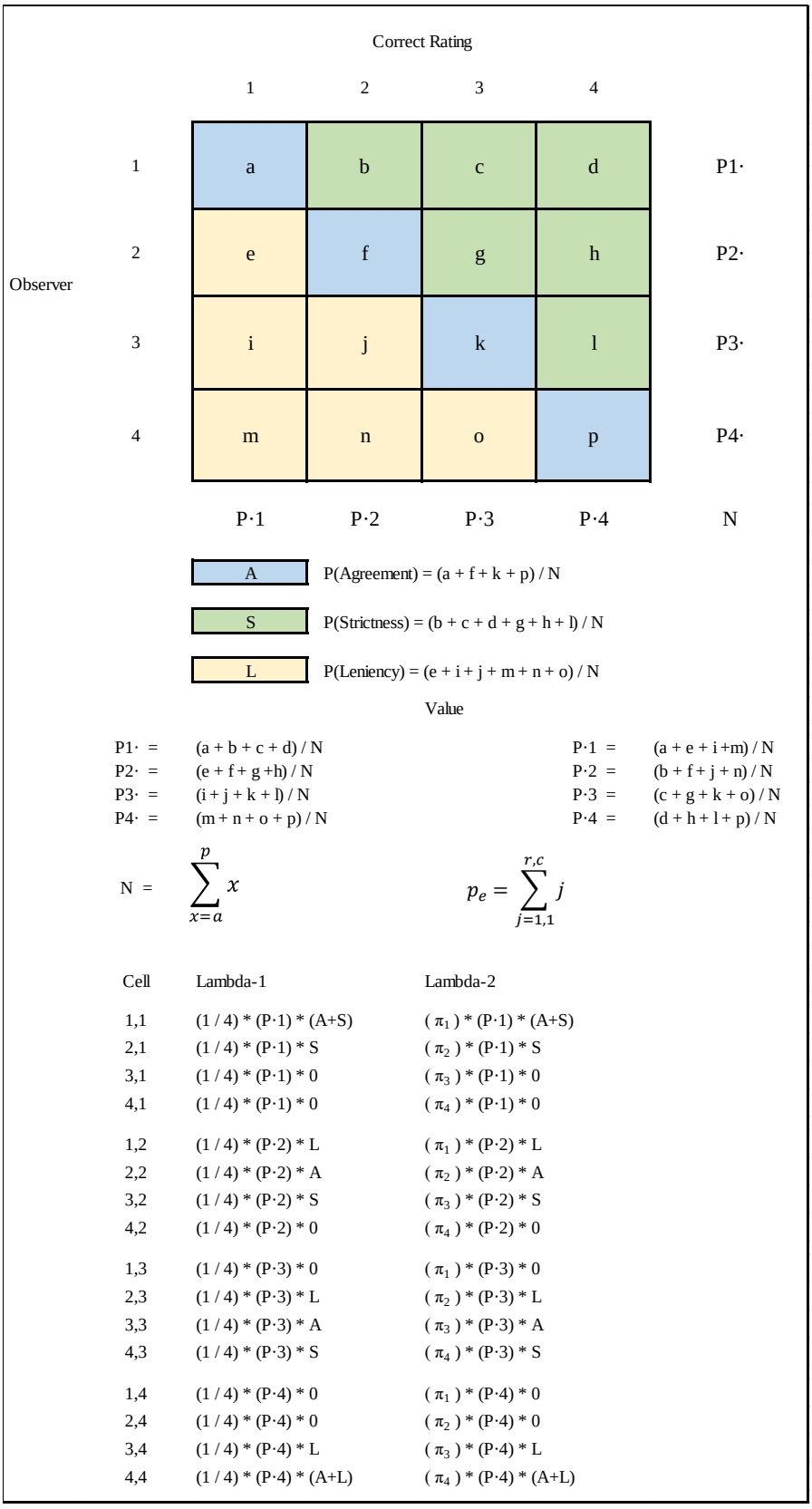


Figure A6. 4x4 agreement matrix probability distribution and Lambda-1 and -2 calculation

APPENDIX B

AGREEMENT MATRICES WITH COEFFICIENT CALCULATIONS

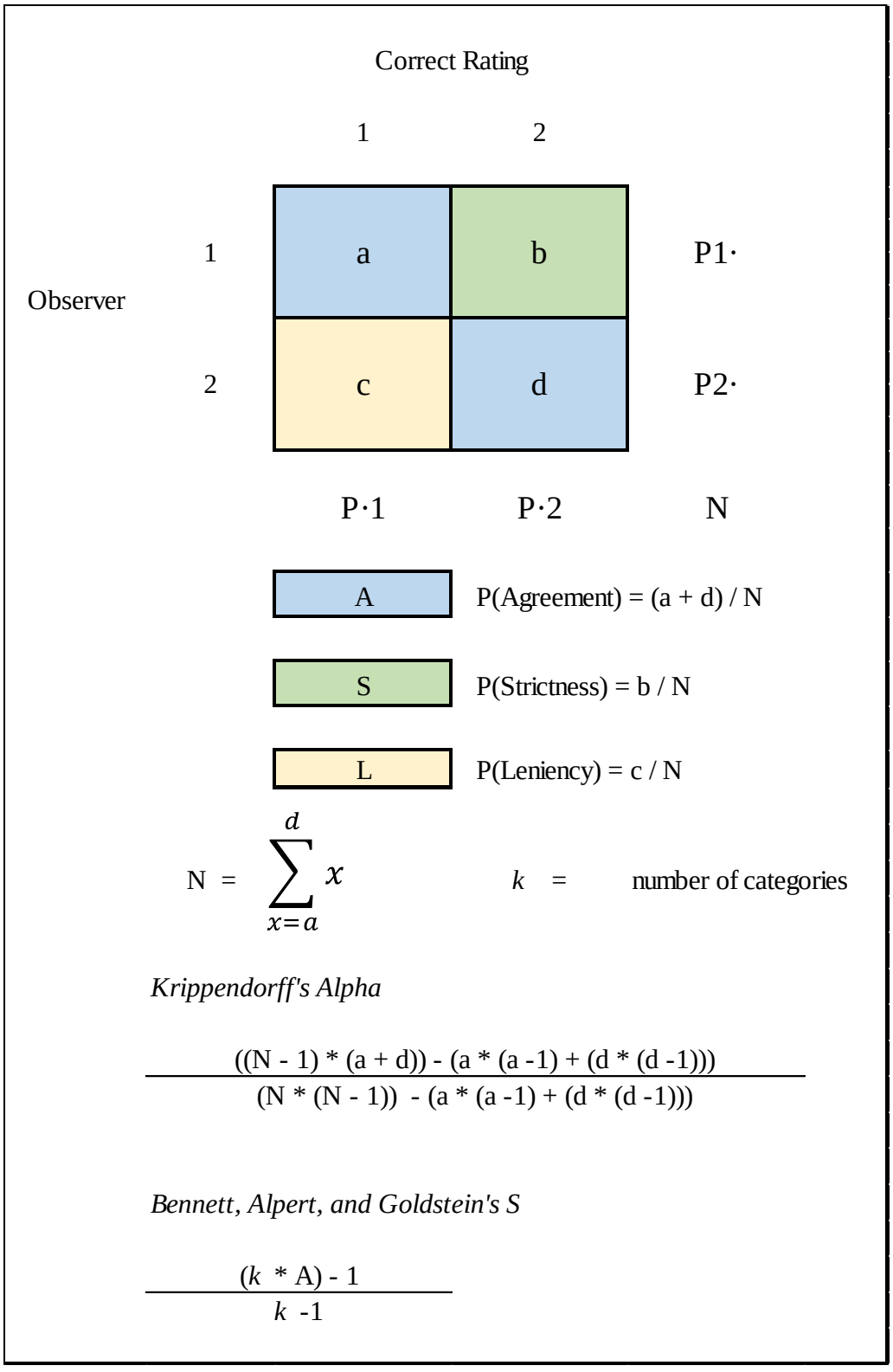


Figure B1. 2x2 agreement matrix probability distribution and Krippendorff's Alpha and Bennett, Alpert, and Goldstein's *S* calculation.

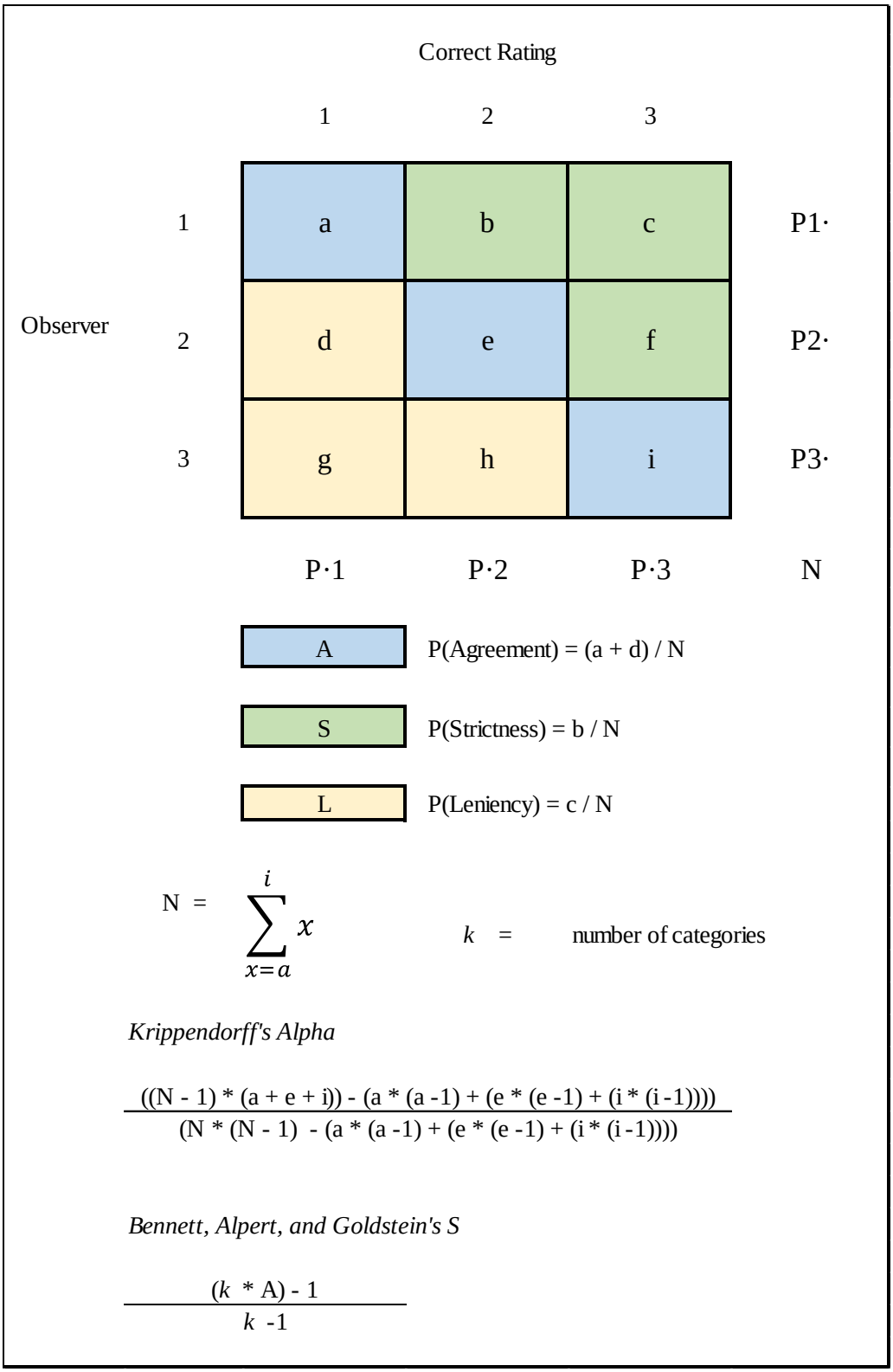


Figure B2. 3x3 agreement matrix probability distribution and Krippendorff's Alpha and Bennett, Alpert, and Goldstein's *S* calculation.

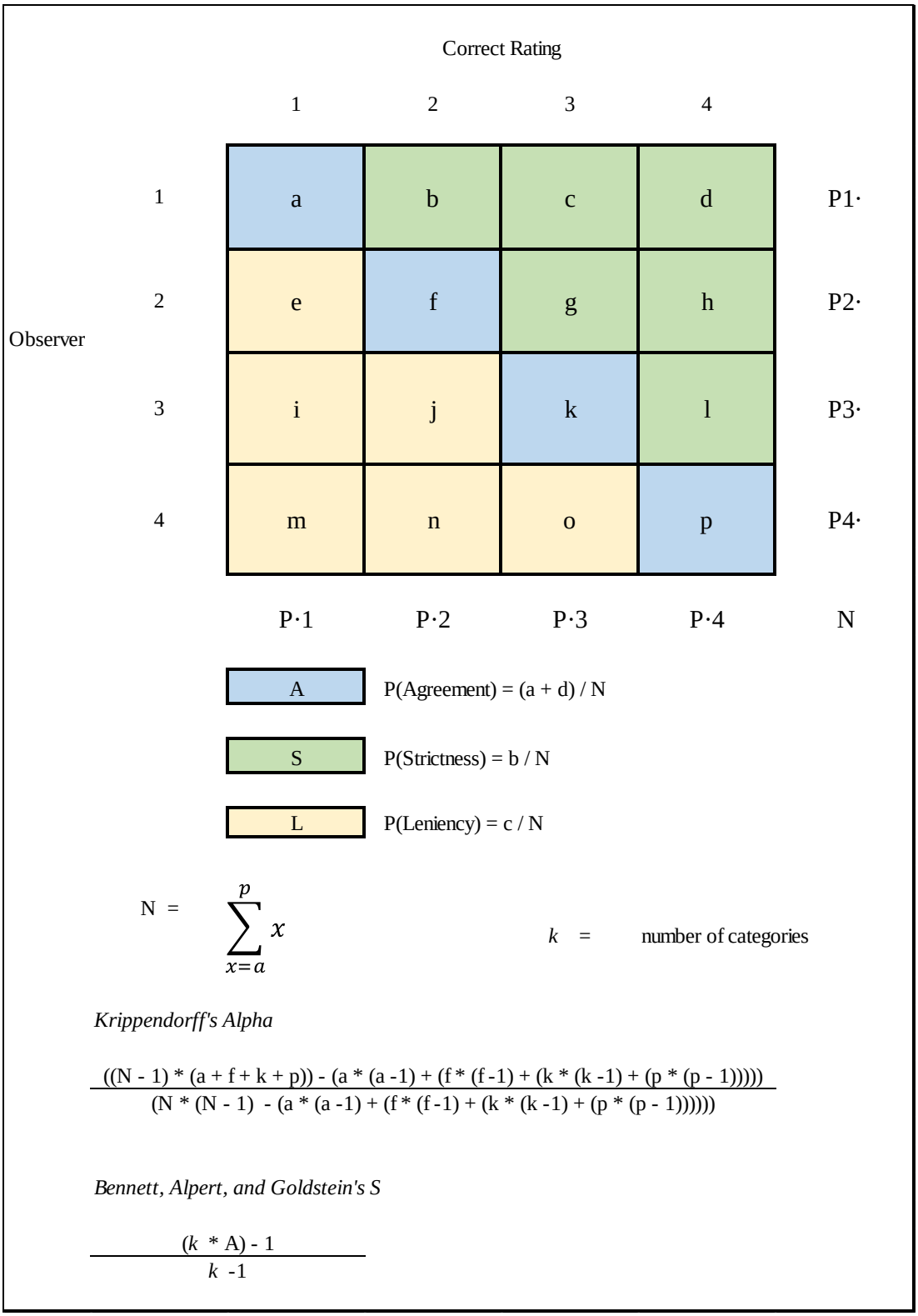


Figure B3. 4x4 agreement matrix probability distribution and Krippendorff's Alpha and Bennett, Alpert, and Goldstein's *S* calculation.

APPENDIX C

EXAMPLE R SCRIPTS

C.1 "2x2 R Code"

```
###Set Working Directory###
setwd("")

###Generate Data for Data Condition###
###adjust cell values as specified in data condition table###
library(data.table)
df <- expand.grid(a = 0:95,b = 5, c = 0,
                  d = 0:95)
setDT(df)
df[, Sum := a + b + c + d]

###Add Data involved in calculating expected chance agreement for each coefficient###
###Adjust Prev. Index###

df[, P.1 := ((a + c) / 100)]
df[, P.2 := ((b + d) / 100)]
df[, P1. := ((a + b) / 100)]
df[, P2. := ((c + d) / 100)]
df[, L1a := ((0.5) * (P.1)* ((a + b + d)/100))]
df[, L1b := ((0.5) * (P.2)* ((c)/100))]
df[, L1c := ((0.5) * (P.1)* ((b)/100))]
df[, L1d := ((0.5) * (P.2)* ((a + c + d)/100))]
df[, L1CAC := (L1a + L1b + L1c + L1d)]
df[, L2a := ((.05)*(P.1)*(.95+.05))]
df[, L2b := ((.05)*(P.2)*(.05))]
df[, L2c := ((.95)*(P.1)*(.05))]
df[, L2d := ((.95)*(P.2)*(.95+.00))]
df[, L2CAC := (L2a + L2b + L2c + L2d)]
df[, AC1Pe := (((P.1*(1-P.1))+(P.2*(1-P.2)))/(1-(1/2)))/2]
df[, Kripp.Numer. := (((a+b+c+d)-1)*(a+d))-(a*(a-1)+(d*(d-1)))]
df[, Kripp.Denom. := (((a+b+c+d)*((a+b+c+d)-1))-(a*(a-1)+(d*(d-1))))]
df[, Pe := ((P1. * P.1) + (P2. * P.2))]
df[, Prev.Index := (a - d)/100]

###Calculate Coefficients###
###change agreement level in calculation of S###
df[, Lambda1 := (((a+d)/100)- L1CAC)/(1-L1CAC)]
df[, Lambda2 := (((a+d)/100)- L2CAC)/(1-L2CAC)]
df[, Kappa := ((((a + d)/100)-Pe)/(1-Pe))]
df[, S := (((2*(.75))-1)/(2-1))]
df[, AC1 := (((a+d)/100)- AC1Pe)/(1-AC1Pe)]
df[, Kripp.Alpha := Kripp.Numer./Kripp.Denom.]

###Add specification for data generation combination that proportions must equal 100###
```

```
df[Sum == 100]

###Write Data frame to CSV Change name to DC# ###
write.csv(df[Sum == 100],"C:/Users/Holcomb Family/Dropbox/My PC (DESKTOP-
UET9RTV)/Desktop/Scott/R/2x2/DC15.csv", row.names = FALSE)

###Read CSV file back to R for data analysis###
###Use .csv2 for large files###
my_data <- read.csv("DC1.csv")
class(my_data)

###Retain columns with Coefficient Values within data set###
df <- my_data[ -c(1,1:24) ]
df

###summary statistics###
library(tidyverse)
CAC <- df %>% select(Lambda1, Lambda2, Kappa, S, AC1, Kripp.Alpha)
summary(CAC)

###Get mean of multiple columns using dplyr###
library(dplyr)
df %>% summarise_if(is.numeric, mean)
df %>% summarise_if(is.numeric, sd)

###Categorize coef. values###
Lambda1Range <- table(cut(my_data$Lambda1,breaks=seq.int(from=-1,to=1.0,by=0.2)))
Lambda2Range <- table(cut(my_data$Lambda2,breaks=seq.int(from=-1,to=1.0,by=0.2)))
KappaRange <- table(cut(my_data$Kappa,breaks=seq.int(from=-1,to=1.0,by=0.2)))
SRange <- table(cut(my_data$S,breaks=seq.int(from=-1,to=1.0,by=0.2)))
AC1Range <- table(cut(my_data$AC1,breaks=seq.int(from=-1,to=1.0,by=0.2)))
Kripp.AlphaRange <- table(cut(my_data$Kripp.Alpha,breaks=seq.int(from=-1,to=1.0,by=0.2)))

Lambda1Range
Lambda2Range
KappaRange
SRange
AC1Range
Kripp.AlphaRange

###Correlation###
###remove columns of coefficients without variance###
###Lambda1=25, Lambda2=26, Kappa=27, S=28, AC1=29, Kripp.Alpha=30###
df <- my_data[ -c(1,1:24, 28) ]

###install.packages("kim")###
###create corr. matrix table and send to file###
###remove Lambda1 when no variance###
library(kim)
correlation_matrix(data = df, var_names = c("Lambda2", "Kappa", "AC1", "Kripp.Alpha"), output_type = "rp")
write_csv(data = correlation_matrix(data = df, var_names = c("Lambda2", "Kappa", "AC1", "Kripp.Alpha"),
                                    output_type = "rp"), name = "corrmatrixDC1", timestamp = NULL)


###Correlation Visualization###
###install.packages("corrplot")###
```

```
library("corrplot")
CorrMat <- cor(df)
head(round(CorrMat,2))
pval <- psych::corr.test(CorrMat, adjust="none")$p

corrplot(CorrMat, type="upper", method="circle", p.mat=pval, insig="p-value", tl.pos="n", sig.level=0)
corrplot(CorrMat, type="lower", method="circle", add=T, tl.pos="d", cl.pos="n")
```

C.2 “3x3 R Code”

```
###Set Working Directory###
setwd("")

###Generate Data for Data Condition###
###adjust cell values as specified in data condition table###
library(data.table)
df <- expand.grid(a = 0:95, b = 5, c = 0,
                  d = 0, e = 0:95, f = 0,
                  g = 0, h = 0, i = 0)
setDT(df)
df[, Sum := a + b + c + d + e + f + g + h + i]

###Add Data involved in calculating expected chance agreement for each coefficient###
###Adjust Prev. Index###

df[, P.1 := ((a + d + g) / 100)]
df[, P.2 := ((b + e + h) / 100)]
df[, P.3 := ((c + f + i) / 100)]
df[, P1. := ((a + b + c) / 100)]
df[, P2. := ((d + e + f) / 100)]
df[, P3. := ((g + h + i) / 100)]
df[, L1a := ((1/3) * (P.1)* ((a + e + i + b + c + f)/100))]
df[, L1b := ((1/3) * (P.2)* ((d + g + h)/100))]
df[, L1c := ((1/3) * (P.3)* (0))]
df[, L1d := ((1/3) * (P.1)* ((b + c + f)/100))]
df[, L1e := ((1/3) * (P.2)* ((a + e + i)/100))]
df[, L1f := ((1/3) * (P.3)* ((d + g + h)/100))]
df[, L1g := ((1/3) * (P.1)* (0))]
df[, L1h := ((1/3) * (P.2)* ((b + c + f)/100))]
df[, L1i := ((1/3) * (P.3)* ((a + e + i)/100))]
df[, L1CAC := (L1a + L1b + L1c + L1d + L1e + L1f + L1g + L1h + L1i)]
df[, L2a := ((.05)*(P.1)*((a + e + i + b + c + f)/100))]
df[, L2b := ((.05)*(P.2)*((d + g + h)/100))]
df[, L2c := ((.05)*(P.3)*(0))]
df[, L2d := ((.90)*(P.1)*((b + c + f)/100))]
df[, L2e := ((.90)*(P.2)*((a + e + i)/100))]
df[, L2f := ((.90)*(P.3)*((d + g + h)/100))]
df[, L2g := ((.05)*(P.1)*(0))]
df[, L2h := ((.05)*(P.2)*((b + c + f)/100))]
df[, L2i := ((.05)*(P.3)*((a + e + i)/100))]
df[, L2CAC := (L2a + L2b + L2c + L2d + L2e + L2f + L2g + L2h + L2i)]
df[, AC1Pe := ((P.1*(1-P.1))+(P.2*(1-P.2))+((P.3*(1-P.3)))/(1-(1/3)))/3]
df[, Kripp.Numer. := ((((a + b + c + d + e + f + g + h + i)-1)*(a+e+i))-((a*(a-1)+(e*(e-1)+(i*(i-1))))))]
```

```
df[, Kripp.Denom. := ((((a + b + c + d + e + f + g + h + i)*((a + b + c + d + e + f + g + h + i)-1))-(a*(a-1)+(e*(e-1)+(i*(i-1))))))]
df[, Pe := ((P1. * P.1) + (P2. * P.2) + (P3. * P.3))]
df[, Prev.Index := (a - e)/100]

###Calculate Coefficients###
###change agreement level in calculation of S###
df[, Lambda1 := (((a+e+i)/100)- L1CAC)/(1-L1CAC)]
df[, Lambda2 := (((a+e+i)/100)- L2CAC)/(1-L2CAC)]
df[, Kappa := ((((a + e + i)/100)-Pe)/(1-Pe))]
df[, S := (((3*(.75))-1)/(3-1))]
df[, AC1 := (((a+e+i)/100)- AC1Pe)/(1-AC1Pe)]
df[, Kripp.Alpha := Kripp.Numer./Kripp.Denom.]

###Add specification for data generation combination that proportions must equal 100###
df[Sum == 100]

###Write Data frame to CSV Change name to DC# ###
write.csv(df[Sum == 100]," ", row.names = FALSE)

###Read CSV file back to R for data analysis###
###Use .csv2 for large files###
my_data <- read.csv("DC1.csv")
class(my_data)

###Retain columns with Coefficient Values within data set###
df <- my_data[ -c(1,1:41) ]
df

###summary statistics###
library(tidyverse)
CAC <- df %>% select(Lambda1, Lambda2, Kappa, S, AC1, Kripp.Alpha)
summary(CAC)

###Get mean of multiple columns using dplyr###
library(dplyr)
df %>% summarise_if(is.numeric, mean)
df %>% summarise_if(is.numeric, sd)


###Categorize coef. values###

Lambda1Range <- table(cut(my_data$Lambda1,breaks=seq.int(from=-1,to=1.0,by=0.2)))
Lambda2Range <- table(cut(my_data$Lambda2,breaks=seq.int(from=-1,to=1.0,by=0.2)))
KappaRange <- table(cut(my_data$Kappa,breaks=seq.int(from=-1,to=1.0,by=0.2)))
SRange <- table(cut(my_data$S,breaks=seq.int(from=-1,to=1.0,by=0.2)))
AC1Range <- table(cut(my_data$AC1,breaks=seq.int(from=-1,to=1.0,by=0.2)))
Kripp.AlphaRange <- table(cut(my_data$Kripp.Alpha,breaks=seq.int(from=-1,to=1.0,by=0.2)))

Lambda1Range
Lambda2Range
KappaRange
SRange
AC1Range
Kripp.AlphaRange
```

```
###Correlation###
###remove columns of coefficients without variance###
###Lambda1=42, Lambda2=43, Kappa=44, S=45, AC1=46, Kripp.Alpha=47###
df <- my_data[ -c(1,1:41, 45) ]

###install.packages("kim")###
library(kim)
correlation_matrix(data = df, var_names = c("Lambda1", "Lambda2", "Kappa", "AC1", "Kripp.Alpha"), output_type
= "rp")
write_csv(data = correlation_matrix(data = df, var_names = c("Lambda1", "Lambda2", "Kappa", "AC1",
"Kripp.Alpha"),
                                    output_type = "rp"), name = "corrmatrixDC1", timestamp = NULL)


###Correlation Visualization###
###install.packages("corrplot")###
library("corrplot")
CorrMat <- cor(df)
head(round(CorrMat,2))
pval <- psych::corr.test(CorrMat, adjust="none")$p

corrplot(CorrMat, type="upper", method="circle", p.mat=pval, insig="p-value", tl.pos="n", sig.level=0)
corrplot(CorrMat, type="lower", method="circle", add=T, tl.pos="d", cl.pos="n")

####other options for correlation visualizations###

corrplot(CorrMat, type="upper", method="color", p.mat=pval, insig="p-value", tl.pos="n", sig.level=0.05)
corrplot(CorrMat, type="lower", method="color", add=T, tl.pos="d", cl.pos="n")

corrplot(CorrMat, type="lower", method="circle", sig.level = 0.05, insig = "blank")
corrplot(CorrMat, type="lower", method="pie", sig.level = 0.05, insig = "blank")
corrplot(CorrMat, type="lower", method="color", sig.level = 0.05, insig = "blank")
corrplot(CorrMat, type="lower", method="number", sig.level = 0.05, insig = "blank")
```

C.3 “4x4 R Code”

```
###Set Working Directory###
setwd(" ")

###Generate Data for Data Condition###
###adjust cell values as specified in data condition table###
library(data.table)
df <- expand.grid(a = 0:95, b = 0, c = 0, d = 5,
                  e = 0, f = 0, g = 0, h = 0,
                  i = 0, j = 0, k = 0, l = 0,
                  m = 0, n = 0, o = 0, p = 0:95)
setDT(df)
df[, Sum := a + b + c + d + e + f + g + h + i + j + k + l + m + n + o + p]

###Add Data involved in calculating expected chance agreement for each coefficient###
###Adjust Prev. Index###

df[, P.1 := ((a + c + i + m) / 100)]
df[, P.2 := ((b + f + j + n) / 100)]
```

```
df[, P.3 := ((c + g + k + o) / 100)]
df[, P.4 := ((d + h + l + p) / 100)]
df[, P1. := ((a + b + c + d) / 100)]
df[, P2. := ((e + f + g + h) / 100)]
df[, P3. := ((i + j + k + l) / 100)]
df[, P4. := ((m + n + o + p) / 100)]
df[, L1a := ((1/4) * (P.1)* ((a + f + k + p + b + c + d + g + h + l)/100))]
df[, L1b := ((1/4) * (P.2)* ((e + i + j + m + n + o)/100))]
df[, L1c := ((1/4) * (P.3)* (0))]
df[, L1d := ((1/4) * (P.4)* (0))]
df[, L1e := ((1/4) * (P.1)* ((b + c + d + g + h + l)/100))]
df[, L1f := ((1/4) * (P.2)* ((a + f + k + p)/100))]
df[, L1g := ((1/4) * (P.3)* ((e + i + j + m + n + o)/100))]
df[, L1h := ((1/4) * (P.4)* (0))]
df[, L1i := ((1/4) * (P.1)* (0))]
df[, L1j := ((1/4) * (P.2)* ((b + c + d + g + h + l)/100))]
df[, L1k := ((1/4) * (P.3)* ((a + f + k + p)/100))]
df[, L1l := ((1/4) * (P.4)* ((e + i + j + m + n + o)/100))]
df[, L1m := ((1/4) * (P.1)* (0))]
df[, L1n := ((1/4) * (P.2)* (0))]
df[, L1o := ((1/4) * (P.3)* ((b + c + d + g + h + l)/100))]
df[, L1p := ((1/4) * (P.4)* ((a + f + k + p + e + i + j + m + n + o)/100))]
df[, L1CAC := (L1a + L1b + L1c + L1d + L1e + L1f + L1g + L1h + L1i + L1j + L1k + L1m + L1n + L1o + L1p)]
df[, L2a := ((.05)*(P.1)*((a + f + k + p + b + c + d + g + h + l)/100))]
df[, L2b := ((.05)*(P.2)*((e + i + j + m + n + o)/100))]
df[, L2c := ((.05)*(P.3)*(0))]
df[, L2d := ((.05)*(P.4)*(0))]
df[, L2e := ((.65)*(P.1)*((b + c + d + g + h + l)/100))]
df[, L2f := ((.65)*(P.2)*((a + f + k + p)/100))]
df[, L2g := ((.65)*(P.3)*((e + i + j + m + n + o)/100))]
df[, L2h := ((.65)*(P.4)*(0))]
df[, L2i := ((.25)*(P.1)*(0))]
df[, L2j := ((.25)*(P.2)*((b + c + d + g + h + l)/100))]
df[, L2k := ((.25)*(P.3)*((a + f + k + p)/100))]
df[, L2l := ((.25)*(P.4)*((e + i + j + m + n + o)/100))]
df[, L2m := ((.05)*(P.1)*(0))]
df[, L2n := ((.05)*(P.2)*(0))]
df[, L2o := ((.05)*(P.3)*((b + c + d + g + h + l)/100))]
df[, L2p := ((.05)*(P.4)*((a + f + k + p + e + i + j + m + n + o)/100))]
df[, L2CAC := (L2a + L2b + L2c + L2d + L2e + L2f + L2g + L2h + L2i + L2j + L2k + L2m + L2n + L2o + L2p)]
df[, AC1Pe := (((P.1*(1-P.1))+(P.2*(1-P.2))+(P.3*(1-P.3))+(P.4*(1-P.4)))/(1-(1/4)))/4]
df[, Kripp.Numer. := ((((a + b + c + d + e + f + g + h + i + j + k + l + m + n + o + p)-1)*(a+f+k+p))-((a*(a-1)+(f*(f-
1)+(k*(k-1)+(p*(p-1)))))))]
df[, Kripp.Denom. := ((((a + b + c + d + e + f + g + h + i + j + k + l + m + n + o + p)*((a + b + c + d + e + f + g + h +
i + j + k + l + m + n + o + p)-1))-(a*(a-1)+(f*(f-1)+(k*(k-1)+(p*(p-1)))))))]
df[, Pe := ((P1. * P.1) + (P2. * P.2) + (P3. * P.3) + (P4. * P.4))]
df[, Prev.Index := ((a + f) - (k + p))/100]

###Calculate Coefficients###
###change agreement level in calculation of S###
df[, Lambda1 := (((a+f+k+p)/100)- L1CAC)/(1-L1CAC)]
df[, Lambda2 := (((a+f+k+p)/100)- L2CAC)/(1-L2CAC)]
df[, Kappa := ((((a+f+k+p)/100)-Pe)/(1-Pe))]
df[, S := (((4*(.90))-1)/(4-1))]
df[, AC1 := (((a+f+k+p)/100)- AC1Pe)/(1-AC1Pe)]
df[, Kripp.Alpha := Kripp.Numer./Kripp.Denom.]
```

```
###Add specification for data generation combination that proportions must equal 100###
df[Sum == 100]

###Write Data frame to CSV Change name to DC# ###
write.csv1(df[Sum == 100],"C:/ ", row.names = FALSE)

###Read CSV file back to R for data analysis###
###Use .csv2 for large files###
my_data <- read.csv("DC1.csv")
class(my_data)

###Retain columns with Coefficient Values within data set###
df <- my_data[ -c(1,1:64) ]
df

###summary statistics###
library(tidyverse)
CAC <- df %>% select(Lambda1, Lambda2, Kappa, S, AC1, Kripp.Alpha)
summary(CAC)


###Get mean of multiple columns using dplyr###
library(dplyr)
df %>% summarise_if(is.numeric, mean)
df %>% summarise_if(is.numeric, sd)


###Categorize coef. values###

Lambda1Range <- table(cut(my_data$Lambda1,breaks=seq.int(from=-1,to=1.0,by=0.2)))
Lambda2Range <- table(cut(my_data$Lambda2,breaks=seq.int(from=-1,to=1.0,by=0.2)))
KappaRange <- table(cut(my_data$Kappa,breaks=seq.int(from=-1,to=1.0,by=0.2)))
SRange <- table(cut(my_data$S,breaks=seq.int(from=-1,to=1.0,by=0.2)))
AC1Range <- table(cut(my_data$AC1,breaks=seq.int(from=-1,to=1.0,by=0.2)))
Kripp.AlphaRange <- table(cut(my_data$Kripp.Alpha,breaks=seq.int(from=-1,to=1.0,by=0.2)))

Lambda1Range
Lambda2Range
KappaRange
SRange
AC1Range
Kripp.AlphaRange

###Correlation###
###remove columns of coefficients without variance###
###Lambda1=65, Lambda2=66, Kappa=67, S=68, AC1=69, Kripp.Alpha=70###
df <- my_data[ -c(1,1:64, 68) ]

###install.packages("kim")###
###change name of output file to match DC###
library(kim)
correlation_matrix(data = df, var_names = c("Lambda1", "Lambda2", "Kappa", "AC1", "Kripp.Alpha"), output_type
= "rp")
write_csv(data = correlation_matrix(data = df, var_names = c("Lambda1", "Lambda2", "Kappa", "AC1",
"Kripp.Alpha"), output_type = "rp"), name = "corrmatrixDC1", timestamp = NULL)
```

```
###Correlation Visualization###
###install.packages("corrplot")###
library("corrplot")
CorrMat <- cor(df)
head(round(CorrMat,2))
pval <- psych::corr.test(CorrMat, adjust="none")$p

corrplot(CorrMat, type="upper", method="circle", p.mat=pval, insig="p-value", tl.pos="n", sig.level=0)
corrplot(CorrMat, type="lower", method="circle", add=T, tl.pos="d", cl.pos="n")


####other options for correlation visualizations###

corrplot(CorrMat, type="upper", method="color", p.mat=pval, insig="p-value", tl.pos="n", sig.level=0.05)
corrplot(CorrMat, type="lower", method="color", add=T, tl.pos="d", cl.pos="n")

corrplot(CorrMat, type="lower", method="circle", sig.level = 0.05, insig = "blank")
corrplot(CorrMat, type="lower", method="pie", sig.level = 0.05, insig = "blank")
corrplot(CorrMat, type="lower", method="color", sig.level = 0.05, insig = "blank")
corrplot(CorrMat, type="lower", method="number", sig.level = 0.05, insig = "blank")
```

C.4 “Line Graph and Box Plot - R Code”

```
###Generate Line Graphs###
###Set Working Directory###
setwd("C:/Users/Holcomb Family/Dropbox/My PC (DESKTOP-UET9RTV)/Desktop/Scott/R/4x4")

###Read CSV file back to R for data analysis###
###Use .csv2 for large files###
my_data <- read.csv("DC53.csv")
class(my_data)

###Retain columns with Prev. Index and Coefficient Values within data set###
###2x2###
###data.frame <- my_data[ -c(1,1:23) ]###
###Prev.Index=24, Lambda1=25, Lambda2=26, Kappa=27, S=28, AC1=29, Kripp.Alpha=30###

###3x3###
###data.frame <- my_data[ -c(1,1:40) ]###
###Prev.Index = 41, Lambda1=42, Lambda2=43, Kappa=44, S=45, AC1=46, Kripp.Alpha=47###

###4x4###
###data.frame <- my_data[ -c(1,1:63) ]###
###Prev.Index = 64, Lambda1=65, Lambda2=66, Kappa=67, S=68, AC1=69, Kripp.Alpha=70###
data.frame <- my_data[ -c(1,1:63) ]

library(tidyr)

###Put data in long format###
###gather(data, renamed column, value, rangeFORcolumn, factor_key=TRUE)###
###adjust rangeFORcolumn if coefficients are removed###
data_long <- gather(data.frame, Coefficient, Value, Lambda1:Kripp.Alpha, factor_key=TRUE)
```

```
data_long

###install.packages("ggplot2")###
library(ggplot2)

###Line plots###
ggplot(data_long, aes(x = Prev.Index, y = Value, color = Coefficient)) +
  geom_line(linetype = 7,
            lwd = 1.1)

###Box plots###
ggplot(data_long, aes(x = Coefficient, y=Value, fill = Coefficient)) +
  geom_boxplot(alpha=0.3)
```

APPENDIX D

BENCHMARK AGREEMENT CLASSIFICATION TABLES (2X2 AGREEMENT MATRICES)

Table D.1

Count Across Benchmark Agreement Levels for Data Condition 1 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	96
Lambda-2	0	0	0	4	16	76
Kappa	0	2	2	6	16	70
S	0	0	0	0	0	96
AC1	0	0	0	0	0	96
Kripp. Alpha	0	0	0	4	14	78

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.2

Percentage of Values within Benchmark Levels for Data Condition 1 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	4.2%	16.7%	79.2%
Kappa	0.0%	2.1%	2.1%	6.3%	16.7%	72.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	4.2%	14.6%	81.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.3

Count Across Benchmark Agreement Levels for Data Condition 2 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	41	50
Lambda-2	1	3	5	11	31	40
Kappa	0	4	4	12	62	9
S	0	0	0	0	0	91
AC1	0	0	0	0	1	90
Kripp. Alpha	0	0	0	8	38	45

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.4

Percentage of Values within Benchmark Levels for Data Condition 2 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	45.1%	54.9%
Lambda-2	1.1%	3.3%	5.5%	12.1%	34.1%	44.0%
Kappa	0.0%	4.4%	4.4%	13.2%	68.1%	9.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	1.1%	98.9%
Kripp. Alpha	0.0%	0.0%	0.0%	8.8%	41.8%	49.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.5

Count Across Benchmark Agreement Levels for Data Condition 3 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	86	0
Lambda-2	7	5	8	15	47	4
Kappa	2	4	8	20	52	0
S	0	0	0	0	0	86
AC1	0	0	0	0	71	15
Kripp. Alpha	0	0	0	14	72	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.6

Percentage of Values within Benchmark Levels for Data Condition 3 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	8.1%	5.8%	9.3%	17.4%	54.7%	4.7%
Kappa	2.3%	4.7%	9.3%	23.3%	60.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	82.6%	17.4%
Kripp. Alpha	0.0%	0.0%	0.0%	16.3%	83.7%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.7

Count Across Benchmark Agreement Levels for Data Condition 4 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	31	50	0
Lambda-2	14	6	10	21	30	0
Kappa	2	6	12	42	19	0
S	0	0	0	0	81	0
AC1	0	0	0	1	80	0
Kripp. Alpha	0	0	0	20	61	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.8

Percentage of Values within Benchmark Levels for Data Condition 4 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	38.3%	61.7%	0.0%
Lambda-2	17.3%	7.4%	12.3%	25.9%	37.0%	0.0%
Kappa	2.5%	7.4%	14.8%	51.9%	23.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	1.2%	98.8%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	24.7%	75.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.9

Count Across Benchmark Agreement Levels for Data Condition 5 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	76	0	0
Lambda-2	20	7	13	26	10	0
Kappa	2	8	18	48	0	0
S	0	0	0	0	76	0
AC1	0	0	0	51	25	0
Kripp. Alpha	0	0	0	32	44	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.10

Percentage of Values within Benchmark Levels for Data Condition 5 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
Lambda-2	26.3%	9.2%	17.1%	34.2%	13.2%	0.0%
Kappa	2.6%	10.5%	23.7%	63.2%	0.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	67.1%	32.9%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	42.1%	57.9%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.11

Count Across Benchmark Agreement Levels for Data Condition 6 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	96
Lambda-2	0	0	0	0	15	81
Kappa	2	0	2	6	16	70
S	0	0	0	0	0	96
AC1	0	0	0	0	0	96
Kripp. Alpha	0	0	0	4	14	78

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.12

Percentage of Values within Benchmark Levels for Data Condition 6 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	15.6%	84.4%
Kappa	2.1%	0.0%	2.1%	6.3%	16.7%	72.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	4.2%	14.6%	81.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.13

Count Across Benchmark Agreement Levels for Data Condition 7 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	41	50
Lambda-2	0	0	0	10	31	50
Kappa	2	2	4	12	62	9
S	0	0	0	0	0	91
AC1	0	0	0	0	1	90
Kripp. Alpha	0	0	0	8	38	45

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.14

Percentage of Values within Benchmark Levels for Data Condition 7 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	45.1%	54.9%
Lambda-2	0.0%	0.0%	0.0%	11.0%	34.1%	54.9%
Kappa	2.2%	2.2%	4.4%	13.2%	68.1%	9.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	1.1%	98.9%
Kripp. Alpha	0.0%	0.0%	0.0%	8.8%	41.8%	49.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.15

Count Across Benchmark Agreement Levels for Data Condition 8 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	86	0
Lambda-2	0	0	5	15	47	19
Kappa	2	4	8	20	52	0
S	0	0	0	0	0	86
AC1	0	0	0	0	71	15
Kripp. Alpha	0	0	0	14	72	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.16

Percentage of Values within Benchmark Levels for Data Condition 8 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	5.8%	17.4%	54.7%	22.1%
Kappa	2.3%	4.7%	9.3%	23.3%	60.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	82.6%	17.4%
Kripp. Alpha	0.0%	0.0%	0.0%	16.3%	83.7%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.17

Count Across Benchmark Agreement Levels for Data Condition 9 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	31	50	0
Lambda-2	0	0	10	21	50	0
Kappa	2	6	12	42	19	0
S	0	0	0	0	81	0
AC1	0	0	0	1	80	0
Kripp. Alpha	0	0	0	20	61	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.18

Percentage of Values within Benchmark Levels for Data Condition 9 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	38.3%	61.7%	0.0%
Lambda-2	0.0%	0.0%	12.3%	25.9%	61.7%	0.0%
Kappa	2.5%	7.4%	14.8%	51.9%	23.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	1.2%	98.8%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	24.7%	75.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.19

Count Across Benchmark Agreement Levels for Data Condition 10 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	76	0	0
Lambda-2	0	2	13	26	35	0
Kappa	2	8	18	48	0	0
S	0	0	0	0	76	0
AC1	0	0	0	51	25	0
Kripp. Alpha	0	0	0	32	44	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.20

Percentage of Values within Benchmark Levels for Data Condition 10 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
Lambda-2	0.0%	2.6%	17.1%	34.2%	46.1%	0.0%
Kappa	2.6%	10.5%	23.7%	63.2%	0.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	67.1%	32.9%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	42.1%	57.9%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.21

Count Across Benchmark Agreement Levels for Data Condition 11 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	96
Lambda-2	0	0	0	2	15	79
Kappa	2	0	2	6	16	70
S	0	0	0	0	0	96
AC1	0	0	0	0	0	96
Kripp. Alpha	0	0	0	4	14	78

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.22

Percentage of Values within Benchmark Levels for Data Condition 11 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	2.1%	15.6%	82.3%
Kappa	2.1%	0.0%	2.1%	6.3%	16.7%	72.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	4.2%	14.6%	81.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.23

Count Across Benchmark Agreement Levels for Data Condition 12 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	91	0
Lambda-2	0	0	4	11	31	45
Kappa	2	2	6	10	71	0
S	0	0	0	0	0	91
AC1	0	0	0	0	1	90
Kripp. Alpha	0	0	0	8	38	45

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.24

Percentage of Values within Benchmark Levels for Data Condition 12 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	4.4%	12.1%	34.1%	49.5%
Kappa	2.2%	2.2%	6.6%	11.0%	78.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	1.1%	98.9%
Kripp. Alpha	0.0%	0.0%	0.0%	8.8%	41.8%	49.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.25

Count Across Benchmark Agreement Levels for Data Condition 13 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	86	0
Lambda-2	0	4	8	16	46	12
Kappa	2	4	10	20	50	0
S	0	0	0	0	0	86
AC1	0	0	0	0	70	16
Kripp. Alpha	0	0	0	14	72	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.26

Percentage of Values within Benchmark Levels for Data Condition 13 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	4.7%	9.3%	18.6%	53.5%	14.0%
Kappa	2.3%	4.7%	11.6%	23.3%	58.1%	0.0%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	81.4%	18.6%
Kripp. Alpha	0.0%	0.0%	0.0%	16.3%	83.7%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.27

Count Across Benchmark Agreement Levels for Data Condition 14 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	81	0	0
Lambda-2	4	6	10	21	40	0
Kappa	4	6	14	57	0	0
S	0	0	0	0	81	0
AC1	0	0	0	1	80	0
Kripp. Alpha	0	0	0	20	61	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.28

Percentage of Values within Benchmark Levels for Data Condition 14 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
Lambda-2	4.9%	7.4%	12.3%	25.9%	49.4%	0.0%
Kappa	4.9%	7.4%	17.3%	70.4%	0.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	1.2%	98.8%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	24.7%	75.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.29

Count Across Benchmark Agreement Levels for Data Condition 15 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	76	0	0
Lambda-2	7	8	13	26	22	0
Kappa	6	8	22	40	0	0
S	0	0	0	0	76	0
AC1	0	0	0	50	26	0
Kripp. Alpha	0	0	0	32	44	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table D.30

Percentage of Values within Benchmark Levels for Data Condition 15 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
Lambda-2	9.2%	10.5%	17.1%	34.2%	28.9%	0.0%
Kappa	7.9%	10.5%	28.9%	52.6%	0.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	65.8%	34.2%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	42.1%	57.9%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

APPENDIX E

BENCHMARK AGREEMENT CLASSIFICATION TABLES (3X3 AGREEMENT MATRICES)

Table E.1

Count Across Benchmark Agreement Levels for Data Condition 1 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	96
Lambda-2	0	0	0	0	15	81
Kappa	2	0	2	6	16	70
S	0	0	0	0	0	96
AC1	0	0	0	0	0	96
Kripp. Alpha	0	0	0	4	14	78

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.2

Percentage of Values within Benchmark Levels for Data Condition 1 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	15.6%	84.4%
Kappa	2.1%	0.0%	2.1%	6.3%	16.7%	72.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	4.2%	14.6%	81.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.3

Count Across Benchmark Agreement Levels for Data Condition 2 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	91
Lambda-2	0	0	0	10	37	44
Kappa	2	2	4	12	62	9
S	0	0	0	0	0	91
AC1	0	0	0	0	0	91
Kripp. Alpha	0	0	0	8	38	45

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.4

Percentage of Values within Benchmark Levels for Data Condition 2 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	11.0%	40.7%	48.4%
Kappa	2.2%	2.2%	4.4%	13.2%	68.1%	9.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	8.8%	41.8%	49.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.5

Count Across Benchmark Agreement Levels for Data Condition 3 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	86	0
Lambda-2	0	0	4	22	60	0
Kappa	2	4	8	20	52	0
S	0	0	0	0	86	0
AC1	0	0	0	0	0	86
Kripp. Alpha	0	0	0	14	72	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.6

Percentage of Values within Benchmark Levels for Data Condition 3 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	4.7%	25.6%	69.8%	0.0%
Kappa	2.3%	4.7%	9.3%	23.3%	60.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	16.3%	83.7%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.7

Count Across Benchmark Agreement Levels for Data Condition 4 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	81	0
Lambda-2	0	0	13	34	34	0
Kappa	2	6	12	42	19	0
S	0	0	0	0	81	0
AC1	0	0	0	0	81	0
Kripp. Alpha	0	0	0	20	61	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.8

Percentage of Values within Benchmark Levels for Data Condition 4 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	16.0%	42.0%	42.0%	0.0%
Kappa	2.5%	7.4%	14.8%	51.9%	23.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	24.7%	75.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.9

Count Across Benchmark Agreement Levels for Data Condition 5 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	26	50	0
Lambda-2	0	1	25	50	0	0
Kappa	0	2	8	18	48	0
S	0	0	0	0	76	0
AC1	0	0	0	0	76	0
Kripp. Alpha	0	0	0	32	44	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.10

Percentage of Values within Benchmark Levels for Data Condition 5 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	34.2%	65.8%	0.0%
Lambda-2	0.0%	1.3%	32.9%	65.8%	0.0%	0.0%
Kappa	0.0%	2.6%	10.5%	23.7%	63.2%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	42.1%	57.9%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.11

Count Across Benchmark Agreement Levels for Data Condition 6 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	96
Lambda-2	0	0	0	0	9	87
Kappa	2	0	2	6	16	70
S	0	0	0	0	0	96
AC1	0	0	0	0	0	96
Kripp. Alpha	0	0	0	4	14	78

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.12

Percentage of Values within Benchmark Levels for Data Condition 6 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	9.4%	90.6%
Kappa	2.1%	0.0%	2.1%	6.3%	16.7%	72.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	4.2%	14.6%	81.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.13

Count Across Benchmark Agreement Levels for Data Condition 7 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	91
Lambda-2	0	0	0	0	31	60
Kappa	2	2	4	12	62	9
S	0	0	0	0	0	91
AC1	0	0	0	0	0	91
Kripp. Alpha	0	0	0	8	38	45

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.14

Percentage of Values within Benchmark Levels for Data Condition 7 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	34.1%	65.9%
Kappa	2.2%	2.2%	4.4%	13.2%	68.1%	9.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	8.8%	41.8%	49.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.15

Count Across Benchmark Agreement Levels for Data Condition 8 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	86	0
Lambda-2	0	0	0	6	51	29
Kappa	2	4	8	20	52	0
S	0	0	0	0	86	0
AC1	0	0	0	0	0	86
Kripp. Alpha	0	0	0	14	72	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.16

Percentage of Values within Benchmark Levels for Data Condition 8 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	7.0%	59.3%	33.7%
Kappa	2.3%	4.7%	9.3%	23.3%	60.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	16.3%	83.7%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.17

Count Across Benchmark Agreement Levels for Data Condition 9 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	81	0
Lambda-2	0	0	0	14	67	0
Kappa	2	6	12	42	19	0
S	0	0	0	0	81	0
AC1	0	0	0	0	81	0
Kripp. Alpha	0	0	0	20	61	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.18

Percentage of Values within Benchmark Levels for Data Condition 9 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	17.3%	82.7%	0.0%
Kappa	2.5%	7.4%	14.8%	51.9%	23.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	24.7%	75.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.19

Count Across Benchmark Agreement Levels for Data Condition 10 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	76	0
Lambda-2	0	0	0	24	52	0
Kappa	2	8	18	48	0	0
S	0	0	0	0	76	0
AC1	0	0	0	0	76	0
Kripp. Alpha	0	0	0	32	44	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.20

Percentage of Values within Benchmark Levels for Data Condition 10 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	31.6%	68.4%	0.0%
Kappa	2.6%	10.5%	23.7%	63.2%	0.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	42.1%	57.9%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.21

Count Across Benchmark Agreement Levels for Data Condition 11 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	96
Lambda-2	0	0	0	0	9	87
Kappa	2	0	2	6	16	70
S	0	0	0	0	0	96
AC1	0	0	0	0	0	96
Kripp. Alpha	0	0	0	4	14	78

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.22

Percentage of Values within Benchmark Levels for Data Condition 11 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	9.4%	90.6%
Kappa	2.1%	0.0%	2.1%	6.3%	16.7%	72.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	4.2%	14.6%	81.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.23

Count Across Benchmark Agreement Levels for Data Condition 12 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	91
Lambda-2	0	0	0	0	31	60
Kappa	2	2	4	12	62	9
S	0	0	0	0	0	91
AC1	0	0	0	0	0	91
Kripp. Alpha	0	0	0	8	38	45

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.24

Percentage of Values within Benchmark Levels for Data Condition 12 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	34.1%	65.9%
Kappa	2.2%	2.2%	4.4%	13.2%	68.1%	9.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	8.8%	41.8%	49.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.25

Count Across Benchmark Agreement Levels for Data Condition 13 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	86	0
Lambda-2	0	0	0	6	51	29
Kappa	2	4	8	20	52	0
S	0	0	0	0	86	0
AC1	0	0	0	0	0	86
Kripp. Alpha	0	0	0	14	72	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.26

Percentage of Values within Benchmark Levels for Data Condition 13 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	7.0%	59.3%	33.7%
Kappa	2.3%	4.7%	9.3%	23.3%	60.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	16.3%	83.7%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.27

Count Across Benchmark Agreement Levels for Data Condition 14 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	81	0
Lambda-2	0	0	0	14	67	0
Kappa	2	6	12	42	19	0
S	0	0	0	0	81	0
AC1	0	0	0	0	81	0
Kripp. Alpha	0	0	0	20	61	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.28

Percentage of Values within Benchmark Levels for Data Condition 14 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	17.3%	82.7%	0.0%
Kappa	2.5%	7.4%	14.8%	51.9%	23.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	24.7%	75.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.29

Count Across Benchmark Agreement Levels for Data Condition 15 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	76	0
Lambda-2	0	0	0	24	52	0
Kappa	2	8	18	48	0	0
S	0	0	0	0	76	0
AC1	0	0	0	0	76	0
Kripp. Alpha	0	0	0	32	44	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.30

Percentage of Values within Benchmark Levels for Data Condition 15 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	31.6%	68.4%	0.0%
Kappa	2.6%	10.5%	23.7%	63.2%	0.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	42.1%	57.9%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.31

Count Across Benchmark Agreement Levels for Data Condition 16 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	96
Lambda-2	0	0	0	0	15	81
Kappa	2	0	2	6	16	70
S	0	0	0	0	0	96
AC1	0	0	0	0	0	96
Kripp. Alpha	0	0	0	4	14	78

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.32

Percentage of Values within Benchmark Levels for Data Condition 16 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	15.6%	84.4%
Kappa	2.1%	0.0%	2.1%	6.3%	16.7%	72.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	4.2%	14.6%	81.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.33

Count Across Benchmark Agreement Levels for Data Condition 17 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	91
Lambda-2	0	0	0	10	37	44
Kappa	2	2	4	12	62	9
S	0	0	0	0	0	91
AC1	0	0	0	0	0	91
Kripp. Alpha	0	0	0	8	38	45

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.34

Percentage of Values within Benchmark Levels for Data Condition 17 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	11.0%	40.7%	48.4%
Kappa	2.2%	2.2%	4.4%	13.2%	68.1%	9.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	8.8%	41.8%	49.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.35

Count Across Benchmark Agreement Levels for Data Condition 18 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	86	0
Lambda-2	0	0	4	21	61	0
Kappa	2	4	8	20	52	0
S	0	0	0	0	86	0
AC1	0	0	0	0	0	86
Kripp. Alpha	0	0	0	14	72	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.36

Percentage of Values within Benchmark Levels for Data Condition 18 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	4.7%	24.4%	70.9%	0.0%
Kappa	2.3%	4.7%	9.3%	23.3%	60.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	16.3%	83.7%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.37

Count Across Benchmark Agreement Levels for Data Condition 19 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	81	0
Lambda-2	0	0	13	33	35	0
Kappa	2	6	12	42	19	0
S	0	0	0	0	81	0
AC1	0	0	0	0	81	0
Kripp. Alpha	0	0	0	20	61	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.38

Percentage of Values within Benchmark Levels for Data Condition 19 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	16.0%	40.7%	43.2%	0.0%
Kappa	2.5%	7.4%	14.8%	51.9%	23.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	24.7%	75.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.39

Count Across Benchmark Agreement Levels for Data Condition 20 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	76	0
Lambda-2	0	1	24	49	2	0
Kappa	2	8	18	48	0	0
S	0	0	0	0	76	0
AC1	0	0	0	0	76	0
Kripp. Alpha	0	0	0	32	44	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.40

Percentage of Values within Benchmark Levels for Data Condition 20 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	1.3%	31.6%	64.5%	2.6%	0.0%
Kappa	2.6%	10.5%	23.7%	63.2%	0.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	42.1%	57.9%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.41

Count Across Benchmark Agreement Levels for Data Condition 21 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	4656
Lambda-2	0	0	0	0	74	4582
Kappa	1	0	4	22	183	4446
S	0	0	0	0	0	4656
AC1	0	0	0	0	0	4656
Kripp. Alpha	0	0	0	9	126	4521

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.42

Percentage of Values within Benchmark Levels for Data Condition 21 (2x2)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	1.6%	98.4%
Kappa	0.0%	0.0%	0.1%	0.5%	3.9%	95.5%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.2%	2.7%	97.1%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.43

Count Across Benchmark Agreement Levels for Data Condition 22 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	4186
Lambda-2	0	0	0	10	764	3412
Kappa	1	2	13	89	1184	2897
S	0	0	0	0	0	4186
AC1	0	0	0	0	0	4186
Kripp. Alpha	0	0	0	30	711	3445

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.44

Percentage of Values within Benchmark Levels for Data Condition 22 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.2%	18.3%	81.5%
Kappa	0.0%	0.0%	0.3%	2.1%	28.3%	69.2%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.7%	17.0%	82.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.45

Count Across Benchmark Agreement Levels for Data Condition 23 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	3741	0
Lambda-2	0	0	0	115	2509	1117
Kappa	1	5	34	248	3453	0
S	0	0	0	0	3741	0
AC1	0	0	0	0	493	3248
Kripp. Alpha	0	0	0	69	3396	276

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.46

Percentage of Values within Benchmark Levels for Data Condition 23 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	3.1%	67.1%	29.9%
Kappa	0.0%	0.1%	0.9%	6.6%	92.3%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	13.2%	86.8%
Kripp. Alpha	0.0%	0.0%	0.0%	1.8%	90.8%	7.4%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.47

Count Across Benchmark Agreement Levels for Data Condition 24 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	3321	0
Lambda-2	0	0	0	404	2917	0
Kappa	1	9	75	613	2623	0
S	0	0	0	0	3321	0
AC1	0	0	0	0	3321	0
Kripp. Alpha	0	0	0	165	3156	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.48

Percentage of Values within Benchmark Levels for Data Condition 24 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	12.2%	87.8%	0.0%
Kappa	0.0%	0.3%	2.3%	18.5%	79.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	5.0%	95.0%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.49

Count Across Benchmark Agreement Levels for Data Condition 25 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	49	2877	0
Lambda-2	0	0	14	1030	1882	0
Kappa	1	20	134	1873	898	0
S	0	0	0	0	2926	0
AC1	0	0	0	0	2926	0
Kripp. Alpha	0	0	0	372	2554	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.50

Percentage of Values within Benchmark Levels for Data Condition 25 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	1.7%	98.3%	0.0%
Lambda-2	0.0%	0.0%	0.5%	35.2%	64.3%	0.0%
Kappa	0.0%	0.7%	4.6%	64.0%	30.7%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	12.7%	87.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.51

Count Across Benchmark Agreement Levels for Data Condition 26 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	4656
Lambda-2	0	0	0	0	73	4583
Kappa	1	0	4	22	183	4446
S	0	0	0	0	0	4656
AC1	0	0	0	0	0	4656
Kripp. Alpha	0	0	0	9	126	4521

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.52

Percentage of Values within Benchmark Levels for Data Condition 26 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	1.6%	98.4%
Kappa	0.0%	0.0%	0.1%	0.5%	3.9%	95.5%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.2%	2.7%	97.1%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.53

Count Across Benchmark Agreement Levels for Data Condition 27 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	4186
Lambda-2	0	0	0	10	760	3416
Kappa	1	2	13	89	1184	2897
S	0	0	0	0	0	4186
AC1	0	0	0	0	0	4186
Kripp. Alpha	0	0	0	30	711	3445

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.54

Percentage of Values within Benchmark Levels for Data Condition 27 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.2%	18.2%	81.6%
Kappa	0.0%	0.0%	0.3%	2.1%	28.3%	69.2%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.7%	17.0%	82.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.55

Count Across Benchmark Agreement Levels for Data Condition 28 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	3741	0
Lambda-2	0	0	0	114	2477	1150
Kappa	1	5	34	248	3453	0
S	0	0	0	0	3741	0
AC1	0	0	0	0	489	3252
Kripp. Alpha	0	0	0	69	3396	276

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.56

Percentage of Values within Benchmark Levels for Data Condition 28 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	3.0%	66.2%	30.7%
Kappa	0.0%	0.1%	0.9%	6.6%	92.3%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	13.1%	86.9%
Kripp. Alpha	0.0%	0.0%	0.0%	1.8%	90.8%	7.4%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.57

Count Across Benchmark Agreement Levels for Data Condition 29 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	3321	0
Lambda-2	0	0	0	396	2925	0
Kappa	1	9	75	613	2623	0
S	0	0	0	0	3321	0
AC1	0	0	0	0	3321	0
Kripp. Alpha	0	0	0	165	3156	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.58

Percentage of Values within Benchmark Levels for Data Condition 29 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	11.9%	88.1%	0.0%
Kappa	0.0%	0.3%	2.3%	18.5%	79.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	5.0%	95.0%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.59

Count Across Benchmark Agreement Levels for Data Condition 30 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	2926	0
Lambda-2	0	0	14	1000	1912	0
Kappa	1	20	134	1873	898	0
S	0	0	0	0	2926	0
AC1	0	0	0	0	2926	0
Kripp. Alpha	0	0	0	372	2554	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.60

Percentage of Values within Benchmark Levels for Data Condition 30 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.5%	34.2%	65.3%	0.0%
Kappa	0.0%	0.7%	4.6%	64.0%	30.7%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	12.7%	87.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.61

Count Across Benchmark Agreement Levels for Data Condition 31 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	4656
Lambda-2	0	0	0	0	78	4578
Kappa	1	0	4	22	183	4446
S	0	0	0	0	0	4656
AC1	0	0	0	0	0	4656
Kripp. Alpha	0	0	0	9	126	4521

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.62

Percentage of Values within Benchmark Levels for Data Condition 31 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	1.7%	98.3%
Kappa	0.0%	0.0%	0.1%	0.5%	3.9%	95.5%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.2%	2.7%	97.1%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.63

Count Across Benchmark Agreement Levels for Data Condition 32 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	4186
Lambda-2	0	0	0	10	770	3406
Kappa	1	2	13	91	1192	2887
S	0	0	0	0	0	4186
AC1	0	0	0	0	0	4186
Kripp. Alpha	0	0	0	30	711	3445

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.64

Percentage of Values within Benchmark Levels for Data Condition 32 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.2%	18.4%	81.4%
Kappa	0.0%	0.0%	0.3%	2.2%	28.5%	69.0%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.7%	17.0%	82.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.65

Count Across Benchmark Agreement Levels for Data Condition 33 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	3741	0
Lambda-2	0	0	0	120	2508	1113
Kappa	1	5	34	258	3443	0
S	0	0	0	0	3741	0
AC1	0	0	0	0	494	3247
Kripp. Alpha	0	0	0	69	3396	276

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.66

Percentage of Values within Benchmark Levels for Data Condition 33 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	3.2%	67.0%	29.8%
Kappa	0.0%	0.1%	0.9%	6.9%	92.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	13.2%	86.8%
Kripp. Alpha	0.0%	0.0%	0.0%	1.8%	90.8%	7.4%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.67

Count Across Benchmark Agreement Levels for Data Condition 34 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	3321	0
Lambda-2	0	0	1	434	2886	0
Kappa	1	11	73	655	2581	0
S	0	0	0	0	3321	0
AC1	0	0	0	0	3321	0
Kripp. Alpha	0	0	0	165	3156	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.68

Percentage of Values within Benchmark Levels for Data Condition 34 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	13.1%	86.9%	0.0%
Kappa	0.0%	0.3%	2.2%	19.7%	77.7%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	5.0%	95.0%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.69

Count Across Benchmark Agreement Levels for Data Condition 35 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	2926	0
Lambda-2	0	0	31	1072	1823	0
Kappa	3	18	158	1992	755	0
S	0	0	0	0	2926	0
AC1	0	0	0	0	2926	0
Kripp. Alpha	0	0	0	372	2554	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table E.70

Percentage of Values within Benchmark Levels for Data Condition 35 (3x3)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	1.1%	36.6%	62.3%	0.0%
Kappa	0.1%	0.6%	5.4%	68.1%	25.8%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	12.7%	87.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

APPENDIX F

BENCHMARK AGREEMENT CLASSIFICATION TABLES (4X4 AGREEMENT MATRICES)

Table F.1

Count Across Benchmark Agreement Levels for Data Condition 1 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	96
Lambda-2	0	0	0	0	0	96
Kappa	2	0	2	6	16	70
S	0	0	0	0	0	96
AC1	0	0	0	0	0	96
Kripp. Alpha	0	0	0	4	14	78

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.2

Percentage of Values within Benchmark Levels for Data Condition 1 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kappa	2.1%	0.0%	2.1%	6.3%	16.7%	72.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	4.2%	14.6%	81.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.3

Count Across Benchmark Agreement Levels for Data Condition 2 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	91
Lambda-2	0	0	0	0	0	91
Kappa	2	2	4	12	62	9
S	0	0	0	0	0	91
AC1	0	0	0	0	0	91
Kripp. Alpha	0	0	0	8	38	45

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.4

Percentage of Values within Benchmark Levels for Data Condition 2 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kappa	2.2%	2.2%	4.4%	13.2%	68.1%	9.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	8.8%	41.8%	49.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.5

Count Across Benchmark Agreement Levels for Data Condition 3 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	36	50
Lambda-2	0	0	0	0	0	86
Kappa	2	4	8	20	52	0
S	0	0	0	0	86	0
AC1	0	0	0	0	0	86
Kripp. Alpha	0	0	0	14	72	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.6

Percentage of Values within Benchmark Levels for Data Condition 3 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	41.9%	58.1%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kappa	2.3%	4.7%	9.3%	23.3%	60.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	16.3%	83.7%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.7

Count Across Benchmark Agreement Levels for Data Condition 4 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	81	0
Lambda-2	0	0	0	0	81	0
Kappa	2	6	12	42	19	0
S	0	0	0	0	81	0
AC1	0	0	0	0	81	0
Kripp. Alpha	0	0	0	20	61	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.8

Percentage of Values within Benchmark Levels for Data Condition 4 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kappa	2.5%	7.4%	14.8%	51.9%	23.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	24.7%	75.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.9

Count Across Benchmark Agreement Levels for Data Condition 5 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	76	0
Lambda-2	0	0	0	0	76	0
Kappa	2	8	18	48	0	0
S	0	0	0	0	76	0
AC1	0	0	0	0	76	0
Kripp. Alpha	0	0	0	32	44	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.10

Percentage of Values within Benchmark Levels for Data Condition 5 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kappa	2.6%	10.5%	23.7%	63.2%	0.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	42.1%	57.9%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.11

Count Across Benchmark Agreement Levels for Data Condition 6 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	96
Lambda-2	0	0	0	0	0	96
Kappa	2	0	2	6	16	70
S	0	0	0	0	0	96
AC1	0	0	0	0	0	96
Kripp. Alpha	0	0	0	4	14	78

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.12

Percentage of Values within Benchmark Levels for Data Condition 6 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kappa	2.1%	0.0%	2.1%	6.3%	16.7%	72.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	4.2%	14.6%	81.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.13

Count Across Benchmark Agreement Levels for Data Condition 7 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	91
Lambda-2	0	0	0	0	0	91
Kappa	2	2	4	12	62	9
S	0	0	0	0	0	91
AC1	0	0	0	0	0	91
Kripp. Alpha	0	0	0	8	38	45

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.14

Percentage of Values within Benchmark Levels for Data Condition 7 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kappa	2.2%	2.2%	4.4%	13.2%	68.1%	9.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	8.8%	41.8%	49.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.15

Count Across Benchmark Agreement Levels for Data Condition 8 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	86
Lambda-2	0	0	0	0	0	86
Kappa	2	4	8	20	52	0
S	0	0	0	0	86	0
AC1	0	0	0	0	0	86
Kripp. Alpha	0	0	0	14	72	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.16

Percentage of Values within Benchmark Levels for Data Condition 8 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kappa	2.3%	4.7%	9.3%	23.3%	60.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	16.3%	83.7%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.17

Count Across Benchmark Agreement Levels for Data Condition 9 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	81	0
Lambda-2	0	0	0	0	81	0
Kappa	2	6	12	42	19	0
S	0	0	0	0	81	0
AC1	0	0	0	0	81	0
Kripp. Alpha	0	0	0	20	61	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.18

Percentage of Values within Benchmark Levels for Data Condition 9 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kappa	2.5%	7.4%	14.8%	51.9%	23.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	24.7%	75.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.19

Count Across Benchmark Agreement Levels for Data Condition 10 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	76	0
Lambda-2	0	0	0	0	76	0
Kappa	2	8	18	48	0	0
S	0	0	0	0	76	0
AC1	0	0	0	0	76	0
Kripp. Alpha	0	0	0	32	44	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.20

Percentage of Values within Benchmark Levels for Data Condition 10 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kappa	2.6%	10.5%	23.7%	63.2%	0.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	42.1%	57.9%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.21

Count Across Benchmark Agreement Levels for Data Condition 11 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	96
Lambda-2	0	0	0	0	0	96
Kappa	2	0	2	6	16	70
S	0	0	0	0	0	96
AC1	0	0	0	0	0	96
Kripp. Alpha	0	0	0	4	14	78

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.22

Percentage of Values within Benchmark Levels for Data Condition 11 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kappa	2.1%	0.0%	2.1%	6.3%	16.7%	72.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	4.2%	14.6%	81.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.23

Count Across Benchmark Agreement Levels for Data Condition 12 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	91
Lambda-2	0	0	0	0	19	72
Kappa	2	2	4	12	62	9
S	0	0	0	0	0	91
AC1	0	0	0	0	0	91
Kripp. Alpha	0	0	0	8	38	45

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.24

Percentage of Values within Benchmark Levels for Data Condition 12 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	20.9%	79.1%
Kappa	2.2%	2.2%	4.4%	13.2%	68.1%	9.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	8.8%	41.8%	49.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.25

Count Across Benchmark Agreement Levels for Data Condition 13 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	86	0
Lambda-2	0	0	0	0	77	9
Kappa	2	4	8	20	52	0
S	0	0	0	0	86	0
AC1	0	0	0	0	0	86
Kripp. Alpha	0	0	0	14	72	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.26

Percentage of Values within Benchmark Levels for Data Condition 13 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	89.5%	10.5%
Kappa	2.3%	4.7%	9.3%	23.3%	60.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	16.3%	83.7%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.27

Count Across Benchmark Agreement Levels for Data Condition 14 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	81	0
Lambda-2	0	0	0	0	81	0
Kappa	2	6	12	42	19	0
S	0	0	0	0	81	0
AC1	0	0	0	0	81	0
Kripp. Alpha	0	0	0	20	61	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.28

Percentage of Values within Benchmark Levels for Data Condition 14 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kappa	2.5%	7.4%	14.8%	51.9%	23.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	24.7%	75.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.29

Count Across Benchmark Agreement Levels for Data Condition 15 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	76	0
Lambda-2	0	0	0	26	50	0
Kappa	2	8	18	48	0	0
S	0	0	0	0	76	0
AC1	0	0	0	0	76	0
Kripp. Alpha	0	0	0	32	44	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.30

Percentage of Values within Benchmark Levels for Data Condition 15 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	34.2%	65.8%	0.0%
Kappa	2.6%	10.5%	23.7%	63.2%	0.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	42.1%	57.9%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.31

Count Across Benchmark Agreement Levels for Data Condition 16 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	96
Lambda-2	0	0	0	0	0	96
Kappa	0	1	2	5	15	73
S	0	0	0	0	0	96
AC1	0	0	0	0	0	96
Kripp. Alpha	0	0	0	4	14	78

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.32

Percentage of Values within Benchmark Levels for Data Condition 16 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kappa	0.0%	1.0%	2.1%	5.2%	15.6%	76.0%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	4.2%	14.6%	81.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.33

Count Across Benchmark Agreement Levels for Data Condition 17 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	91
Lambda-2	0	0	0	0	21	70
Kappa	0	2	4	10	50	25
S	0	0	0	0	0	91
AC1	0	0	0	0	0	91
Kripp. Alpha	0	0	0	8	38	45

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.34

Percentage of Values within Benchmark Levels for Data Condition 17 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	23.1%	76.9%
Kappa	0.0%	2.2%	4.4%	11.0%	54.9%	27.5%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	8.8%	41.8%	49.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.35

Count Across Benchmark Agreement Levels for Data Condition 18 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	86
Lambda-2	0	0	0	0	80	6
Kappa	0	2	6	20	58	0
S	0	0	0	0	86	0
AC1	0	0	0	0	0	86
Kripp. Alpha	0	0	0	14	72	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.36

Percentage of Values within Benchmark Levels for Data Condition 18 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	93.0%	7.0%
Kappa	0.0%	2.3%	7.0%	23.3%	67.4%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	16.3%	83.7%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.37

Count Across Benchmark Agreement Levels for Data Condition 19 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	81	0
Lambda-2	0	0	0	3	78	0
Kappa	0	2	11	33	35	0
S	0	0	0	0	81	0
AC1	0	0	0	0	81	0
Kripp. Alpha	0	0	0	20	61	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.38

Percentage of Values within Benchmark Levels for Data Condition 19 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	3.7%	96.3%	0.0%
Kappa	0.0%	2.5%	13.6%	40.7%	43.2%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	24.7%	75.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.39

Count Across Benchmark Agreement Levels for Data Condition 20 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	76	0
Lambda-2	0	0	0	30	46	0
Kappa	0	4	16	56	0	0
S	0	0	0	0	76	0
AC1	0	0	0	0	76	0
Kripp. Alpha	0	0	0	32	44	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.40

Percentage of Values within Benchmark Levels for Data Condition 20 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	39.5%	60.5%	0.0%
Kappa	0.0%	5.3%	21.1%	73.7%	0.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	42.1%	57.9%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.41

Count Across Benchmark Agreement Levels for Data Condition 21 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	96
Lambda-2	0	0	0	0	0	96
Kappa	2	0	2	6	16	70
S	0	0	0	0	0	96
AC1	0	0	0	0	0	96
Kripp. Alpha	0	0	0	4	14	78

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.42

Percentage of Values within Benchmark Levels for Data Condition 21 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kappa	2.1%	0.0%	2.1%	6.3%	16.7%	72.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	4.2%	14.6%	81.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.43

Count Across Benchmark Agreement Levels for Data Condition 22 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	91
Lambda-2	0	0	0	0	31	60
Kappa	2	2	4	12	62	9
S	0	0	0	0	0	91
AC1	0	0	0	0	0	91
Kripp. Alpha	0	0	0	8	38	45

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.44

Percentage of Values within Benchmark Levels for Data Condition 22 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	34.1%	65.9%
Kappa	2.2%	2.2%	4.4%	13.2%	68.1%	9.9%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	8.8%	41.8%	49.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.45

Count Across Benchmark Agreement Levels for Data Condition 23 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	86	0
Lambda-2	0	0	0	0	86	0
Kappa	2	4	8	20	52	0
S	0	0	0	0	86	0
AC1	0	0	0	0	0	86
Kripp. Alpha	0	0	0	14	72	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.46

Percentage of Values within Benchmark Levels for Data Condition 23 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kappa	2.3%	4.7%	9.3%	23.3%	60.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	16.3%	83.7%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.47

Count Across Benchmark Agreement Levels for Data Condition 24 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	81	0
Lambda-2	0	0	0	16	65	0
Kappa	2	6	12	42	19	0
S	0	0	0	0	81	0
AC1	0	0	0	0	81	0
Kripp. Alpha	0	0	0	20	61	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.48

Percentage of Values within Benchmark Levels for Data Condition 24 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	19.8%	80.2%	0.0%
Kappa	2.5%	7.4%	14.8%	51.9%	23.5%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	24.7%	75.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.49

Count Across Benchmark Agreement Levels for Data Condition 25 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	76	0
Lambda-2	0	0	0	76	0	0
Kappa	2	8	18	48	0	0
S	0	0	0	0	76	0
AC1	0	0	0	0	76	0
Kripp. Alpha	0	0	0	32	44	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.50

Percentage of Values within Benchmark Levels for Data Condition 25 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%
Kappa	2.6%	10.5%	23.7%	63.2%	0.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	42.1%	57.9%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.51

Count Across Benchmark Agreement Levels for Data Condition 26 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	96
Lambda-2	0	0	0	0	0	96
Kappa	0	1	2	5	15	73
S	0	0	0	0	0	96
AC1	0	0	0	0	0	96
Kripp. Alpha	0	0	0	4	14	78

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.52

Percentage of Values within Benchmark Levels for Data Condition 26 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kappa	0.0%	1.0%	2.1%	5.2%	15.6%	76.0%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	4.2%	14.6%	81.3%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.53

Count Across Benchmark Agreement Levels for Data Condition 27 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	91
Lambda-2	0	0	0	0	22	69
Kappa	0	2	4	10	50	25
S	0	0	0	0	0	91
AC1	0	0	0	0	0	91
Kripp. Alpha	0	0	0	8	38	45

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.54

Percentage of Values within Benchmark Levels for Data Condition 27 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	24.2%	75.8%
Kappa	0.0%	2.2%	4.4%	11.0%	54.9%	27.5%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	8.8%	41.8%	49.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.55

Count Across Benchmark Agreement Levels for Data Condition 28 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	86
Lambda-2	0	0	0	0	86	0
Kappa	0	2	6	20	58	0
S	0	0	0	0	86	0
AC1	0	0	0	0	0	86
Kripp. Alpha	0	0	0	14	72	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.56

Percentage of Values within Benchmark Levels for Data Condition 28 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kappa	0.0%	2.3%	7.0%	23.3%	67.4%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	16.3%	83.7%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.57

Count Across Benchmark Agreement Levels for Data Condition 29 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	81	0
Lambda-2	0	0	0	0	81	0
Kappa	0	2	11	33	35	0
S	0	0	0	0	81	0
AC1	0	0	0	0	81	0
Kripp. Alpha	0	0	0	20	61	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.58

Percentage of Values within Benchmark Levels for Data Condition 29 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kappa	0.0%	2.5%	13.6%	40.7%	43.2%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	24.7%	75.3%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.59

Count Across Benchmark Agreement Levels for Data Condition 30 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	76	0
Lambda-2	0	0	0	49	27	0
Kappa	0	4	16	56	0	0
S	0	0	0	0	76	0
AC1	0	0	0	0	76	0
Kripp. Alpha	0	0	0	32	44	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.60

Percentage of Values within Benchmark Levels for Data Condition 30 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	64.5%	35.5%	0.0%
Kappa	0.0%	5.3%	21.1%	73.7%	0.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	42.1%	57.9%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.61

Count Across Benchmark Agreement Levels for Data Condition 51 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	42608
Lambda-2	0	0	0	0	0	42608
Kappa	0	0	0	0	0	42608
S	0	0	0	0	0	42608
AC1	0	0	0	0	0	42608
Kripp. Alpha	0	0	0	0	0	42680

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.62

Percentage of Values within Benchmark Levels for Data Condition 51 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kappa	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.0%	0.0%	100.2%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.63

Count Across Benchmark Agreement Levels for Data Condition 52 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	39480
Lambda-2	0	0	0	0	0	39480
Kappa	0	0	0	0	0	39480
S	0	0	0	0	0	39480
AC1	0	0	0	0	0	39480
Kripp. Alpha	0	0	0	0	0	39480

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.64

Percentage of Values within Benchmark Levels for Data Condition 52 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kappa	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.65

Count Across Benchmark Agreement Levels for Data Condition 53 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	36864	16112
Lambda-2	0	0	0	0	25121	27855
Kappa	0	0	0	0	52976	0
S	0	0	0	0	52976	0
AC1	0	0	0	0	2893	50083
Kripp. Alpha	0	0	0	0	11940	41036

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.66

Percentage of Values within Benchmark Levels for Data Condition 53 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	69.6%	30.4%
Lambda-2	0.0%	0.0%	0.0%	0.0%	47.4%	52.6%
Kappa	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	5.5%	94.5%
Kripp. Alpha	0.0%	0.0%	0.0%	0.0%	22.5%	77.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.67

Count Across Benchmark Agreement Levels for Data Condition 54 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	30014	0
Lambda-2	0	0	0	0	30014	0
Kappa	0	0	0	0	30014	0
S	0	0	0	0	30014	0
AC1	0	0	0	0	30014	0
Kripp. Alpha	0	0	0	0	30014	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.68

Percentage of Values within Benchmark Levels for Data Condition 54 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	903.8%	0.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	903.8%	0.0%
Kappa	0.0%	0.0%	0.0%	0.0%	903.8%	0.0%
S	0.0%	0.0%	0.0%	0.0%	903.8%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	903.8%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.0%	903.8%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.69

Count Across Benchmark Agreement Levels for Data Condition 55 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	24860	0
Lambda-2	0	0	0	0	24860	0
Kappa	0	0	0	6	24854	0
S	0	0	0	0	24860	0
AC1	0	0	0	0	24860	0
Kripp. Alpha	0	0	0	0	24860	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.70

Percentage of Values within Benchmark Levels for Data Condition 55 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kappa	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.71

Count Across Benchmark Agreement Levels for Data Condition 56 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	12557
Lambda-2	0	0	0	0	0	12557
Kappa	0	0	0	0	0	12557
S	0	0	0	0	0	12557
AC1	0	0	0	0	0	12557
Kripp. Alpha	0	0	0	0	0	12557

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.72

Percentage of Values within Benchmark Levels for Data Condition 56 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kappa	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.73

Count Across Benchmark Agreement Levels for Data Condition 57 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	12848
Lambda-2	0	0	0	0	0	12848
Kappa	0	0	0	0	0	12848
S	0	0	0	0	0	12848
AC1	0	0	0	0	0	12848
Kripp. Alpha	0	0	0	0	0	12848

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.74

Percentage of Values within Benchmark Levels for Data Condition 57 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kappa	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.75

Count Across Benchmark Agreement Levels for Data Condition 58 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	52976
Lambda-2	0	0	0	0	24168	28808
Kappa	0	0	0	0	51345	1631
S	0	0	0	0	52976	0
AC1	0	0	0	0	0	52976
Kripp. Alpha	0	0	0	0	11940	41036

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.76

Percentage of Values within Benchmark Levels for Data Condition 58 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	45.6%	54.4%
Kappa	0.0%	0.0%	0.0%	0.0%	96.9%	3.1%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.0%	22.5%	77.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.77

Count Across Benchmark Agreement Levels for Data Condition 59 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	9985
Lambda-2	0	0	0	0	0	9985
Kappa	0	0	0	0	0	9985
S	0	0	0	0	0	9985
AC1	0	0	0	0	0	9985
Kripp. Alpha	0	0	0	0	0	9985

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.78

Percentage of Values within Benchmark Levels for Data Condition 59 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kappa	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.79

Count Across Benchmark Agreement Levels for Data Condition 60 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	9152	0
Lambda-2	0	0	0	0	9152	0
Kappa	0	0	0	0	9152	0
S	0	0	0	0	9152	0
AC1	0	0	0	0	9152	0
Kripp. Alpha	0	0	0	0	9152	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.80

Percentage of Values within Benchmark Levels for Data Condition 60 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kappa	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.81

Count Across Benchmark Agreement Levels for Data Condition 61 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	26557
Lambda-2	0	0	0	0	0	26557
Kappa	0	0	0	0	0	26557
S	0	0	0	0	0	26557
AC1	0	0	0	0	0	26557
Kripp. Alpha	0	0	0	0	0	26557

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.82

Percentage of Values within Benchmark Levels for Data Condition 61 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	570.4%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	570.4%
Kappa	0.0%	0.0%	0.0%	0.0%	0.0%	570.4%
S	0.0%	0.0%	0.0%	0.0%	0.0%	570.4%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	570.4%
Kripp. Alpha	0.0%	0.0%	0.0%	0.0%	0.0%	570.4%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.83

Count Across Benchmark Agreement Levels for Data Condition 62 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	23994
Lambda-2	0	0	0	0	0	23994
Kappa	0	0	0	0	0	23994
S	0	0	0	0	0	23994
AC1	0	0	0	0	0	23994
Kripp. Alpha	0	0	0	0	0	23994

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.84

Percentage of Values within Benchmark Levels for Data Condition 62 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kappa	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
S	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.85

Count Across Benchmark Agreement Levels for Data Condition 63 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	0	52976
Lambda-2	0	0	0	0	25132	27844
Kappa	0	0	0	0	52976	0
S	0	0	0	0	52976	0
AC1	0	0	0	0	0	52976
Kripp. Alpha	0	0	0	0	11940	41036

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.86

Percentage of Values within Benchmark Levels for Data Condition 63 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	47.4%	52.6%
Kappa	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.0%	22.5%	77.5%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.87

Count Across Benchmark Agreement Levels for Data Condition 64 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	17679	0
Lambda-2	0	0	0	0	17679	0
Kappa	0	0	0	0	17679	0
S	0	0	0	0	17679	0
AC1	0	0	0	0	17679	0
Kripp. Alpha	0	0	0	0	17679	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.88

Percentage of Values within Benchmark Levels for Data Condition 64 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kappa	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.89

Count Across Benchmark Agreement Levels for Data Condition 65 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0	0	0	0	14022	0
Lambda-2	0	0	0	0	14022	0
Kappa	0	0	0	16	14006	0
S	0	0	0	0	14022	0
AC1	0	0	0	0	14022	0
Kripp. Alpha	0	0	0	0	14022	0

Note. Benchmark agreement categories from Landis & Koch (1977).

Table F.90

Percentage of Values within Benchmark Levels for Data Condition 65 (4x4)

	No Agreement	Slight	Fair	Moderate	Substantial	Almost Perfect
	< 0.00	0.00 - 0.20	0.21 - 0.40	0.41 - 0.60	0.61 - 0.80	0.81 - 1.00
Lambda-1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Lambda-2	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kappa	0.0%	0.0%	0.0%	0.1%	99.9%	0.0%
S	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
AC1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Kripp. Alpha	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%

Note. Benchmark agreement categories from Landis & Koch (1977).

APPENDIX G

CORRELATION TABLES AND FIGURES (2X2 AGREEMENT MATRICES)

Table G.1

Correlation Matrix for Data Condition 1 (2x2)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	-0.84***	-			
Kappa	0.02	0.44***	-		
AC1	0.14	-0.54***	-0.73***	-	
Kripp.Alpha	0.02	0.46***	0.99***	-0.81***	-

Note. ***p < .001.

Table G.2

Correlation Matrix for Data Condition 2 (2x2)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	-0.83***	-			
Kappa	0.04	0.45***	-		
AC1	0.29**	-0.66***	-0.80***	-	
Kripp.Alpha	0.04	0.46***	0.99***	-0.85***	-

Note. **p < .01; ***p < .001.

Table G.3

Correlation Matrix for Data Condition 3 (2x2)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	-0.83***	-			
Kappa	0.06	0.45***	-		
AC1	0.45***	-0.77***	-0.77***	-	
Kripp.Alpha	0.06	0.45***	0.99***	-0.80***	-

Note. ***p < .001.

Table G.4

Correlation Matrix for Data Condition 4 (2x2)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	-0.83***	-			
Kappa	0.08	0.45***	-		
AC1	0.59***	-0.85***	-0.69***	-	
Kripp.Alpha	0.08	0.45***	1.00***	-0.71***	-

Note. ***p < .001.

Table G.5

Correlation Matrix for Data Condition 5 (2x2)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	-0.83***	-			
Kappa	0.09	0.44***	-		
AC1	0.72***	-0.90***	-0.58***	-	
Kripp.Alpha	0.09	0.44***	1.00***	-0.59***	-

Note. ***p < .001.

Table G.6

Correlation Matrix for Data Condition 6 (2x2)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.90***	-			
Kappa	0.02	0.39***	-		
AC1	0.14	-0.26*	-0.73***	-	
Kripp.Alpha	0.02	0.41***	0.99***	-0.81***	-

Note. *p < .05; ***p < .001.

Table G.7

Correlation Matrix for Data Condition 7 (2x2)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.93***	-			
Kappa	0.04	0.36***	-		
AC1	0.29**	-0.05	-0.80***	-	
Kripp.Alpha	0.04	0.37***	0.99***	-0.85***	-

Note. **p < .01; ***p < .001.

Table G.8

Correlation Matrix for Data Condition 8 (2x2)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.95***	-			
Kappa	0.06	0.33**	-		
AC1	0.45***	0.17	-0.77***	-	
Kripp.Alpha	0.06	0.34**	0.99***	-0.80***	-

Note. **p < .01; ***p < .001.

Table G.9

Correlation Matrix for Data Condition 9 (2x2)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.97***	-			
Kappa	0.08	0.30**	-		
AC1	0.59***	0.39***	-0.69***	-	
Kripp.Alpha	0.08	0.30**	1.00***	-0.71***	-

Note. **p < .01; ***p < .001.

Table G.10

Correlation Matrix for Data Condition 10 (2x2)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.98***	-			
Kappa	0.09	0.27*	-		
AC1	0.72***	0.57***	-0.58***	-	
Kripp.Alpha	0.09	0.27*	1.00***	-0.59***	-

Note. *p < .05; ***p < .001.

Table G.11

Correlation Matrix for Data Condition 11 (2x2)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.41***	-		
AC1	-0.41***	-0.74***	-	
Kripp.Alpha	0.43***	0.98***	-0.83***	-

Note. ***p < .001.

Table G.12

Correlation Matrix for Data Condition 12 (2x2)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.40***	-		
AC1	-0.39***	-0.84***	-	
Kripp.Alpha	0.41***	0.99***	-0.90***	-

Note. ***p < .001.

Table G.13

Correlation Matrix for Data Condition 13 (2x2)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.38***	-		
AC1	-0.37***	-0.89***	-	
Kripp.Alpha	0.39***	0.99***	-0.94***	-

Note. ***p < .001.

Table G.14

Correlation Matrix for Data Condition 14 (2x2)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.36**	-		
AC1	-0.35**	-0.92***	-	
Kripp.Alpha	0.36***	0.99***	-0.96***	-

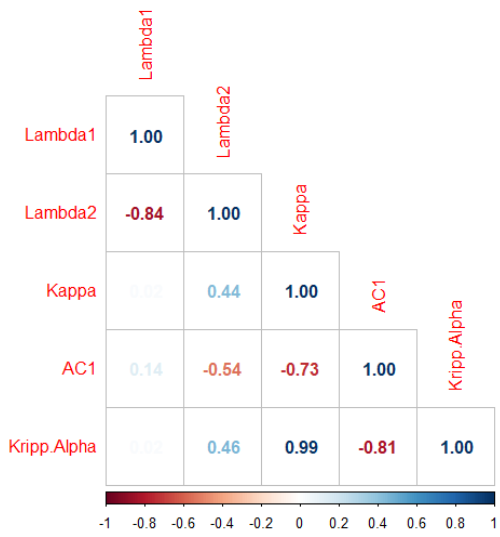
Note. **p < .01; ***p < .001.

Table G.15

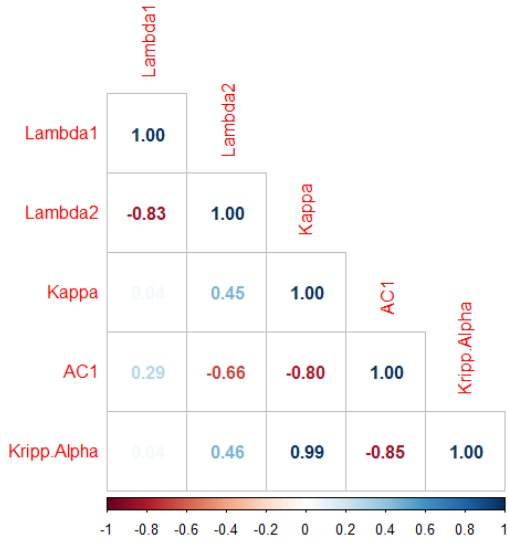
Correlation Matrix for Data Condition 15 (2x2)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.33**	-		
AC1	-0.33**	-0.95***	-	
Kripp.Alpha	0.34**	1.00***	-0.97***	-

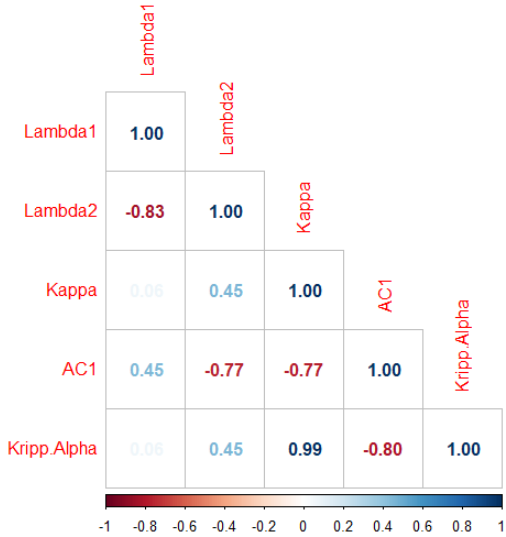
Note. **p < .01; ***p < .001.



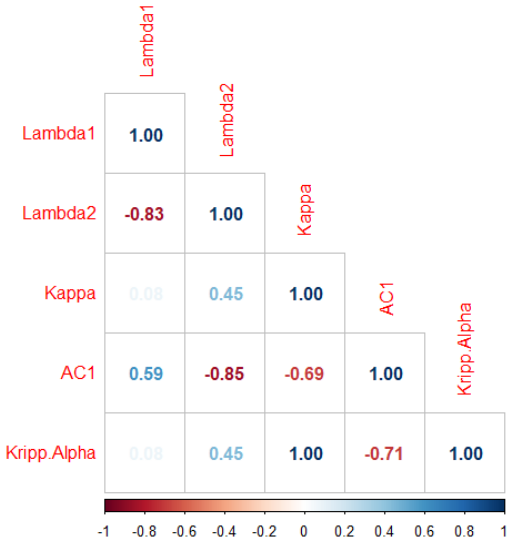
DC1



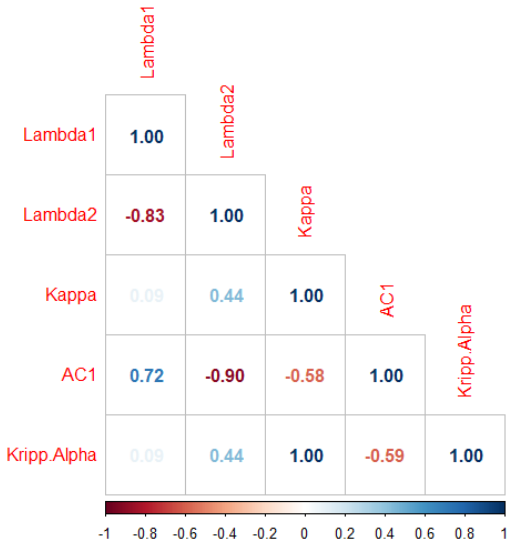
DC2



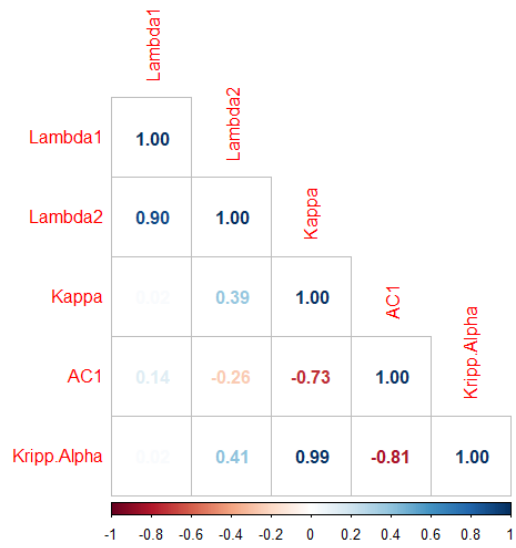
DC3



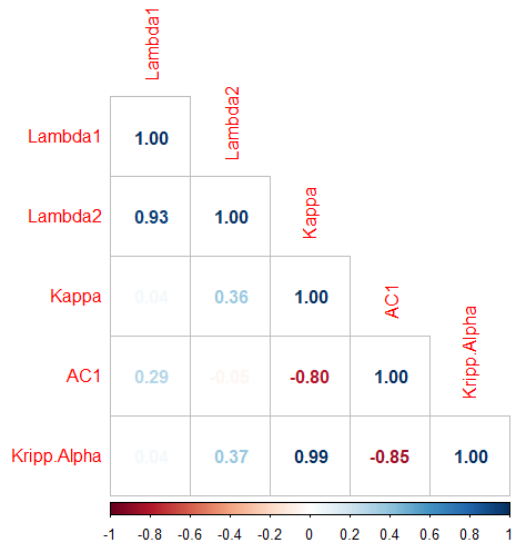
DC4



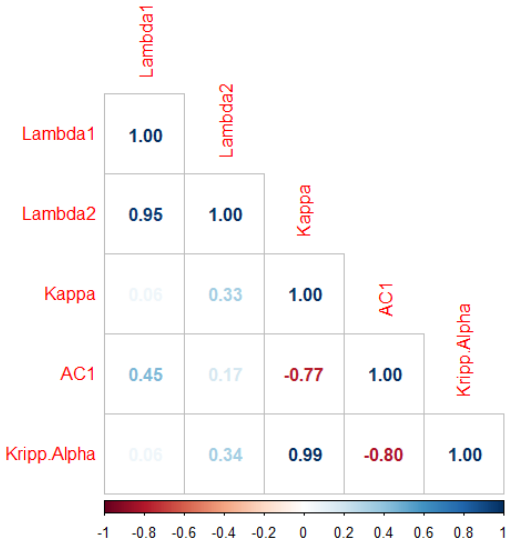
DC5



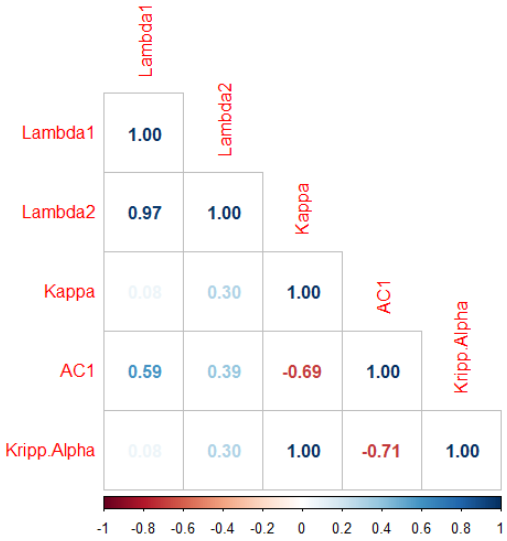
DC6



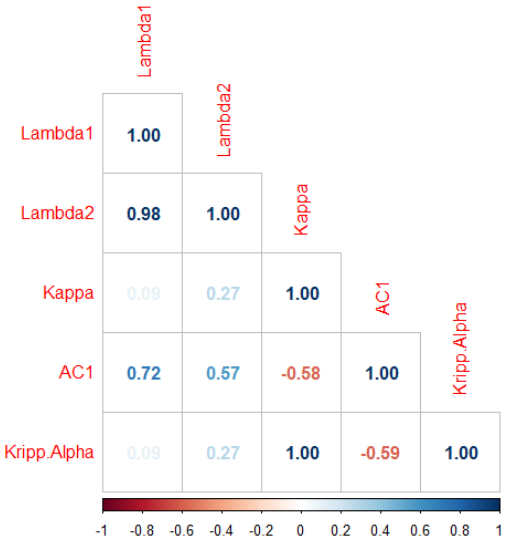
DC7



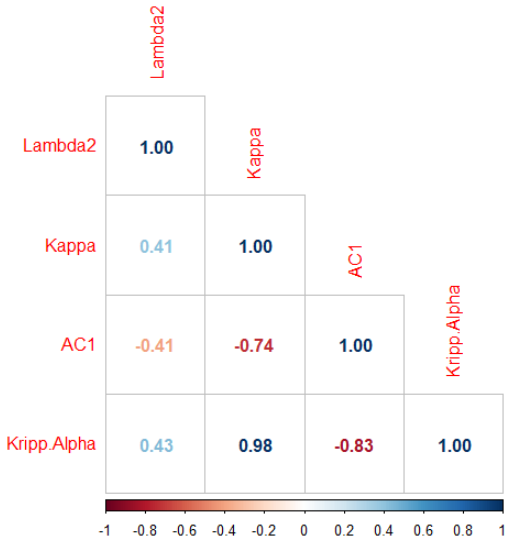
DC8



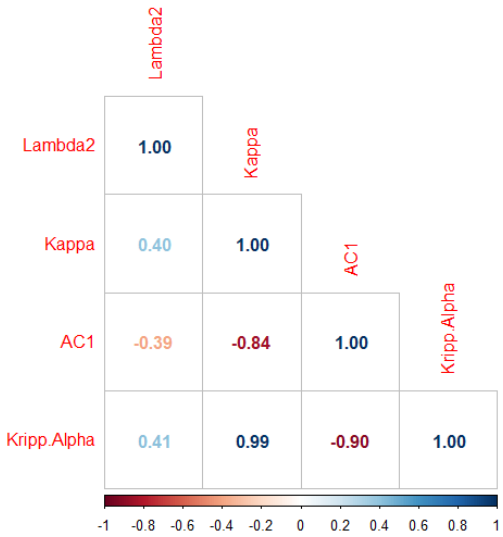
DC9



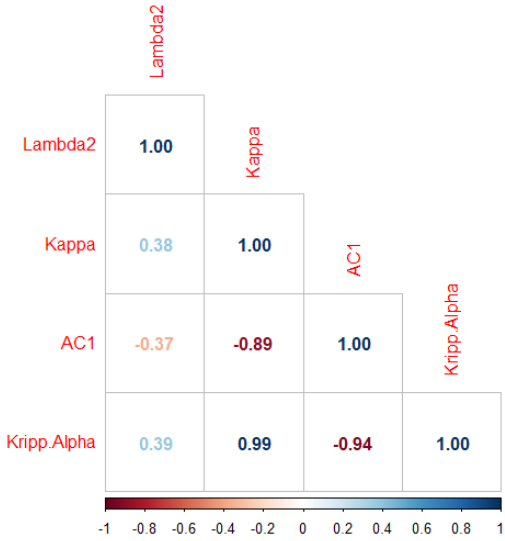
DC10



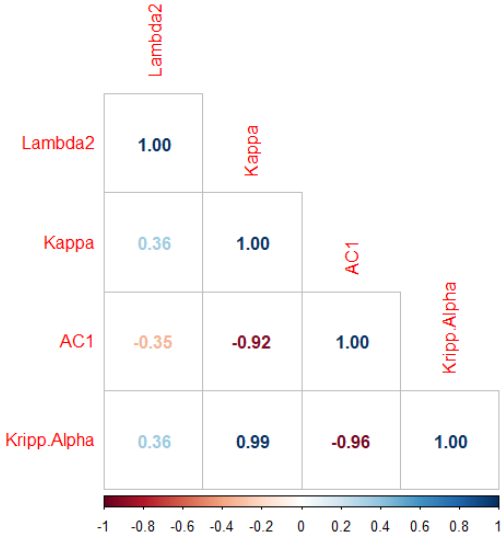
DC11



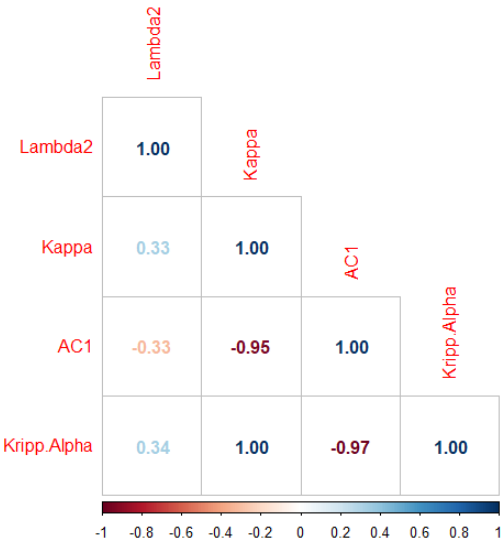
DC12



DC13



DC14



DC15

APPENDIX H

CORRELATION TABLES AND FIGURES (3X3 AGREEMENT MATRICES)

Table H.1

Correlation Matrix for Data Condition 1 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	-0.89***	-			
Kappa	0.01	0.37***	-		
AC1	0.18	-0.55***	-0.79***	-	
Kripp.Alpha	0.01	0.39***	0.99***	-0.86***	-

Note. *** $p < .001$.

Table H.2

Correlation Matrix for Data Condition 2 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	-0.93***	-			
Kappa	0.01	0.33**	-		
AC1	0.37***	-0.66***	-0.83***	-	
Kripp.Alpha	0.01	0.34**	0.99***	-0.87***	-

Note. ** $p < .01$; *** $p < .001$.

Table H.3

Correlation Matrix for Data Condition 3 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	-0.95***	-			
Kappa	0.02	0.28*	-		
AC1	0.53***	-0.76***	-0.78***	-	
Kripp.Alpha	0.02	0.28**	0.99***	-0.81***	-

Note. * $p < .05$; ** $p < .01$; *** $p < .001$.

Table H.4

Correlation Matrix for Data Condition 4 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	-0.97***	-			
Kappa	0.02	0.22*	-		
AC1	0.67***	-0.83***	-0.70***	-	
Kripp.Alpha	0.02	0.23*	1.00***	-0.71***	-

Note. * $p < .05$; *** $p < .001$.

Table H.5

Correlation Matrix for Data Condition 5 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	-0.98***	-			
Kappa	0.03	0.17	-		
AC1	0.77***	-0.88***	-0.60***	-	
Kripp.Alpha	0.03	0.17	1.00***	-0.61***	-

Note. *** $p < .001$.

Table H.6

Correlation Matrix for Data Condition 6 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.91***	-			
Kappa	0.01	0.34***	-		
AC1	0.18	-0.20	-0.79***	-	
Kripp.Alpha	0.01	0.36***	0.99***	-0.86***	-

Note. *** $p < .001$.

Table H.7

Correlation Matrix for Data Condition 7 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.95***	-			
Kappa	0.01	0.29**	-		
AC1	0.37***	0.07	-0.83***	-	
Kripp.Alpha	0.01	0.31**	0.99***	-0.87***	-

Note. ** $p < .01$; *** $p < .001$.

Table H.8

Correlation Matrix for Data Condition 8 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.97***	-			
Kappa	0.02	0.25*	-		
AC1	0.53***	0.31**	-0.78***	-	
Kripp.Alpha	0.02	0.26*	0.99***	-0.81***	-

Note. * $p < .05$; ** $p < .01$; *** $p < .001$.

Table H.9

Correlation Matrix for Data Condition 9 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.98***	-			
Kappa	0.02	0.21	-		
AC1	0.67***	0.51***	-0.70***	-	
Kripp.Alpha	0.02	0.21	1.00***	-0.71***	-

Note. *** $p < .001$.

Table H.10

Correlation Matrix for Data Condition 10 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.99***	-			
Kappa	0.02	0.18	-		
AC1	0.77***	0.66***	-0.60***	-	
Kripp.Alpha	0.02	0.18	1.00***	-0.61***	-

Note. *** $p < .001$.

Table H.11

Correlation Matrix for Data Condition 11 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.91***	-			
Kappa	0.01	0.34***	-		
AC1	0.18	-0.20	-0.78***	-	
Kripp.Alpha	0.01	0.36***	0.99***	-0.86***	-

Note. *** $p < .001$.

Table H.12

Correlation Matrix for Data Condition 12 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.95***	-			
Kappa	0.01	0.29**	-		
AC1	0.36***	0.06	-0.83***	-	
Kripp.Alpha	0.01	0.31**	0.99***	-0.87***	-

Note. ** $p < .01$; *** $p < .001$.

Table H.13

Correlation Matrix for Data Condition 13 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.97***	-			
Kappa	0.02	0.25*	-		
AC1	0.53***	0.31**	-0.78***	-	
Kripp.Alpha	0.02	0.26*	0.99***	-0.81***	-

Note. * $p < .05$; ** $p < .01$; *** $p < .001$.

Table H.14

Correlation Matrix for Data Condition 14 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.98***	-			
Kappa	0.02	0.21	-		
AC1	0.67***	0.51***	-0.70***	-	
Kripp.Alpha	0.02	0.21	1.00***	-0.71***	-

Note. *** $p < .001$.

Table H.15

Correlation Matrix for Data Condition 15 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.99***	-			
Kappa	0.02	0.18	-		
AC1	0.77***	0.66***	-0.60***	-	
Kripp.Alpha	0.02	0.18	1.00***	-0.60***	-

Note. *** $p < .001$.

Table H.16

Correlation Matrix for Data Condition 16 (3x3)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.37***	-		
AC1	-0.55***	-0.78***	-	
Kripp.Alpha	0.39***	0.99***	-0.86***	-

Note. *** $p < .001$.

Table H.17

Correlation Matrix for Data Condition 17 (3x3)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.33**	-		
AC1	-0.66***	-0.83***	-	
Kripp.Alpha	0.34***	0.99***	-0.87***	-

Note. ** $p < .01$; *** $p < .001$.

Table H.18

Correlation Matrix for Data Condition 18 (3x3)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.28**	-		
AC1	-0.75***	-0.78***	-	
Kripp.Alpha	0.28**	0.99***	-0.81***	-

Note. ** $p < .01$; *** $p < .001$.

Table H.19

Correlation Matrix for Data Condition 19 (3x3)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.23*	-		
AC1	-0.83***	-0.70***	-	
Kripp.Alpha	0.23*	1.00***	-0.71***	-

Note. * $p < .05$; *** $p < .001$.

Table H.20

Correlation Matrix for Data Condition 20 (3x3)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.18	-		
AC1	-0.88***	-0.60***	-	
Kripp.Alpha	0.18	1.00***	-0.60***	-

Note. *** $p < .001$.

Table H.21

Correlation Matrix for Data Condition 21 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.02	-			
Kappa	0.01	0.34***	-		
AC1	-0.07***	-0.35***	-0.79***	-	
Kripp.Alpha	0.01	0.27***	0.99***	-0.84***	-

Note. *** $p < .001$.

Table H.22

Correlation Matrix for Data Condition 22 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.06***	-			
Kappa	0.01	0.35***	-		
AC1	0.00	-0.36***	-0.87***	-	
Kripp.Alpha	0.01	0.23***	0.98***	-0.90***	-

Note. *** $p < .001$.

Table H.23

Correlation Matrix for Data Condition 23 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.11***	-			
Kappa	0.01	0.37***	-		
AC1	0.08***	-0.36***	-0.90***	-	
Kripp.Alpha	0.02	0.19***	0.97***	-0.92***	-

Note. *** $p < .001$.

Table H.24

Correlation Matrix for Data Condition 24 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.18***	-			
Kappa	0.02	0.40***	-		
AC1	0.16***	-0.35***	-0.91***	-	
Kripp.Alpha	0.03	0.16***	0.96***	-0.91***	-

Note. *** $p < .001$.

Table H.25

Correlation Matrix for Data Condition 25 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.26***	-			
Kappa	0.02	0.42***	-		
AC1	0.26***	-0.32***	-0.91***	-	
Kripp.Alpha	0.03	0.13***	0.94***	-0.89***	-

Note. *** $p < .001$.

Table H.26

Correlation Matrix for Data Condition 26 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.48***	-			
Kappa	0.03*	0.34***	-		
AC1	0.13***	-0.35***	-0.77***	-	
Kripp.Alpha	0.00	0.27***	0.99***	-0.82***	-

Note. * $p < .05$; *** $p < .001$.

Table H.27

Correlation Matrix for Data Condition 27 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.53***	-			
Kappa	0.07***	0.35***	-		
AC1	0.18***	-0.35***	-0.82***	-	
Kripp.Alpha	0.01	0.23***	0.98***	-0.85***	-

Note. *** $p < .001$.

Table H.28

Correlation Matrix for Data Condition 28 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.58***	-			
Kappa	0.10***	0.37***	-		
AC1	0.23***	-0.33***	-0.82***	-	
Kripp.Alpha	0.01	0.19***	0.97***	-0.82***	-

Note. *** $p < .001$.

Table H.29

Correlation Matrix for Data Condition 29 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.64***	-			
Kappa	0.13***	0.40***	-		
AC1	0.28***	-0.30***	-0.79***	-	
Kripp.Alpha	0.01	0.16***	0.96***	-0.77***	-

Note. *** $p < .001$.

Table H.30

Correlation Matrix for Data Condition 30 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.71***	-			
Kappa	0.17***	0.42***	-		
AC1	0.33***	-0.25***	-0.76***	-	
Kripp.Alpha	0.02	0.13***	0.94***	-0.71***	-

Note. *** $p < .001$.

Table H.31

Correlation Matrix for Data Condition 31 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.46***	-			
Kappa	0.03*	0.34***	-		
AC1	-0.19***	-0.35***	-0.78***	-	
Kripp.Alpha	0.00	0.27***	0.99***	-0.83***	-

Note. *** $p < .001$.

Table H.32

Correlation Matrix for Data Condition 32 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.47***	-			
Kappa	0.06***	0.35***	-		
AC1	-0.24***	-0.37***	-0.86***	-	
Kripp.Alpha	0.00	0.23***	0.98***	-0.89***	-

Note. *** $p < .001$.

Table H.33

Correlation Matrix for Data Condition 33 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.49***	-			
Kappa	0.09***	0.38***	-		
AC1	-0.28***	-0.39***	-0.89***	-	
Kripp.Alpha	0.01	0.19***	0.97***	-0.90***	-

Note. *** $p < .001$.

Table H.34

Correlation Matrix for Data Condition 34 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.49***	-			
Kappa	0.13***	0.40***	-		
AC1	-0.33***	-0.42***	-0.91***	-	
Kripp.Alpha	0.01	0.16***	0.96***	-0.90***	-

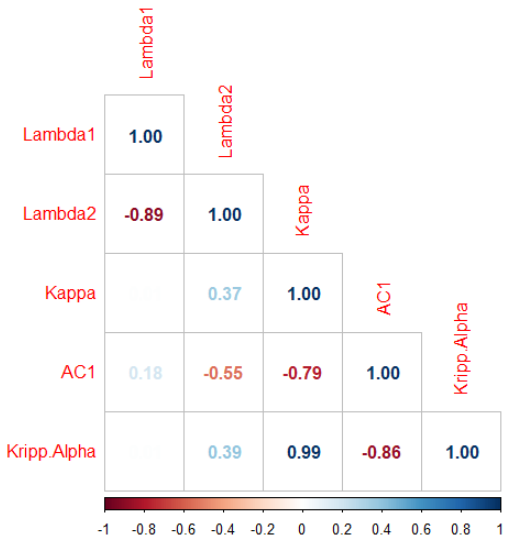
Note. *** $p < .001$.

Table H.35

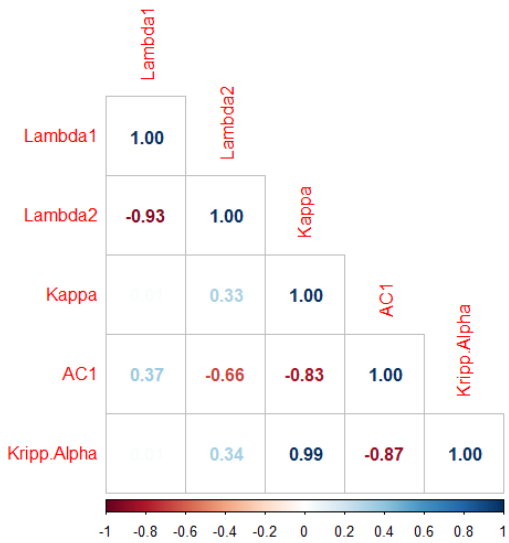
Correlation Matrix for Data Condition 35 (3x3)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.50***	-			
Kappa	0.16***	0.44***	-		
AC1	-0.38***	-0.46***	-0.92***	-	
Kripp.Alpha	0.01	0.13***	0.94***	-0.88***	-

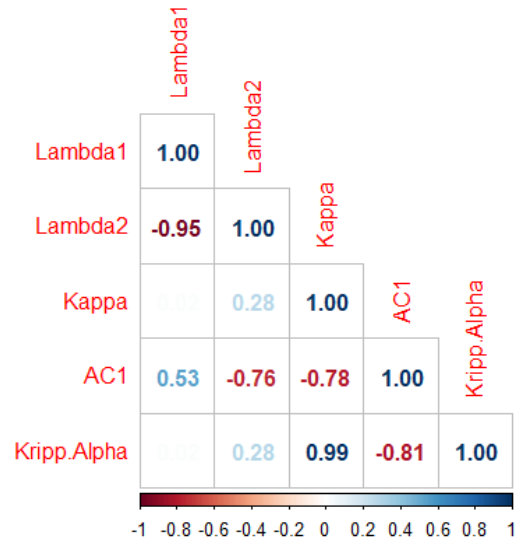
Note. *** $p < .001$.



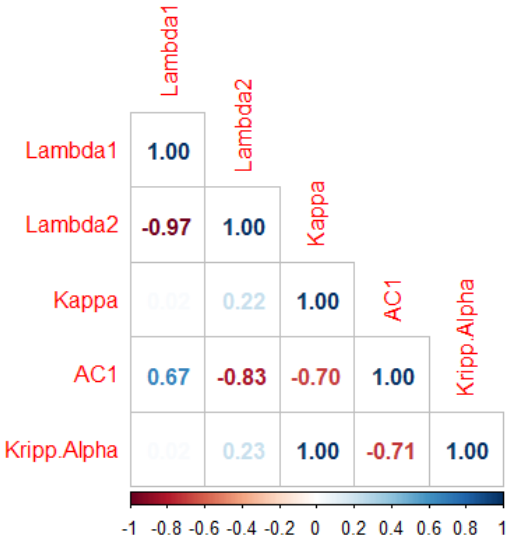
DC1



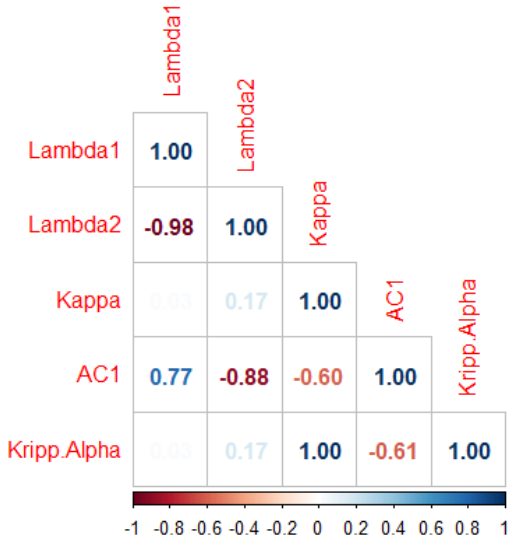
DC2



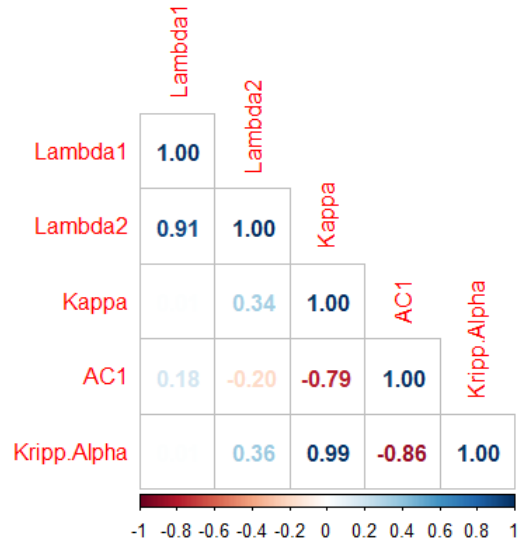
DC3



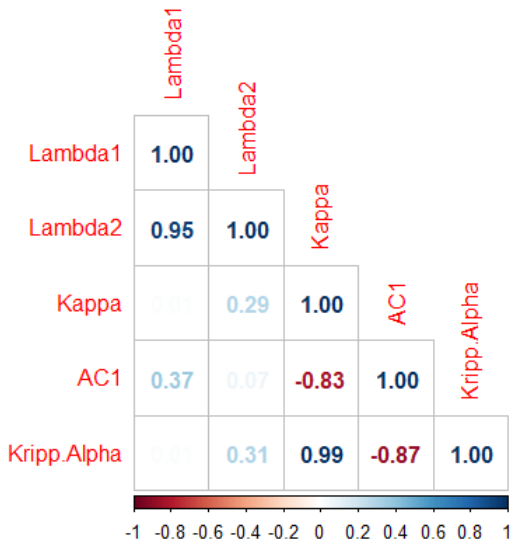
DC4



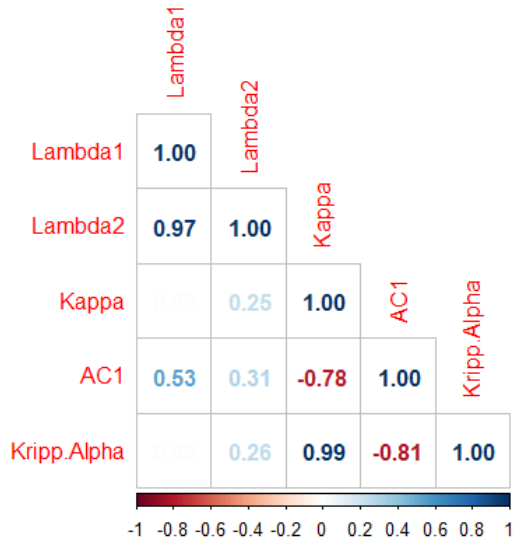
DC5



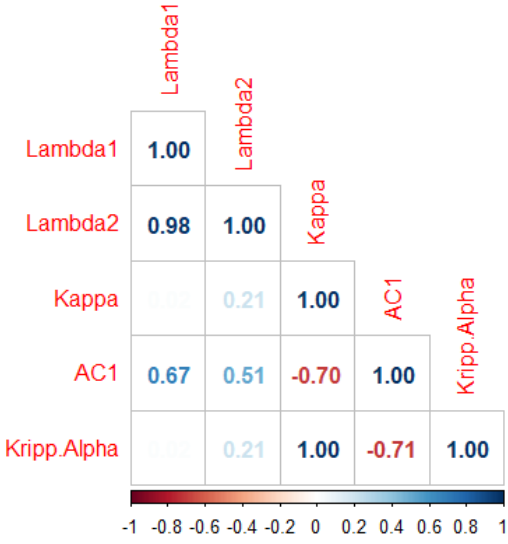
DC6



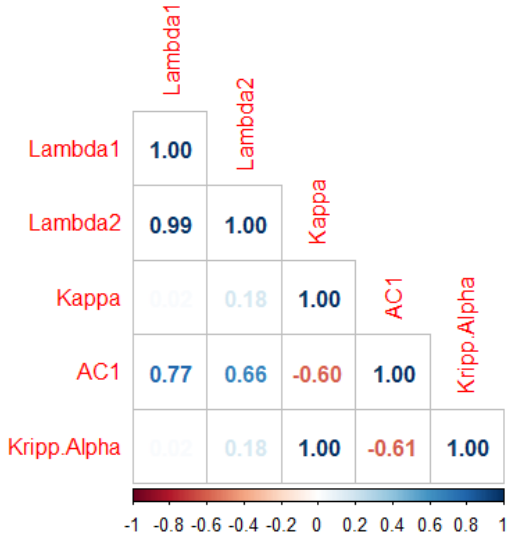
DC7



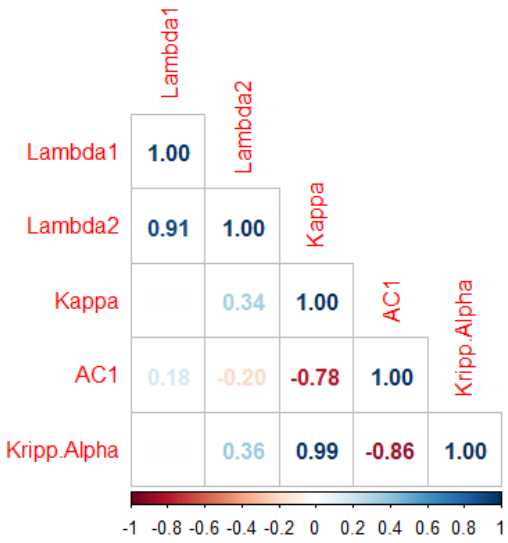
DC8



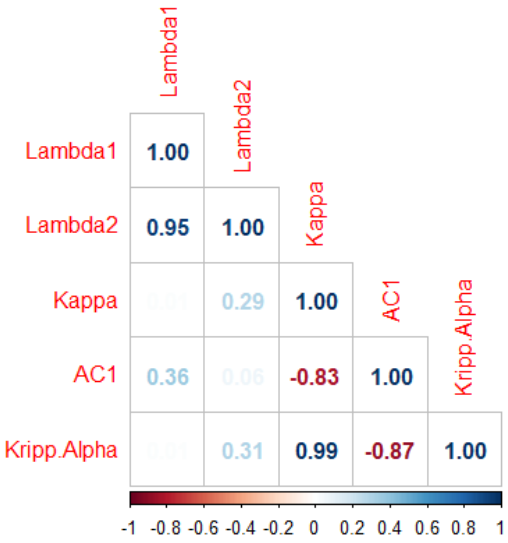
DC9



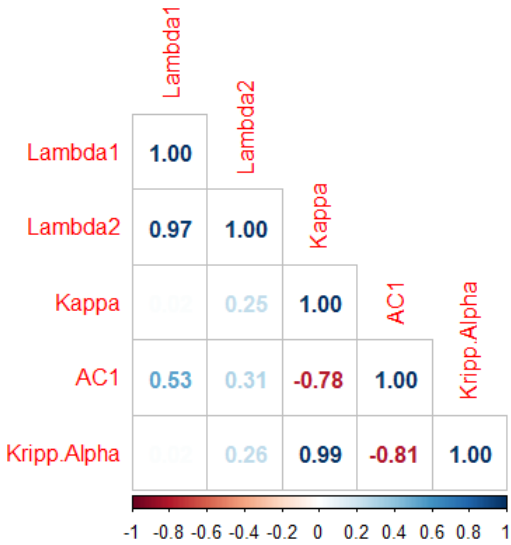
DC10



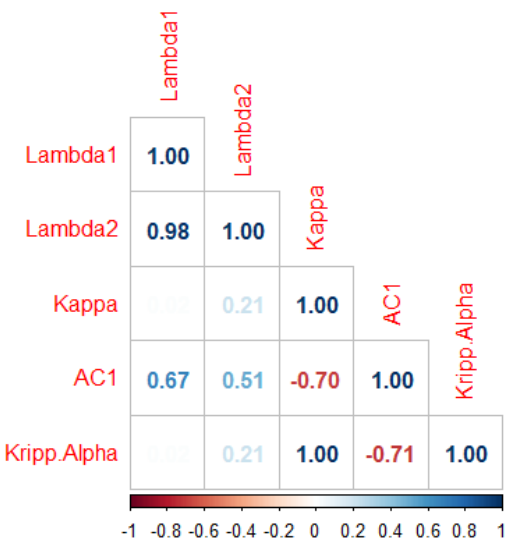
DC11



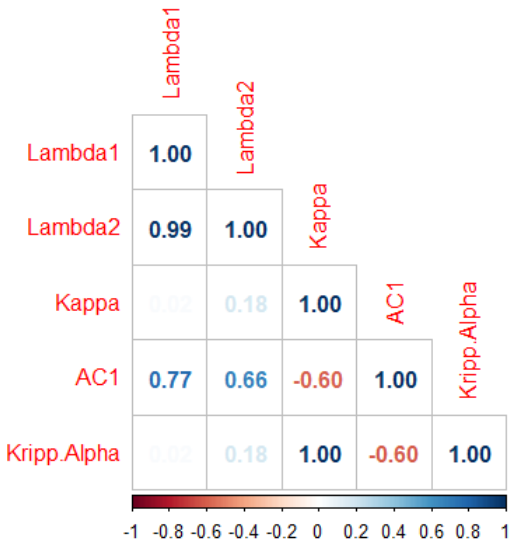
DC12



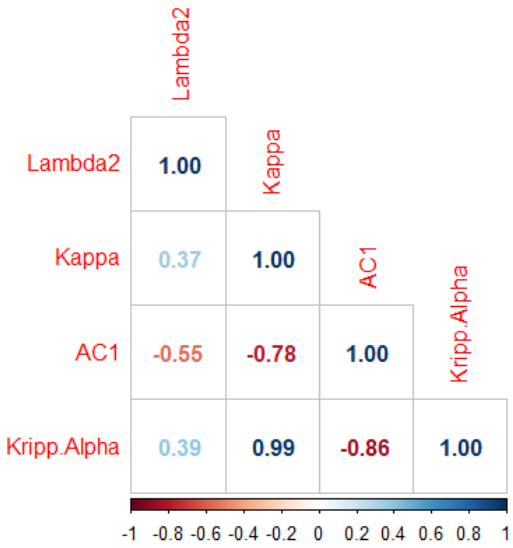
DC13



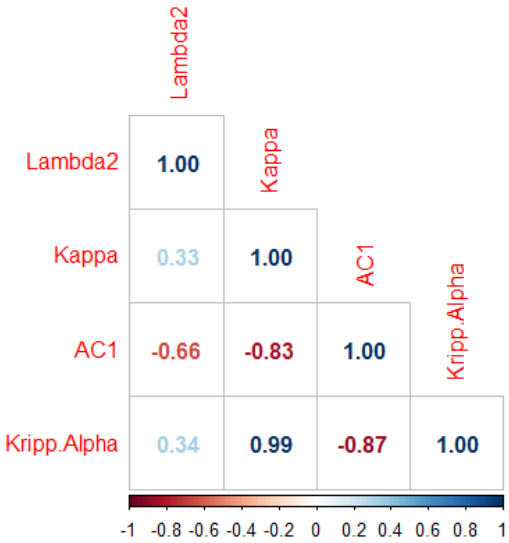
DC14



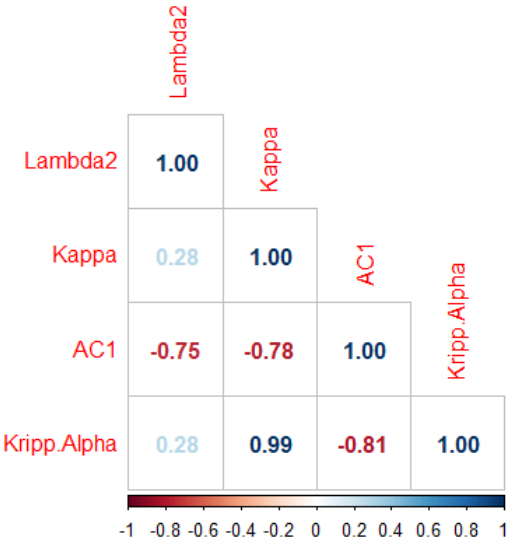
DC15



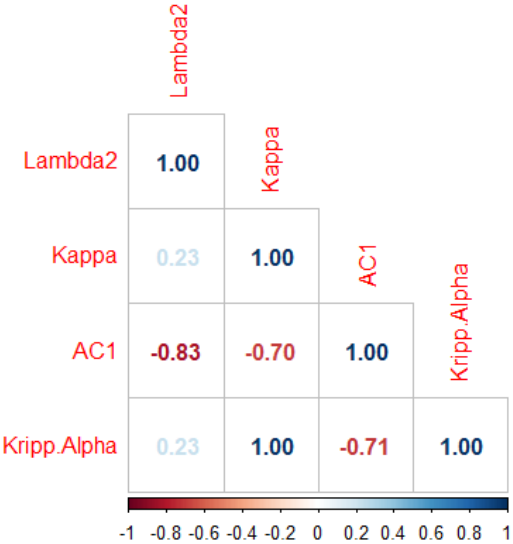
DC16



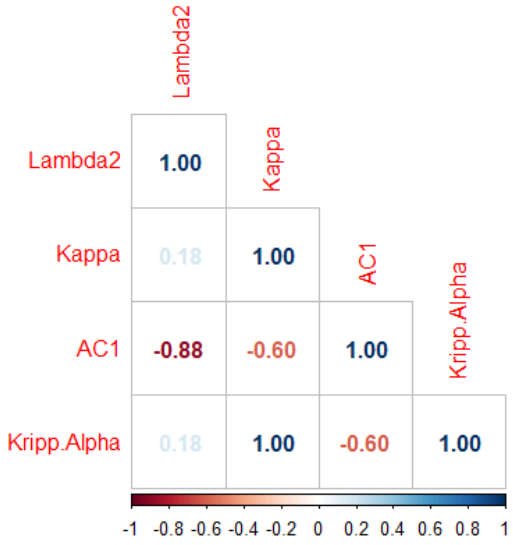
DC17



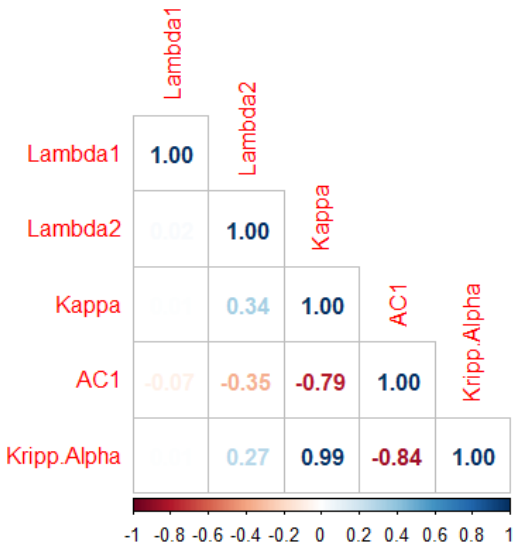
DC18



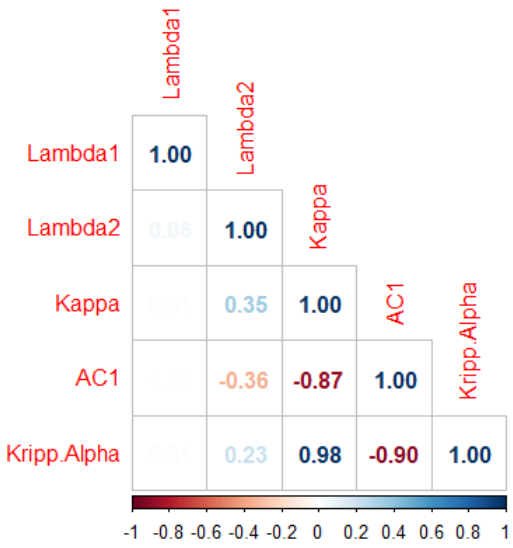
DC19



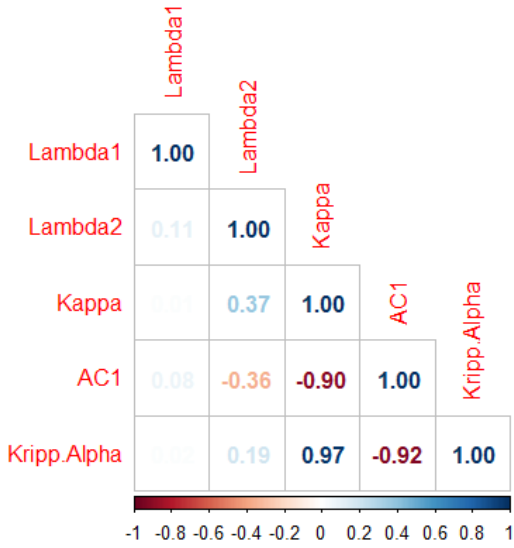
DC20



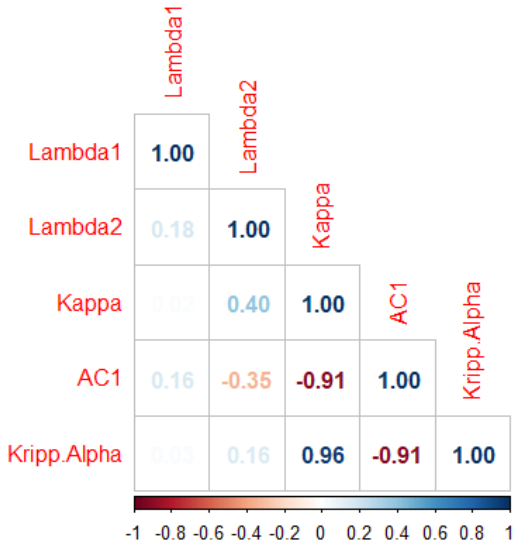
DC21



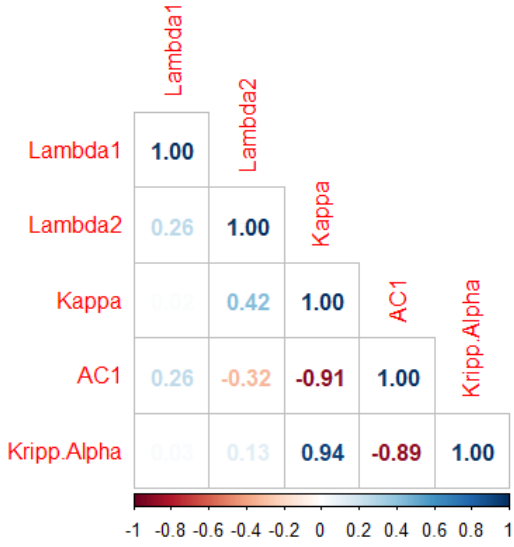
DC22



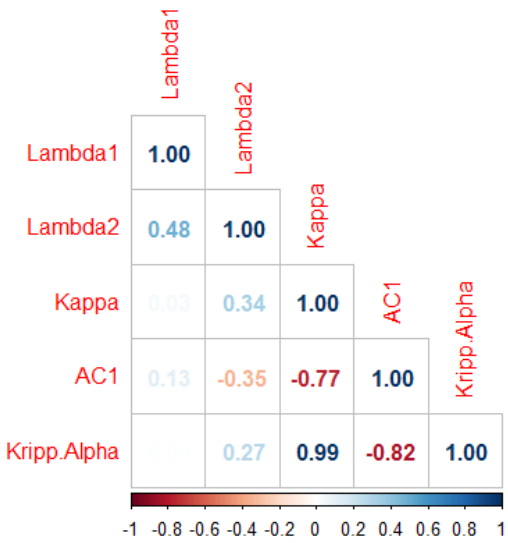
DC23



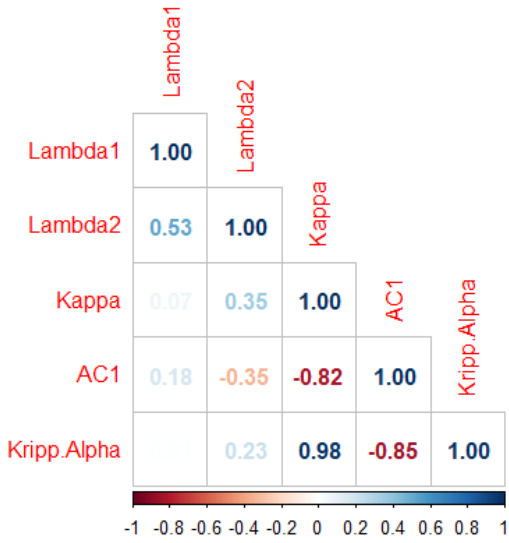
DC24



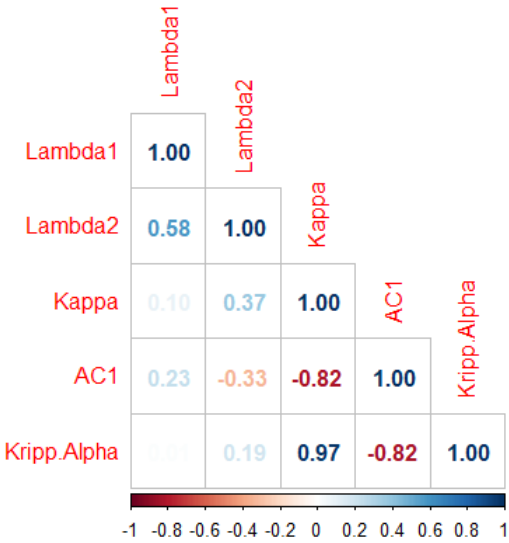
DC25



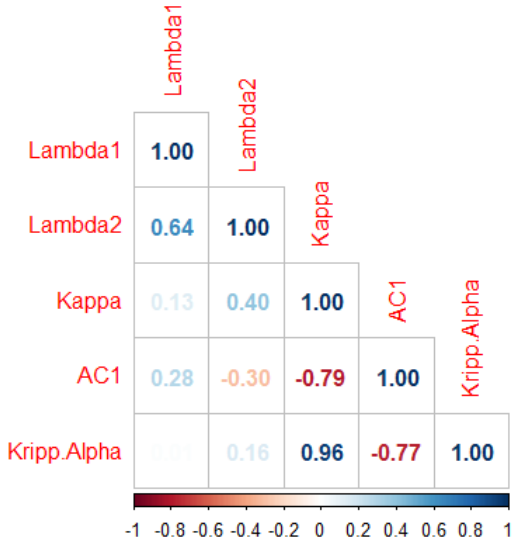
DC26



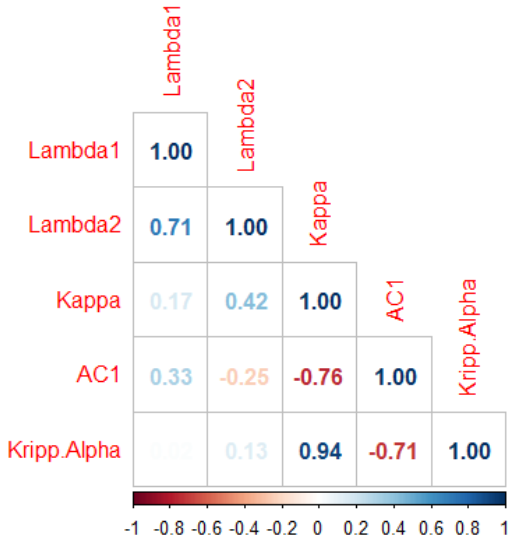
DC27



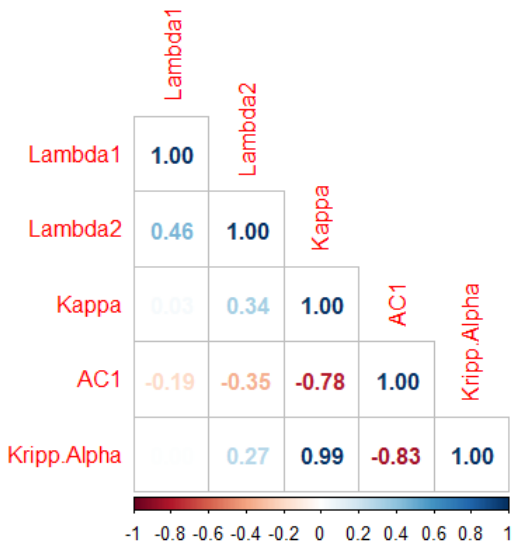
DC28



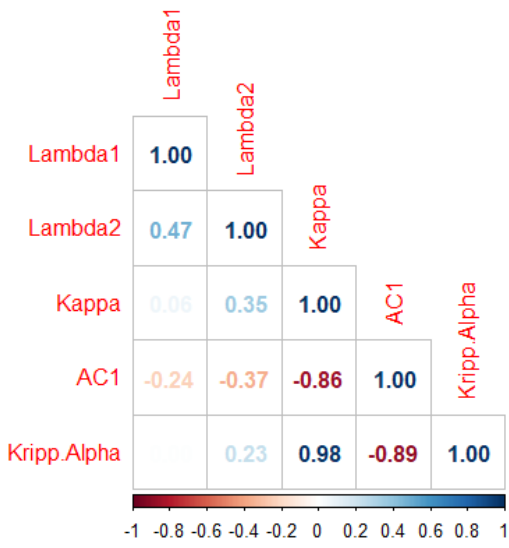
DC29



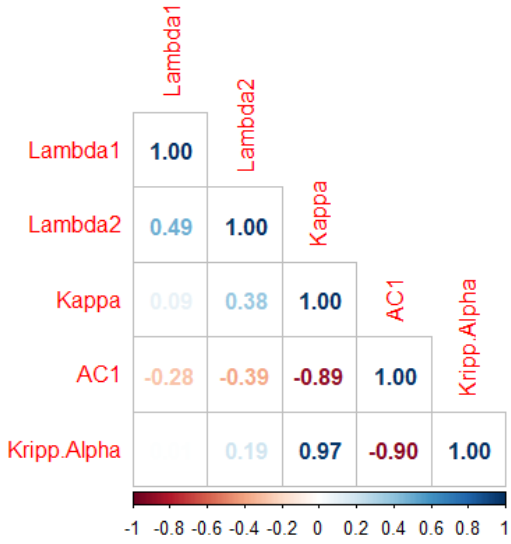
DC30



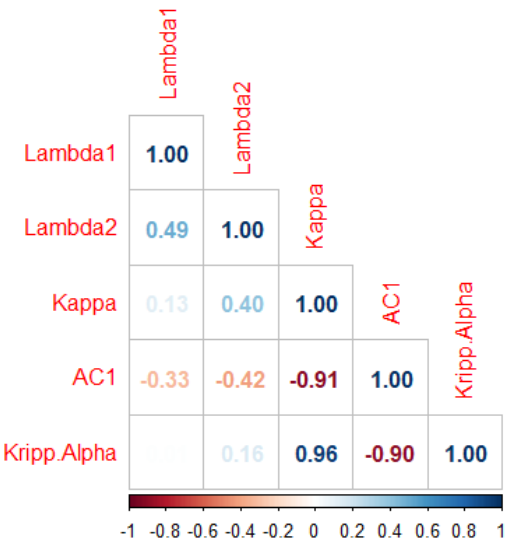
DC31



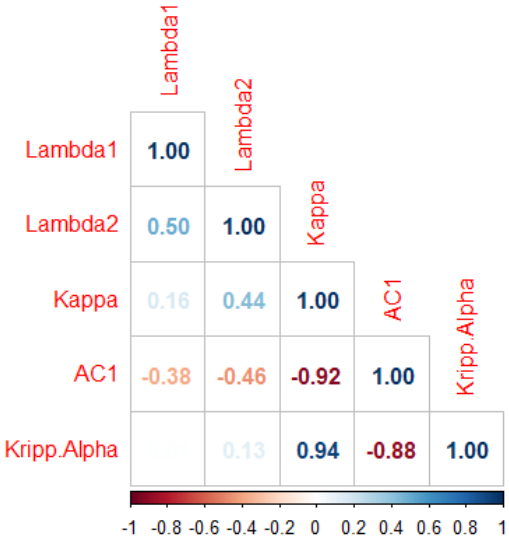
DC32



DC33



DC34



DC35

APPENDIX I

CORRELATION TABLES AND FIGURES (4X4 AGREEMENT MATRICES)

Table I.1

Correlation Matrix for Data Condition 1 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	1.00***	-			
Kappa	0.01	0.01	-		
AC1	0.18	0.18	-0.79***	-	
Kripp.Alpha	0.01	0.01	0.99***	-0.86***	-

Note. *** $p < .001$.

Table I.2

Correlation Matrix for Data Condition 2 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	1.00***	-			
Kappa	0.01	0.02	-		
AC1	0.36***	0.36***	-0.83***	-	
Kripp.Alpha	0.01	0.02	0.99***	-0.87***	-

Note. *** $p < .001$.

Table I.3

Correlation Matrix for Data Condition 3 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	1.00***	-			
Kappa	0.02	0.02	-		
AC1	0.53***	0.53***	-0.78***	-	
Kripp.Alpha	0.02	0.02	0.99***	-0.81***	-

Note. *** $p < .001$.

Table I.4

Correlation Matrix for Data Condition 4 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	1.00***	-			
Kappa	0.03	0.03	-		
AC1	0.67***	0.66***	-0.70***	-	
Kripp.Alpha	0.03	0.03	1.00***	-0.71***	-

Note. *** $p < .001$.

Table I.5

Correlation Matrix for Data Condition 5 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	1.00***	-			
Kappa	0.03	0.04	-		
AC1	0.77***	0.76***	-0.60***	-	
Kripp.Alpha	0.03	0.04	1.00***	-0.61***	-

Note. *** $p < .001$.

Table I.6

Correlation Matrix for Data Condition 6 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	1.00***	-			
Kappa	0.00	0.00	-		
AC1	0.18	0.19	-0.79***	-	
Kripp.Alpha	0.00	0.00	0.99***	-0.86***	-

Note. *** $p < .001$.

Table I.7

Correlation Matrix for Data Condition 7 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	1.00***	-			
Kappa	0.01	0.00	-		
AC1	0.37***	0.38***	-0.83***	-	
Kripp.Alpha	0.01	0.00	0.99***	-0.87***	-

Note. *** $p < .001$.

Table I.8

Correlation Matrix for Data Condition 8 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	1.00***	-			
Kappa	0.01	0.00	-		
AC1	0.54***	0.55***	-0.78***	-	
Kripp.Alpha	0.01	0.00	0.99***	-0.81***	-

Note. *** $p < .001$.

Table I.9

Correlation Matrix for Data Condition 9 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	1.00***	-			
Kappa	0.01	0.00	-		
AC1	0.68***	0.68***	-0.70***	-	
Kripp.Alpha	0.01	0.00	1.00	-0.71***	-

Note. *** $p < .001$.

Table I.10

Correlation Matrix for Data Condition 10 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	1.00***	-			
Kappa	0.02	0.00	-		
AC1	0.78***	0.79***	-0.60***	-	
Kripp.Alpha	0.02	0.00	1.00***	-0.61***	-

Note. *** $p < .001$.

Table I.11

Correlation Matrix for Data Condition 11 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.14	-		
AC1	0.02	-0.79***	-	
Kripp.Alpha	0.15	0.99***	-0.86***	-

Note. *** $p < .001$.

Table I.12

Correlation Matrix for Data Condition 12 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.14	-		
AC1	0.23*	-0.83***	-	
Kripp.Alpha	0.14	0.99***	-0.87***	-

Note. * $p < .05$; *** $p < .001$.

Table I.13

Correlation Matrix for Data Condition 13 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.13	-		
AC1	0.43***	-0.78***	-	
Kripp.Alpha	0.13	0.99***	-0.81***	-

Note. *** $p < .001$.

Table I.14

Correlation Matrix for Data Condition 14 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.11	-		
AC1	0.60***	-0.70***	-	
Kripp.Alpha	0.12	1.00***	-0.71***	-

Note. *** $p < .001$.

Table I.15

Correlation Matrix for Data Condition 15 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.10	-		
AC1	0.72***	-0.60***	-	
Kripp.Alpha	0.10	1.00***	-0.61***	-

Note. *** $p < .001$.

Table I.16

Correlation Matrix for Data Condition 16 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	-0.72***	-		
AC1	0.88***	-0.87***	-	
Kripp.Alpha	-0.76***	1.00***	-0.91***	-

Note. *** $p < .001$.

Table I.17

Correlation Matrix for Data Condition 17 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	-0.79***	-		
AC1	0.87***	-0.94***	-	
Kripp.Alpha	-0.82***	1.00***	-0.96***	-

Note. *** $p < .001$.

Table I.18

Correlation Matrix for Data Condition 18 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.13	-		
AC1	-0.13	-0.95***	-	
Kripp.Alpha	0.13	1.00***	-0.97***	-

Note. *** $p < .001$.

Table I.19

Correlation Matrix for Data Condition 19 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.08	-		
AC1	-0.07	-0.97***	-	
Kripp.Alpha	0.12	1.00***	-0.98***	-

Note. *** $p < .001$.

Table I.20

Correlation Matrix for Data Condition 20 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.10	-		
AC1	-0.06	-0.98***	-	
Kripp.Alpha	0.10	1.00***	-0.99***	-

Note. *** $p < .001$.

Table I.21

Correlation Matrix for Data Condition 21 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.13	-		
AC1	-0.34***	-0.79***	-	
Kripp.Alpha	0.15	0.99***	-0.86***	-

Note. *** $p < .001$.

Table I.22

Correlation Matrix for Data Condition 22 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.12	-		
AC1	-0.49***	-0.83***	-	
Kripp.Alpha	0.13	0.99***	-0.87***	-

Note. *** $p < .001$.

Table I.23

Correlation Matrix for Data Condition 23 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.10	-		
AC1	-0.63***	-0.78***	-	
Kripp.Alpha	0.10	0.99***	-0.81***	-

Note. *** $p < .001$.

Table I.24

Correlation Matrix for Data Condition 24 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.07	-		
AC1	-0.74***	-0.70***	-	
Kripp.Alpha	0.08	1.00***	-0.71***	-

Note. *** $p < .001$.

Table I.25

Correlation Matrix for Data Condition 25 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.05	-		
AC1	-0.82***	-0.60***	-	
Kripp.Alpha	0.05	1.00***	-0.61***	-

Note. *** $p < .001$.

Table I.26

Correlation Matrix for Data Condition 26 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.09	-		
AC1	-0.16	-0.83***	-	
Kripp.Alpha	0.14	0.99***	-0.88***	-

Note. *** $p < .001$.

Table I.27

Correlation Matrix for Data Condition 27 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.12	-		
AC1	-0.16	-0.91***	-	
Kripp.Alpha	0.12	1.00***	-0.94***	-

Note. *** $p < .001$.

Table I.28

Correlation Matrix for Data Condition 28 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.09	-		
AC1	-0.09	-0.95***	-	
Kripp.Alpha	0.09	1.00***	-0.97***	-

Note. *** $p < .001$.

Table I.29

Correlation Matrix for Data Condition 29 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.04	-		
AC1	-0.07	-0.97***	-	
Kripp.Alpha	0.07	1.00***	-0.98***	-

Note. *** $p < .001$.

Table I.30

Correlation Matrix for Data Condition 30 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda2	-			
Kappa	0.05	-		
AC1	-0.10	-0.98***	-	
Kripp.Alpha	0.05	1.00***	-0.99***	-

Note. *** $p < .001$.

Table I.31

Correlation Matrix for Data Condition 51 (4x4)

	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-			
Lambda2	0.17***	-		
Kappa	-0.35***	0.20***	-	
AC1	0.43***	-0.14***	-0.99***	-
Kripp.Alpha	-0.36***	0.21***	1.00***	-0.99***

Note. *** $p < .001$.

Table I.32

Correlation Matrix for Data Condition 52 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.16***	-			
Kappa	-0.30***	0.16***	-		
AC1	0.47***	-0.04***	-0.97***	-	
Kripp.Alpha	-0.33***	0.18***	1.00***	-0.98***	-

Note. *** $p < .001$.

Table I.33

Correlation Matrix for Data Condition 52 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.18***	-			
Kappa	0.09***	0.01*	-		
AC1	0.30***	0.20***	-0.90***	-	
Kripp.Alpha	0.01	0.04***	0.99***	-0.93***	-

Note. * $p < .05$; *** $p < .001$.

Table I.34

Correlation Matrix for Data Condition 54 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.20***	-			
Kappa	-0.21***	0.07***	-		
AC1	0.57***	0.18***	-0.90***	-	
Kripp.Alpha	-0.29***	0.12***	0.99***	-0.92***	-

Note. *** $p < .001$.

Table I.35

Correlation Matrix for Data Condition 55 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.22***	-			
Kappa	-0.11***	0.06***	-		
AC1	0.61***	0.27***	-0.82***	-	
Kripp.Alpha	-0.24***	0.11***	0.98***	-0.87***	-

Note. *** $p < .001$.

Table I.36

Correlation Matrix for Data Condition 56 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.56***	-			
Kappa	0.38***	-0.09***	-		
AC1	-0.31***	0.09***	-0.99***	-	
Kripp.Alpha	0.36***	-0.11***	1.00***	-0.99***	-

Note. *** $p < .001$.

Table I.37

Correlation Matrix for Data Condition 57 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.66***	-			
Kappa	0.33***	-0.04***	-		
AC1	-0.17***	0.05***	-0.97***	-	
Kripp.Alpha	0.27***	-0.09***	1.00***	-0.98***	-

Note. *** $p < .001$.

Table I.38

Correlation Matrix for Data Condition 58 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.53***	-			
Kappa	0.11***	0.10***	-		
AC1	0.19***	-0.20***	-0.90***	-	
Kripp.Alpha	0.00	0.04***	0.99***	-0.93***	-

Note. *** $p < .001$.

Table I.39

Correlation Matrix for Data Condition 59 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.80***	-			
Kappa	0.30***	0.06***	-		
AC1	0.07***	0.04***	-0.87***	-	
Kripp.Alpha	0.16***	-0.06***	0.99***	-0.91***	-

Note. *** $p < .001$.

Table I.40

Correlation Matrix for Data Condition 60 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.88***	-			
Kappa	0.37***	0.24***	-		
AC1	0.14***	0.01	-0.78***	-	
Kripp.Alpha	0.17***	0.06***	0.98***	-0.86***	-

Note. *** $p < .001$.

Table I.41

Correlation Matrix for Data Condition 61 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.33***	-			
Kappa	0.21***	0.09***	-		
AC1	-0.19***	-0.08***	-1.00***	-	
Kripp.Alpha	0.21***	0.09***	1.00***	-1.00***	-

Note. *** $p < .001$.

Table I.42

Correlation Matrix for Data Condition 62 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.36***	-			
Kappa	0.22***	0.11***	-		
AC1	-0.21***	-0.10***	-1.00***	-	
Kripp.Alpha	0.22***	0.11***	1.00***	-1.00***	-

Note. *** $p < .001$.

Table I.43

Correlation Matrix for Data Condition 63 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.55***	-			
Kappa	0.00	0.04***	-		
AC1	0.00	-0.04***	-1.00***	-	
Kripp.Alpha	0.00	0.04***	1.00***	-1.00***	-

Note. *** $p < .001$.

Table I.44

Correlation Matrix for Data Condition 64 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.42***	-			
Kappa	0.20***	0.10***	-		
AC1	-0.19***	-0.09***	-1.00***	-	
Kripp.Alpha	0.20***	0.10***	1.00***	-1.00***	-

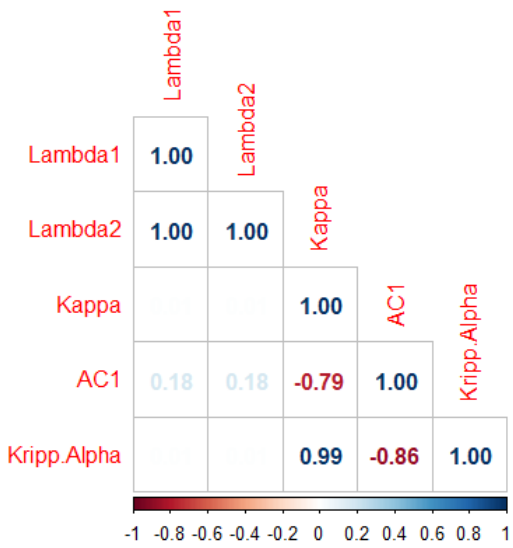
Note. *** $p < .001$.

Table I.45

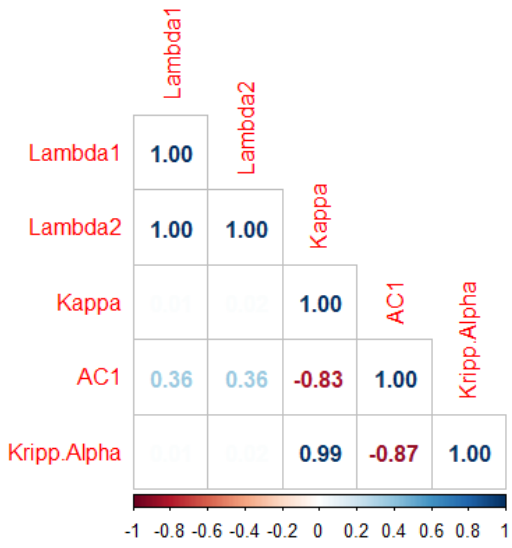
Correlation Matrix for Data Condition 65 (4x4)

	Lambda1	Lambda2	Kappa	AC1	Kripp.Alpha
Lambda1	-				
Lambda2	0.44***	-			
Kappa	0.21***	0.16***	-		
AC1	-0.21***	-0.16***	-1.00***	-	
Kripp.Alpha	0.21***	0.16***	1.00***	-1.00***	-

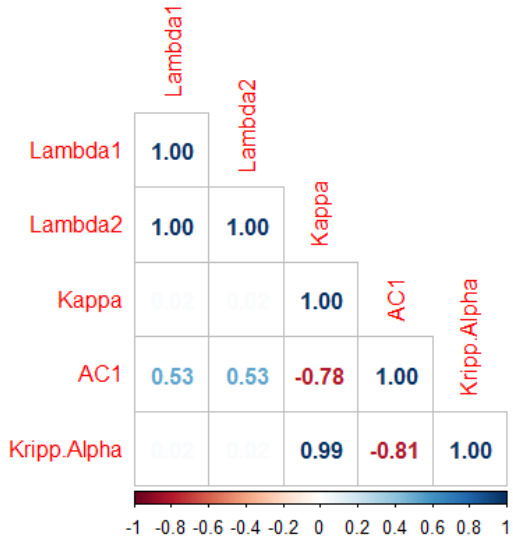
Note. *** $p < .001$.



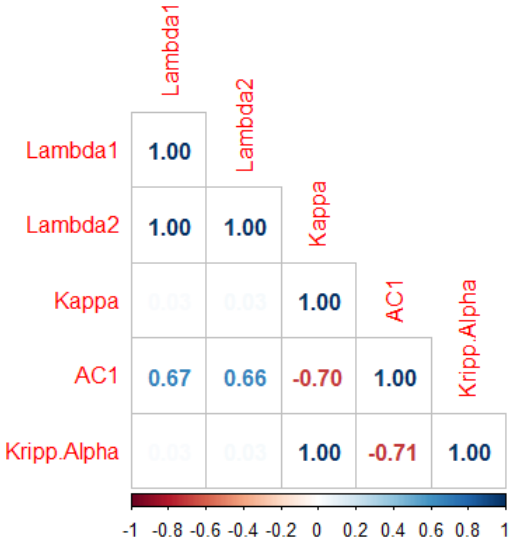
DC1



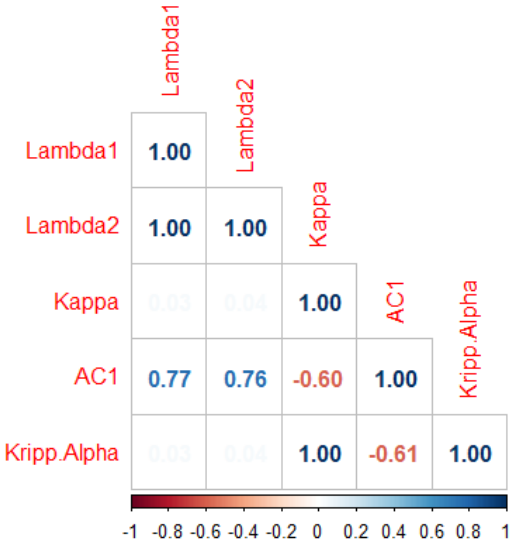
DC2



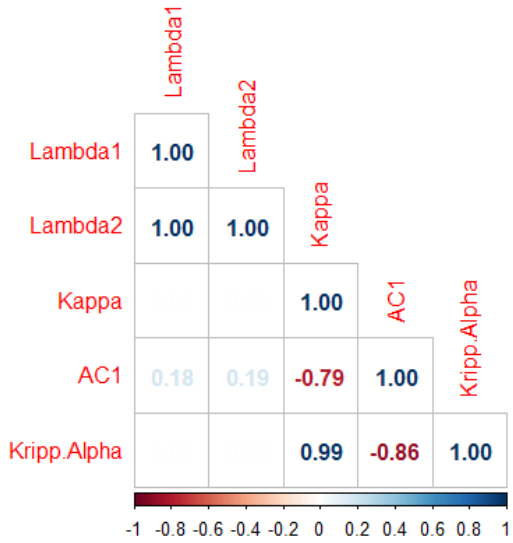
DC3



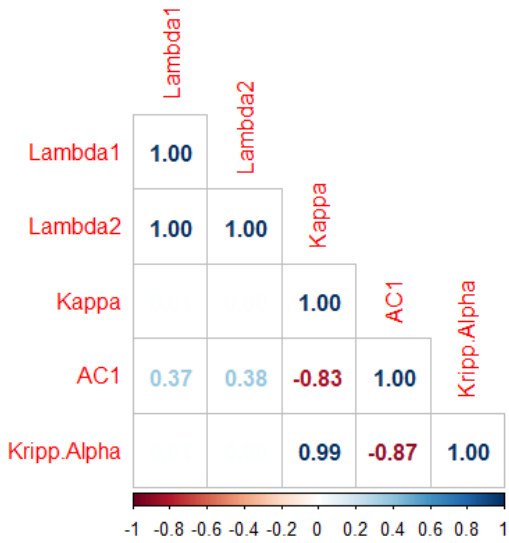
DC4



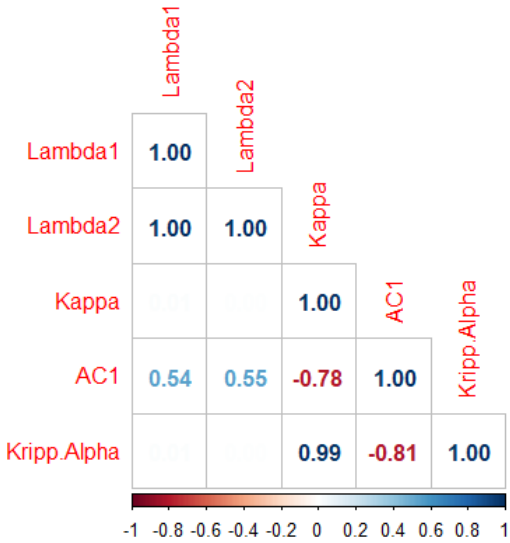
DC5



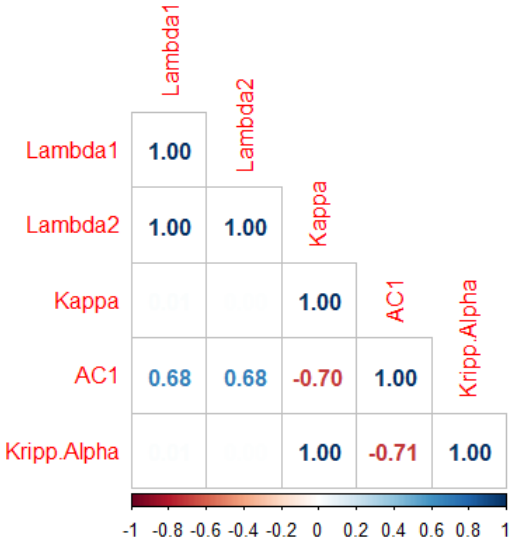
DC6



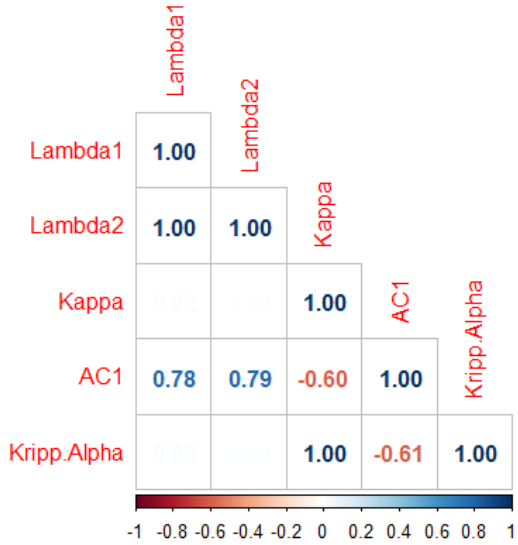
DC7



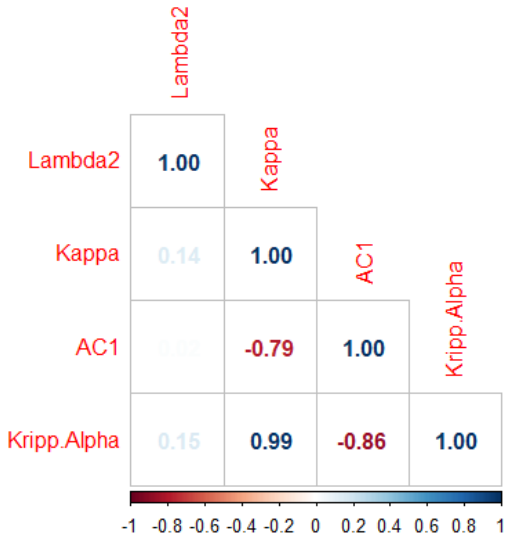
DC8



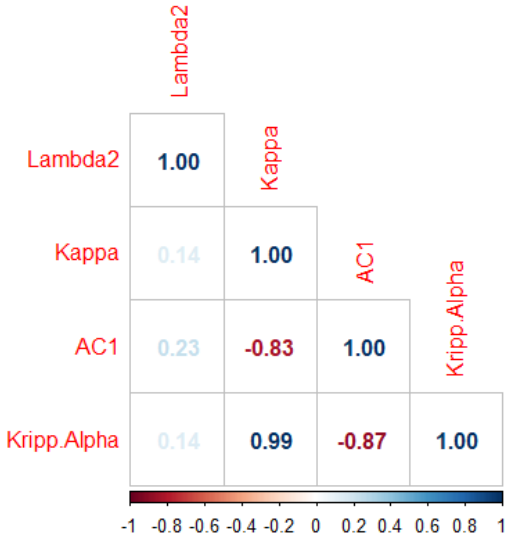
DC9



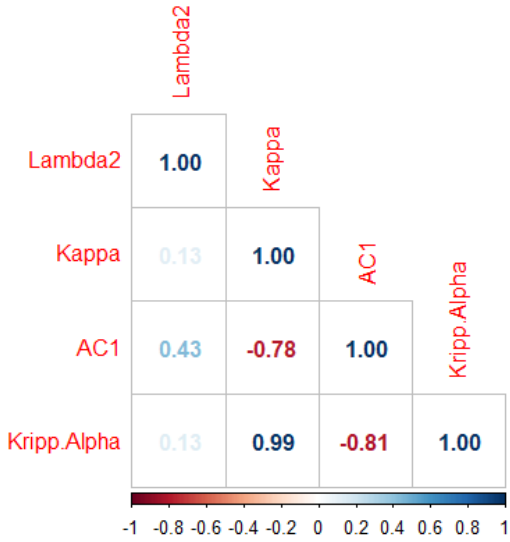
DC10



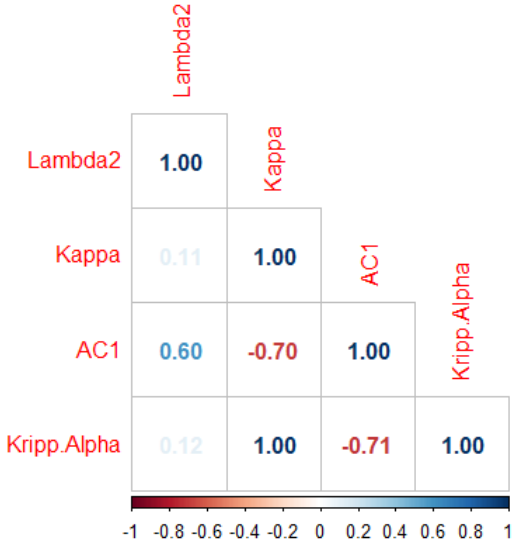
DC11



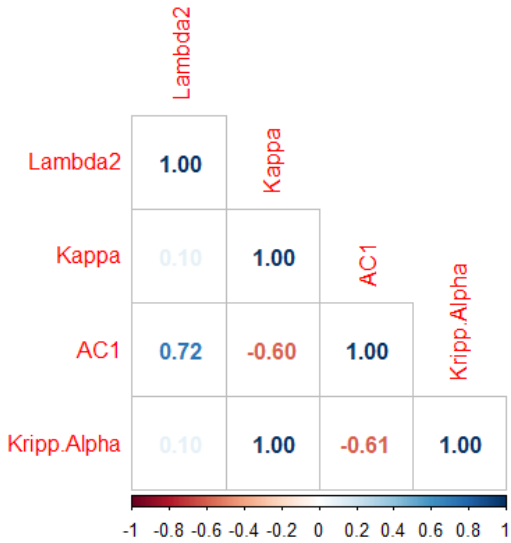
DC12



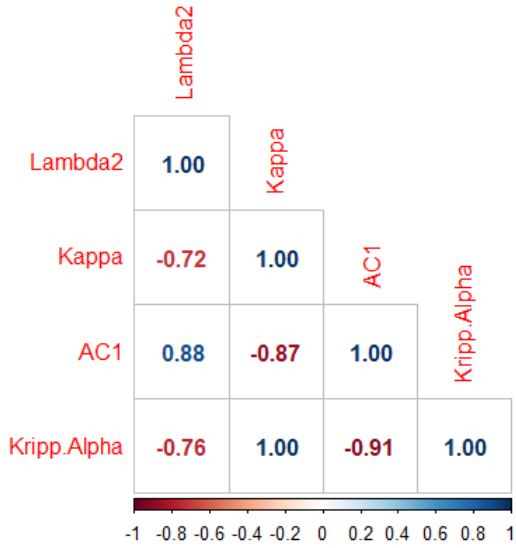
DC13



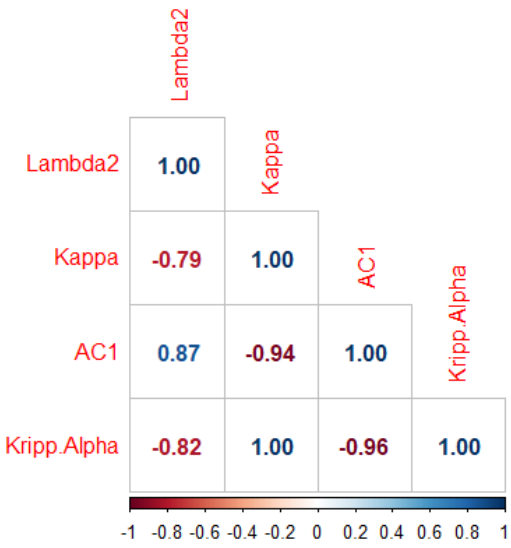
DC14



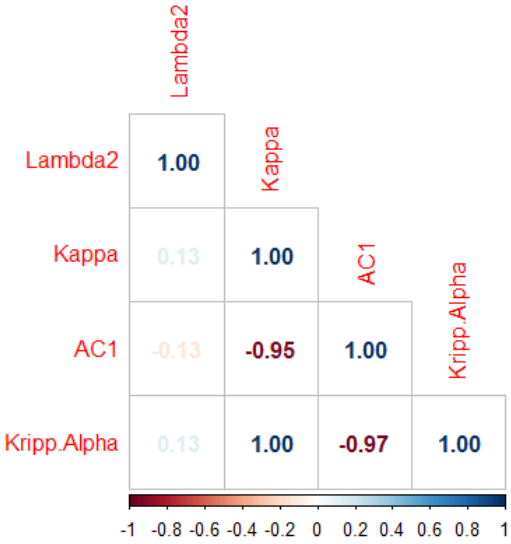
DC15



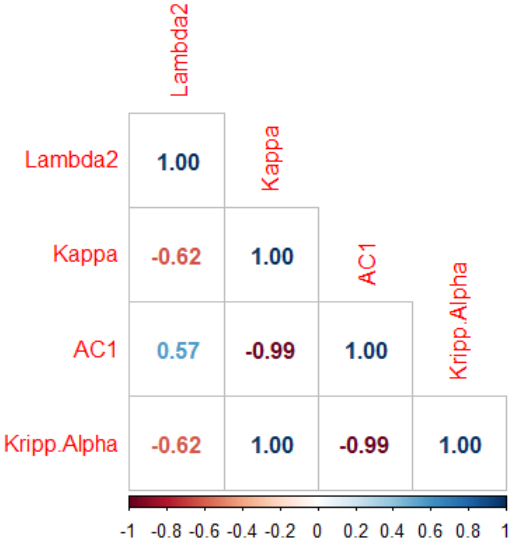
DC16



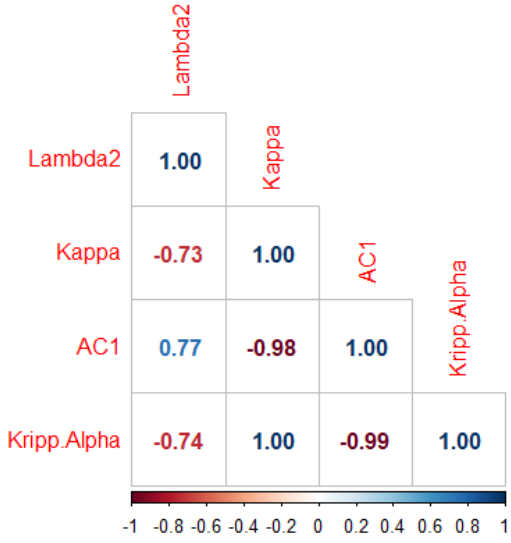
DC17



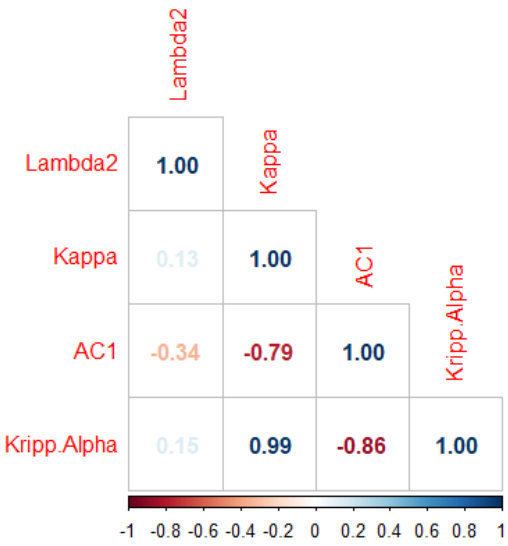
DC18



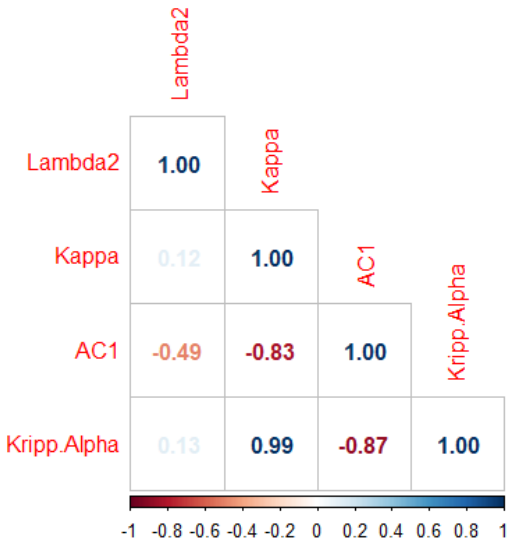
DC19



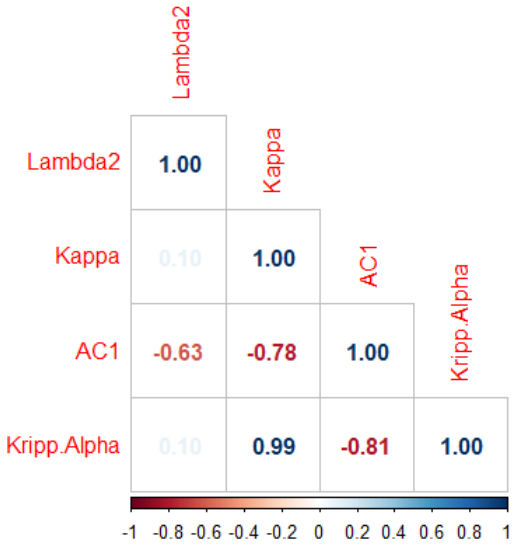
DC20



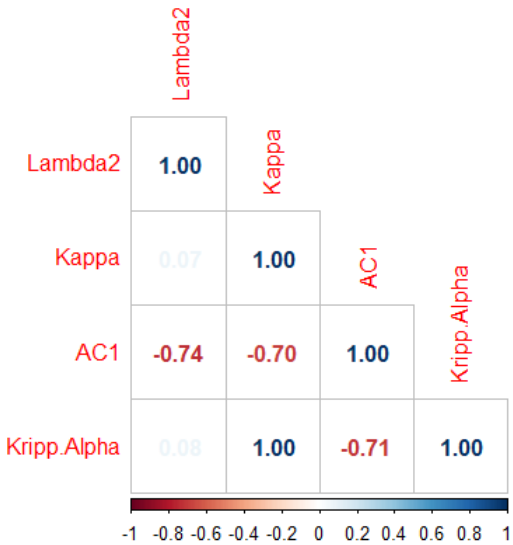
DC21



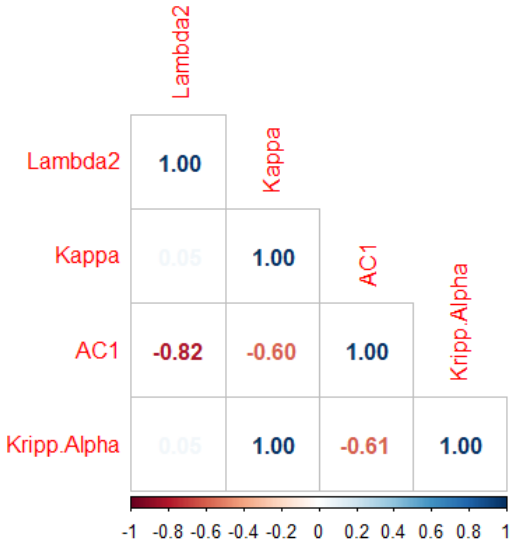
DC22



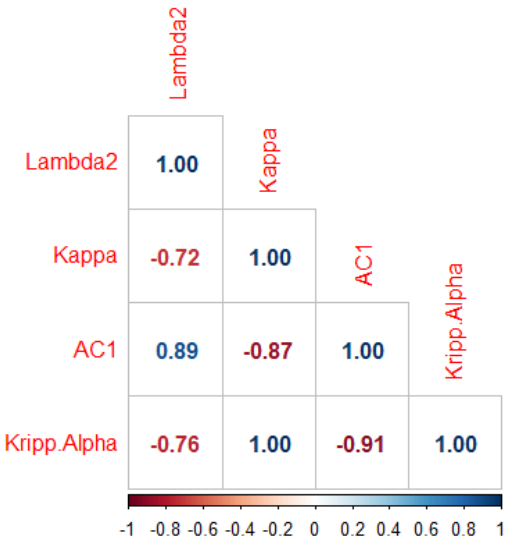
DC23



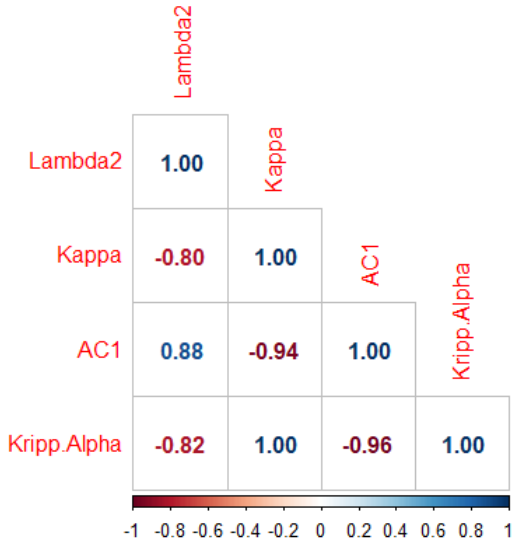
DC24



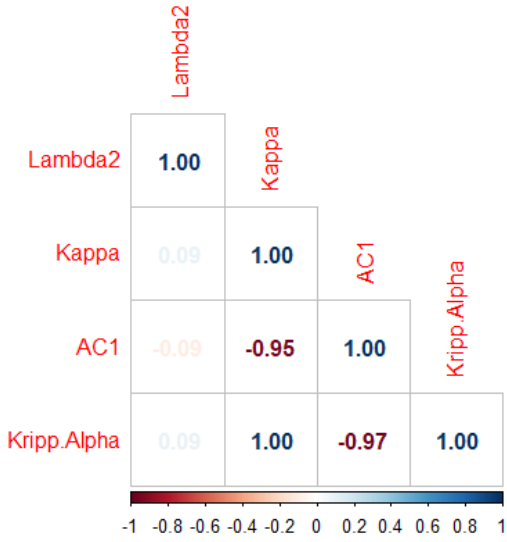
DC25



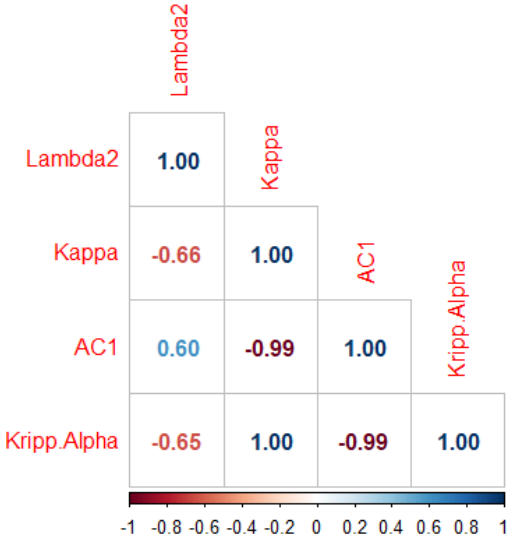
DC26



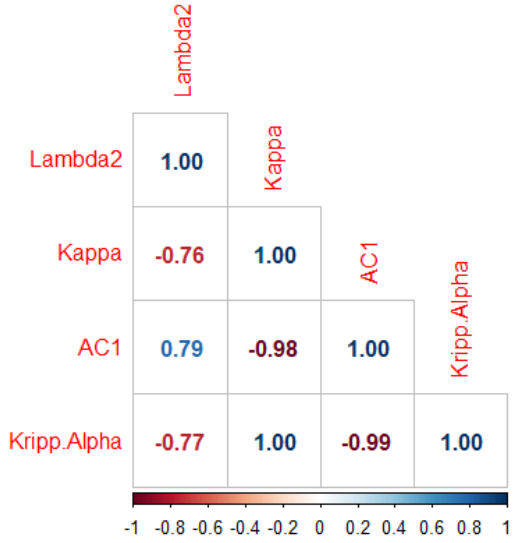
DC27



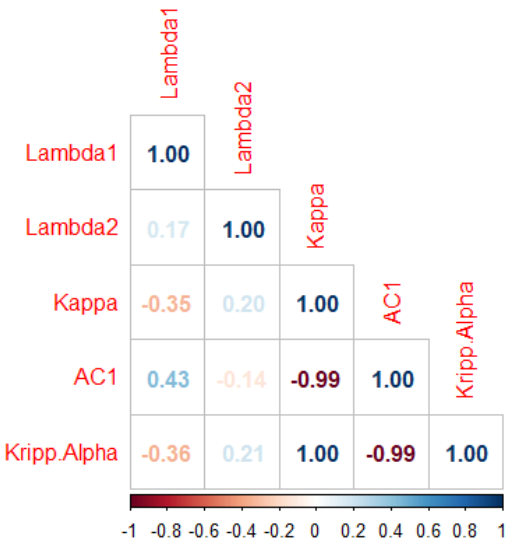
DC28



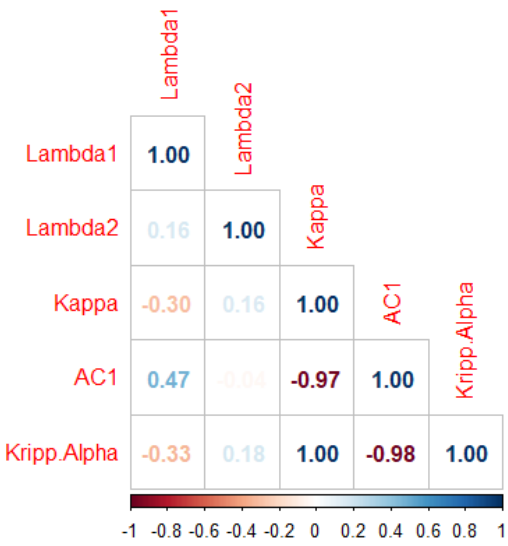
DC29



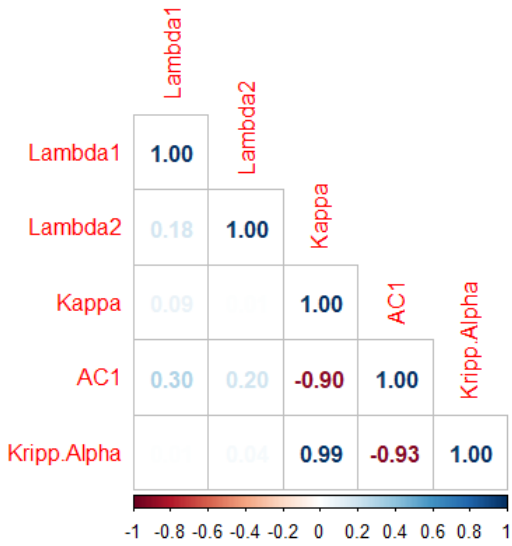
DC30



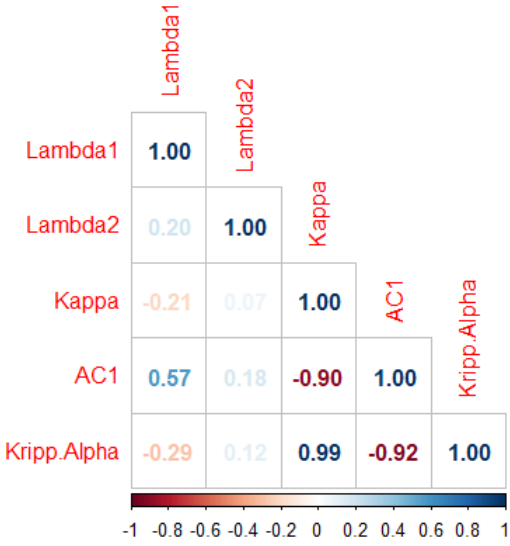
DC51



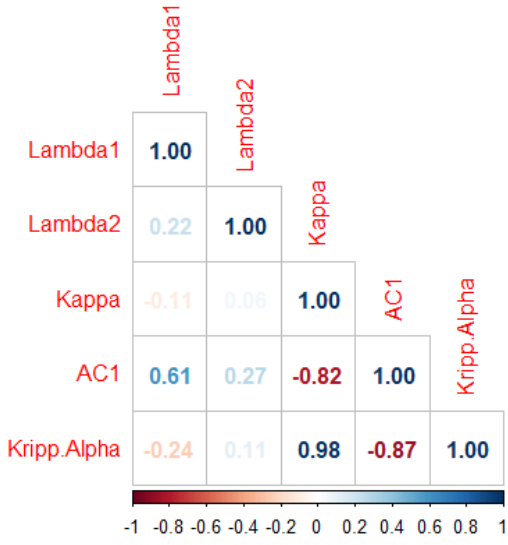
DC52



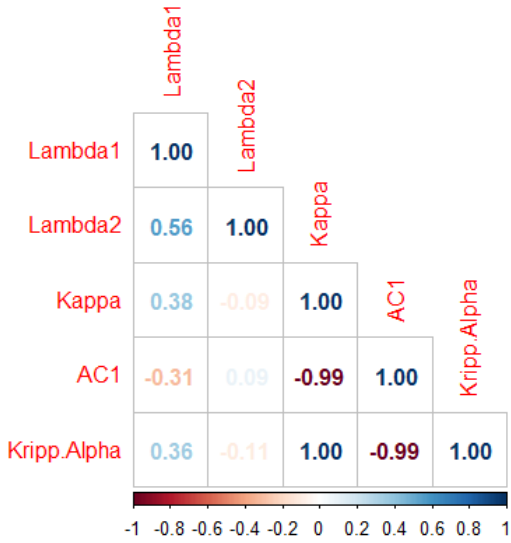
DC53



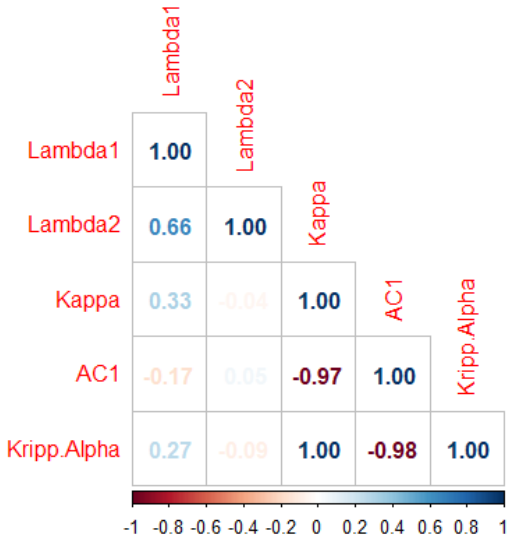
DC54



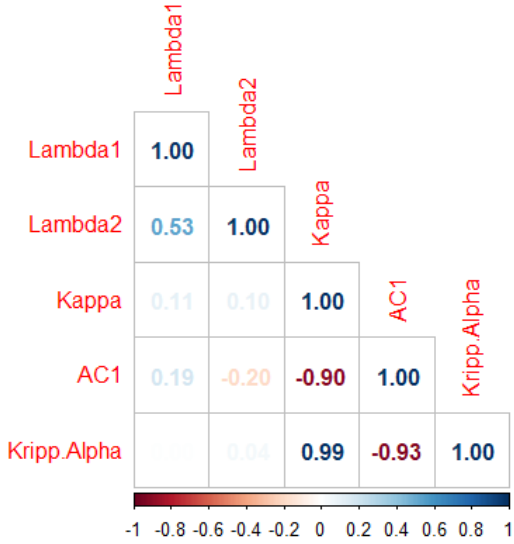
DC55



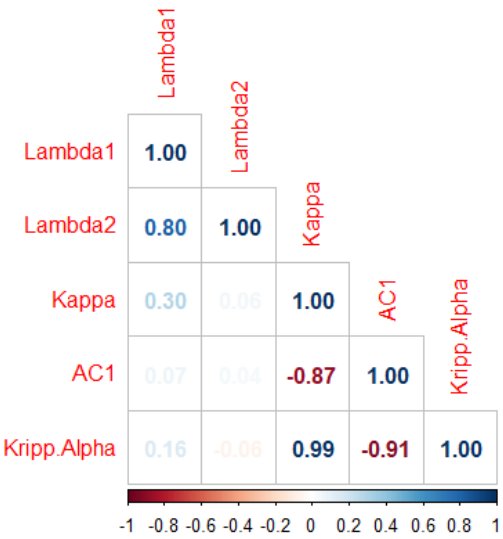
DC56



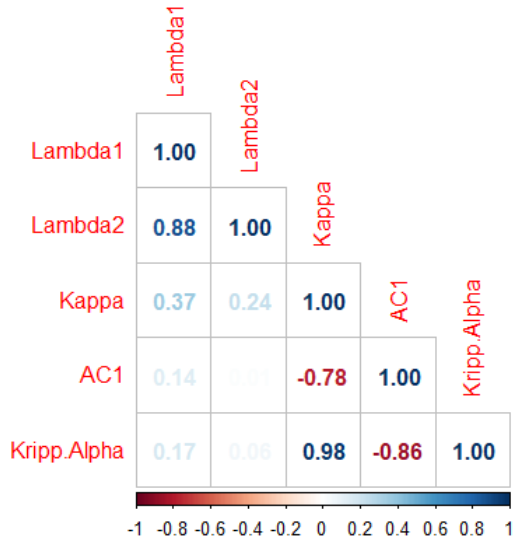
DC57



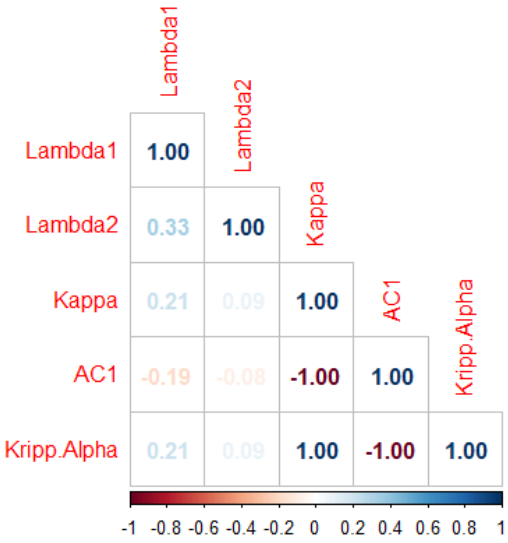
DC58



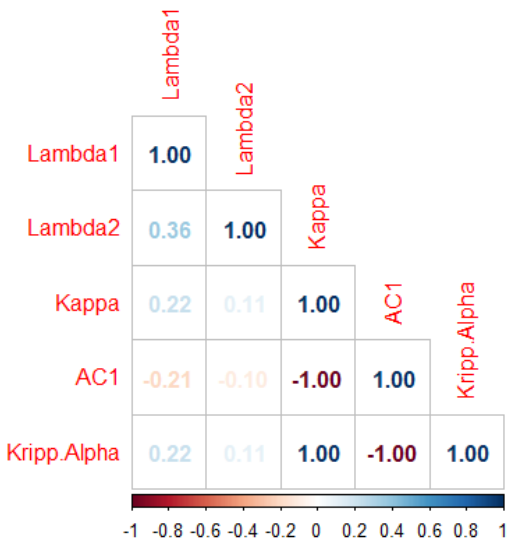
DC59



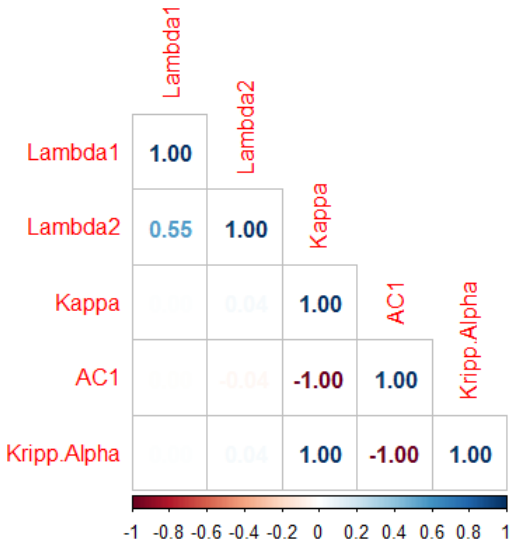
DC60



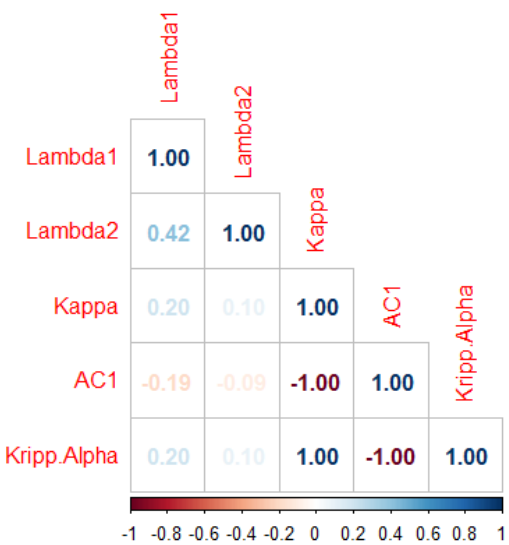
DC61



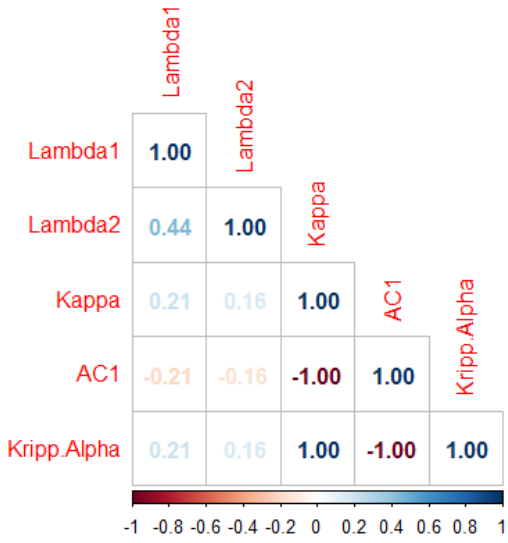
DC62



DC63



DC64



DC65

APPENDIX J

COEFFICIENT LINE GRAPHS

Line graphs demonstrating the performance of coefficients under select data conditions were produced to visualize how each coefficient measures agreement across the same conditions.

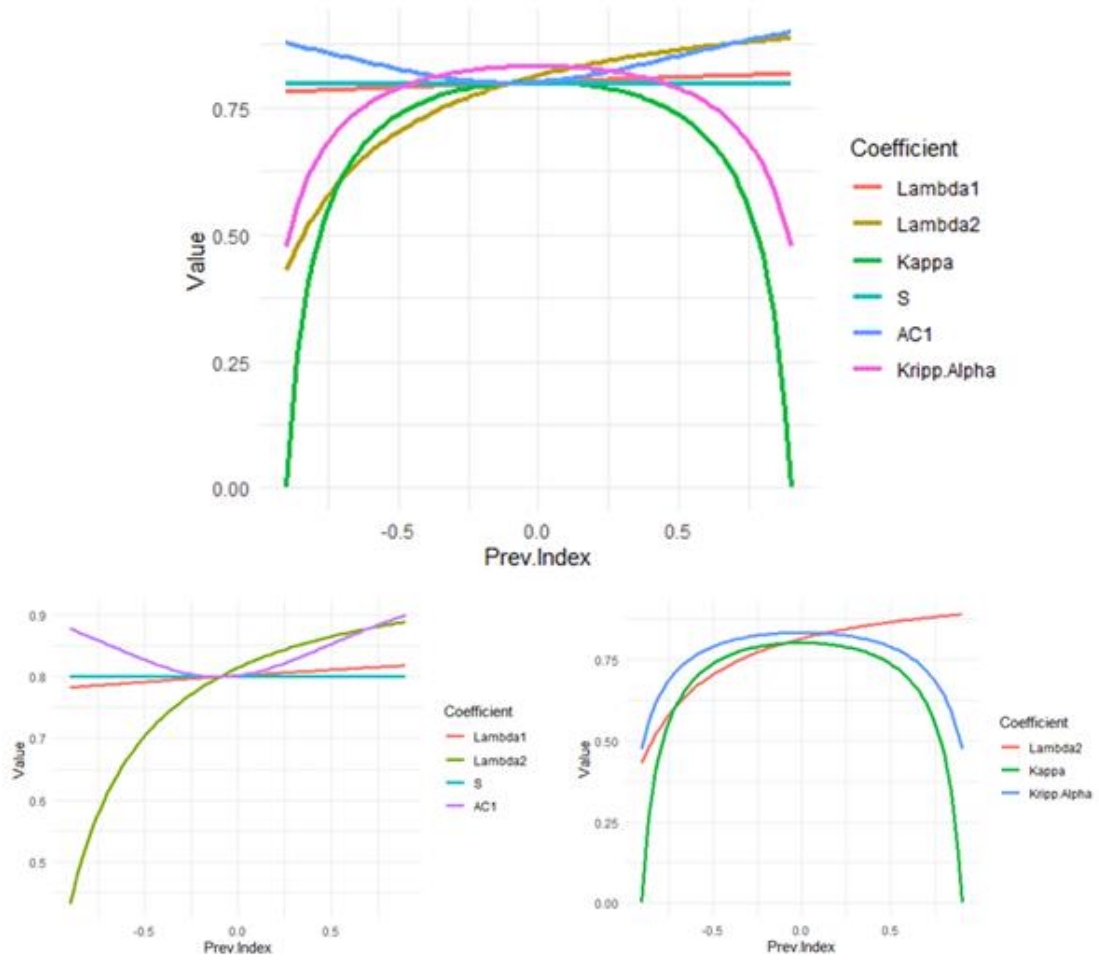


Figure J.1. Line Graphs for Data Condition 7 (2x2)

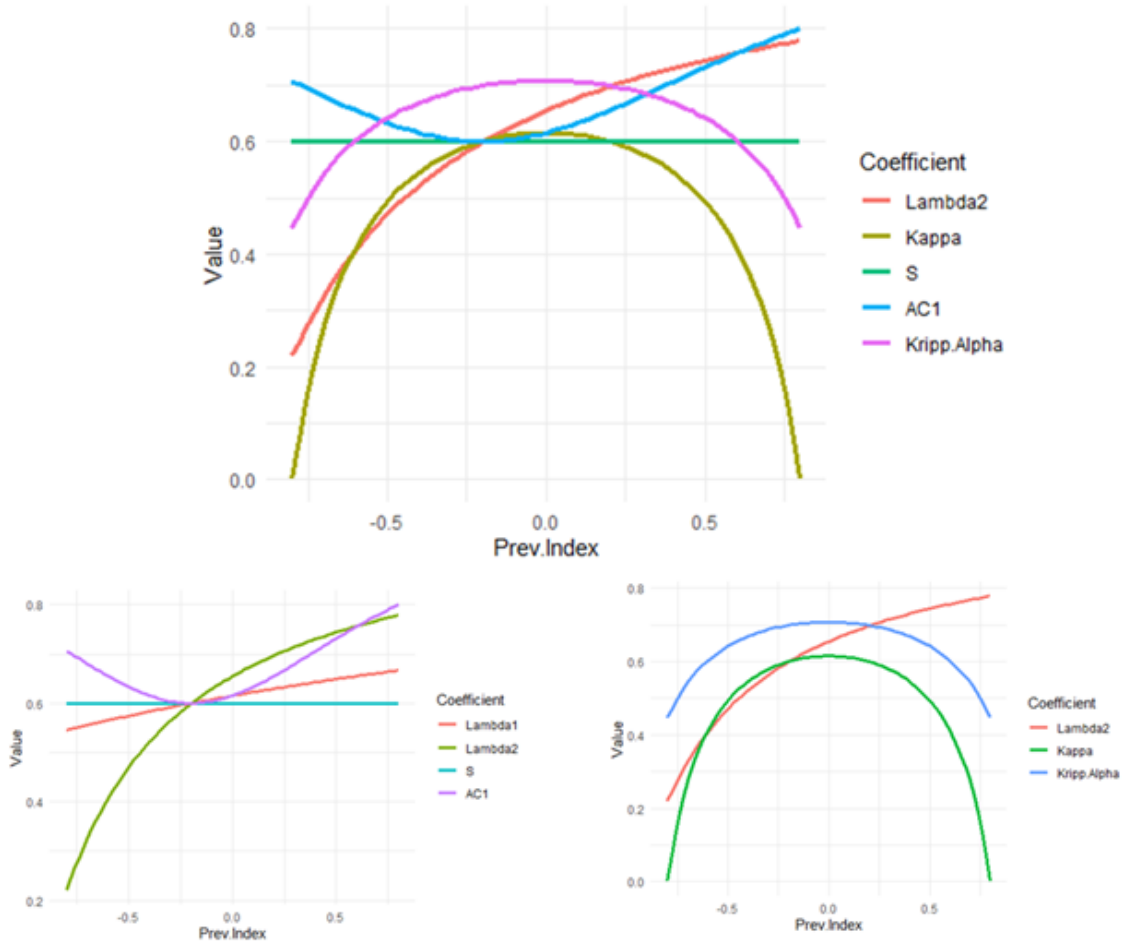


Figure J.2. Line Graphs for Data Condition 9 (2x2).

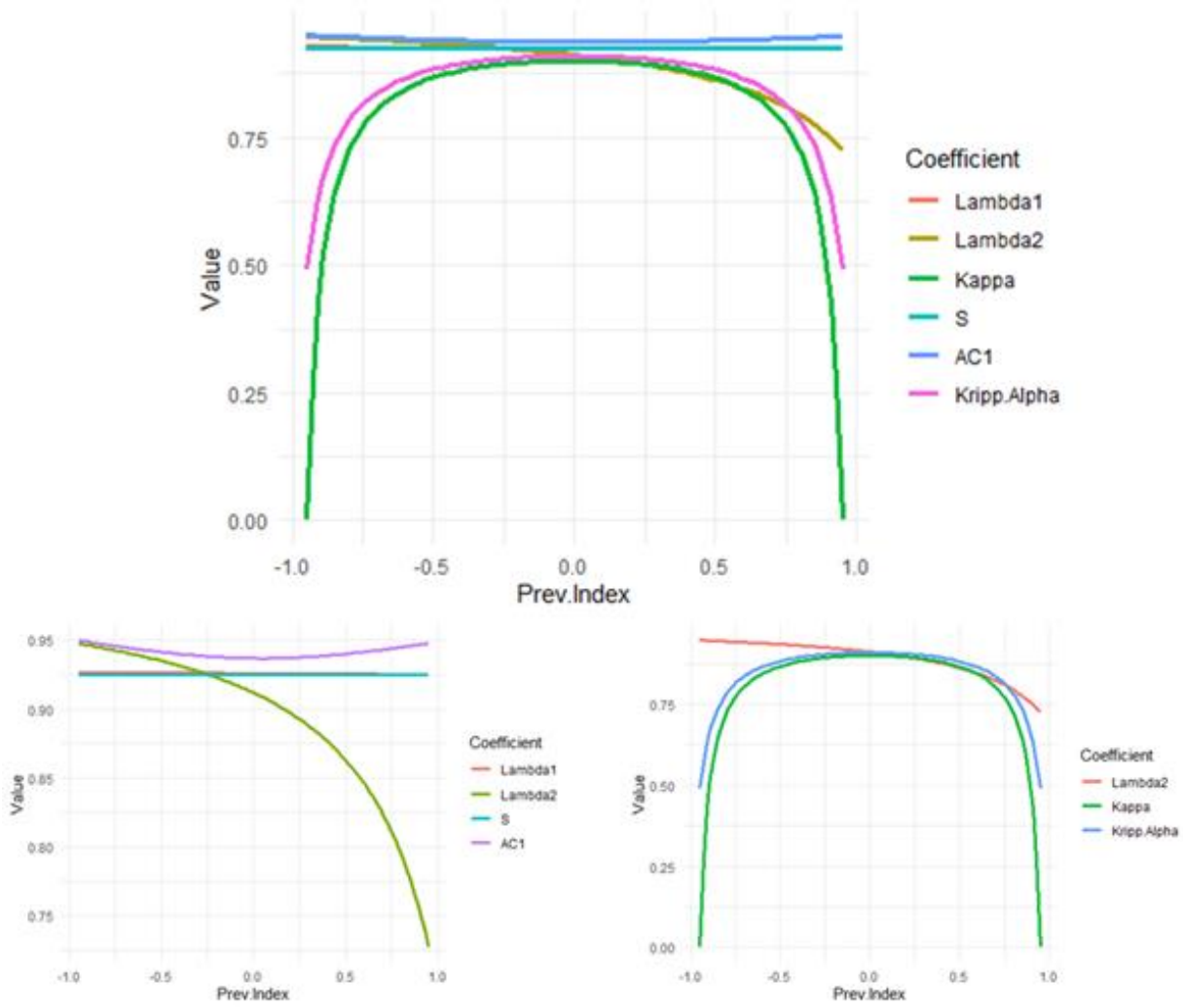


Figure J.3. Line Graphs for Data Condition 11 (3x3).

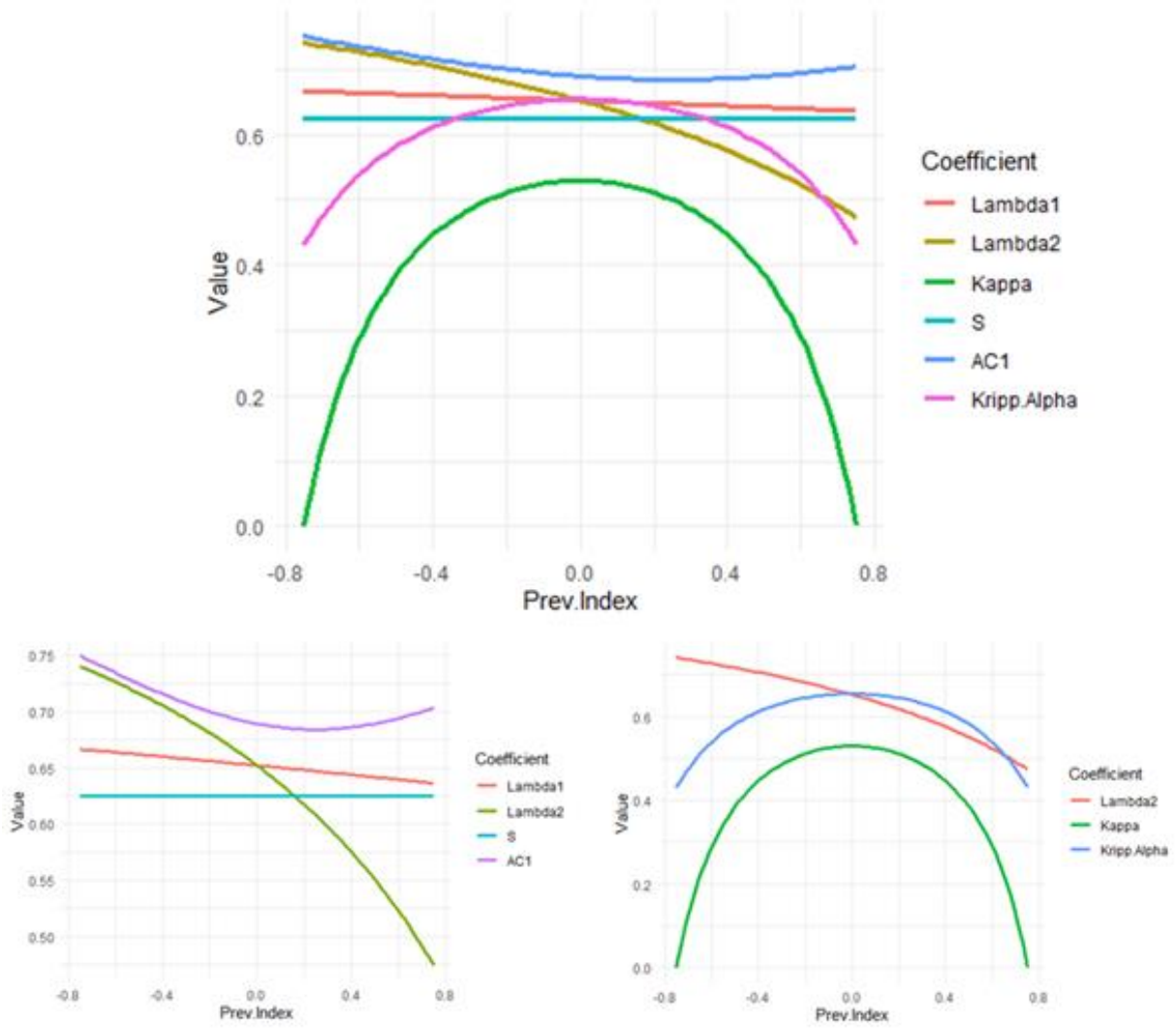


Figure J.4. Line Graphs for Data Condition 15 (3x3).

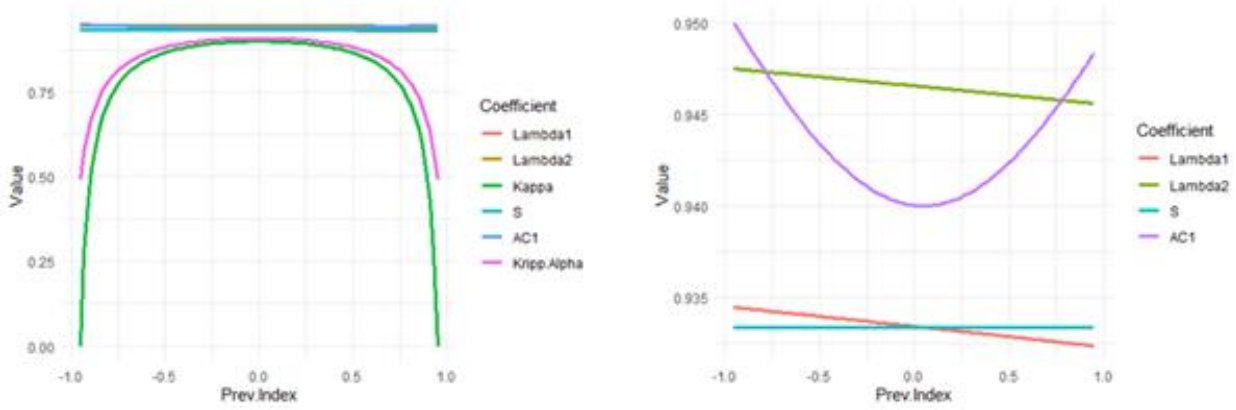


Figure J.5. Line Graphs for Data Condition 1 (4x4).

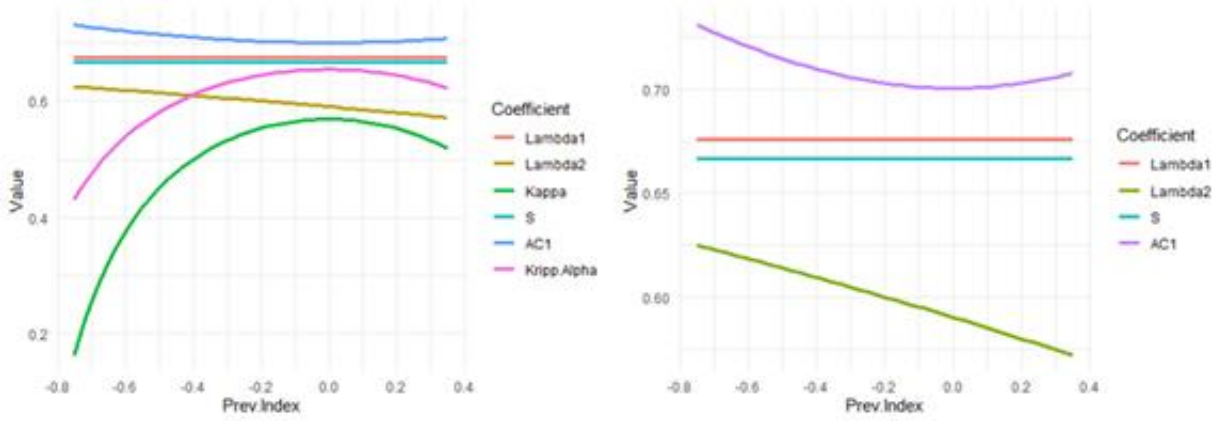


Figure J.6. Line Graphs for Data Condition 30 (4x4).