A COMPARISON OF COHEN'S KAPPA AND GWET'S AC1 WITH A MASS SHOOTING CLASSIFICATION INDEX: A STUDY OF RATER UNCERTAINTY

By

ASHLEY KEENER

Bachelor of Arts in Psychology
Northeastern State University
Broken Arrow, Oklahoma
2011

Master of Science in Educational Psychology
Oklahoma State University
Stillwater, Oklahoma
2015

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2020

A COMPARISON OF COHEN'S KAPPA AND GWET'S AC1 WITH A MASS SHOOTING CLASSIFICATION INDEX: A STUDY OF RATER UNCERTAINTY

Dissertation Approved:

Dr. Laura Barnes

Dissertation Adviser

Dr. Mwarumba Mwavita

Dr. Denna Wheeler

Dr. Jason Beaman

ACKNOWLEDGEMENTS

I would like to express the utmost appreciation to Dr. Laura Barnes who was my Committee Chair and Advisor prior to her retirement. I am forever grateful for her patience, understanding, and mentorship throughout my academic career.

Further, I would like to thank my other committee members: Dr. Mwarumba Mwavita, Dr. Denna Wheeler, and Dr. Jason Beaman for their continual support and insightful comments. My sincere thanks goes to my Department Chair, Dr. Jason Beaman, for providing me with the opportunity to grow as a researcher over the last few years.

Above all, I would like to thank my mom, Lori Keener, and dad, Dennis Keener, for always believing in me. I am beyond grateful to have such a loving and supportive family.

Name: ASHLEY KEENER

Date of Degree: MAY, 2020

Title of Study: A COMPARISON OF COHEN'S KAPPA AND GWET'S AC1 WITH A MASS SHOOTING CLASSIFICATION INDEX: A STUDY OF RATER UNCERTAINTY

Major Field: EDUCATIONAL PSYCHOLOGY

Abstract: In order to quantify the degree of agreement between raters when classifying subjects into predefined categories, inter-rater reliability (IRR) experiments are often conducted in the medical field. Originally, percent agreement was used to calculate the extent of agreement between raters; however, it was criticized for not taking into account chance-agreement. Chance-agreement refers to the propensity for raters to guess when classifying nondeterministic subjects to categories. In other words, raters can be certain that some subjects are textbook and are associated with a true category membership, whereas, other subjects are ambiguous and require true random guessing (Schuster & Smith, 2002). A commonly used chance-corrected agreement coefficient has been Cohen's Kappa. Limitations have been associated with the Kappa statistic such as Kappa's tendency to overcorrect for chance-agreement in the presence of high prevalence rates (i.e., highly skewed data). Due to such issues, Gwet (2014) proposed a new chance-corrected agreement coefficient called the AC_1 statistic. The purpose of this study was to examine Cohen's Kappa and Gwet's AC_1 with respect to prevalence rates and rater uncertainty using a newly developed classification system for mass shooters. A new methodology for identifying textbook and ambiguous subjects was demonstrated. Specifically, the purposes of the present study were (1) to examine how Cohen's Kappa and Gwet's AC_1 are affected by prevalence rates and (2) to determine whether there are differences in the observable discrepancies between Cohen's Kappa and Gwet's AC_1 for subjects classified as textbook compared to subjects classified as ambiguous. Findings indicated that observable discrepancies between Cohen's Kappa and Gwet's AC_1 could be seen in both the textbook and ambiguous conditions. Specifically, analyses suggested that percent agreement was likely to overestimate the extent of true agreement among raters and Cohen's Kappa was likely to underestimate the extent of true agreement among raters. The ambiguous analysis revealed larger discrepancies between Gwet's AC_1 and Cohen's Kappa in the presence of highly skewed data, however, discrepancies between Gwet's AC_1 and Cohen's Kappa appeared to be more dependent on the number of observable disagreements between raters during the textbook analysis. Recommendations for practice and future research are discussed.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Background

An inter-rater reliability (IRR) experiment involves asking two or more individuals, referred to as raters, to independently classify the same set of subjects into predefined categories. This process is expected to produce two or more categorizations of the same subjects. The objective is to produce high agreement between the raters, meaning, the raters can be used interchangeably without categorization being affected by a significant rater factor (Gwet, 2014). In other words, if interchangeability is guaranteed, one can have confidence that the categorization of subjects is due to the characteristics associated with the subjects as opposed to the raters. IRR studies are important to scientific investigations where the research subjects are the main focus and the data should not be affected by the raters analyzing the subjects (Gwet, 2014). For clarification, the term *subjects* may refer to people, things, or events being rated by a given set of raters in this study.

IRR studies are frequently conducted in the medical field (Gwet, 2014; McHugh, 2012). For example, in the healthcare setting, it is common for multiple people to

collect clinical laboratory data or patient information; variability among human observations and/or procedures during these processes may have severe consequences on patients. Therefore, research has focused on analyses that quantify the degree of agreement between two or more raters (e.g., healthcare providers) (McHugh, 2012). Study designs may involve training healthcare professionals to observe patients in a specific way and then measuring the extent to which they record the same scores for the same phenomenon. Hence, the objective is to observe the amount of disagreement or error that the individuals have introduced into the procedure or data collection process. If a substantial amount of disagreement can be observed, the disagreement may stem from multiple people interpreting the phenomenon differently.

Specifically, in the area of forensic psychiatry, classification systems can be utilized to analyze offender behavior. For example, the *Crime Classification Manual: A Standard System for Investigating and Classifying Violent Crimes,* Third Edition (CCM-III) allows FBI investigative profilers, law enforcement officers, and mental health practitioners to organize and classify criminal behavior based on previously defined characteristics (Douglas, Burgess, Burgess, & Ressler, 2013). Additionally, the manual helps to "standardize terminology, facilitate communication, educate, and establish a database for investigative research within the criminal justice field" (Douglas et al., 2013, p. viii). Scholars have also attempted to categorize mass murders and create classification systems based on the motivation(s) of the offender (i.e., the mass shooter) in an effort to examine characteristics associated with the identified shooter and features related to the event (Petee, Padgett, & York, 1997) Recently, much attention has been devoted to incidents of mass murder due to heightened media coverage and public interest (Meindl & Ivy, 2017; Towers, Gomez-Lievano, Khan, Mubayi,

& Castillo-Chavez, 2015); ergo, classification systems designed to analyze offender behavior may aid scholars and forensic specialists in the identification and prevention of tragic events. However, when organizing incidents according to a specific categorical system, it is essential to understand whether multiple experts are utilizing the classification system in a consistent manner. IRR studies can help to quantify this aspect of the classification process by determining whether there are consistent responses across a pool of raters to the same set of mass shooting incidents.

Statement of the Problem

Traditionally, IRR was measured as percent agreement. In order to calculate percent agreement, the number of subjects upon which raters agree in their categorization is simply divided by the total number of subjects rated (McHugh, 2012). However, percent agreement was criticized for not taking into account chance agreement – that is, the notion that raters may guess during the classification process due to uncertainty (McHugh, 2012). In response, a new IRR coefficient, called Cohen's Kappa, was developed in 1960 and was designed to address uncertainty (Cohen, 1960; McHugh, 2012). Cohen's Kappa has gained considerable popularity over the decades and is used when assessments produce categorical outcomes (Gwet, 2014; Wongpakaran, Wongpakaran, Wedding, Gwet, 2013). However, weaknesses related to Cohen's Kappa have been documented in the literature. For example, the "Kappa paradox" occurs when low Kappa values are seen despite high percent agreement (Wongpakaran et al., 2013). Some investigators have noted that the Kappa coefficient may be affected by prevalence rates (i.e., skew in the distribution of rating categories); this occurs when the "distributions of observed ratings fall under one category of ratings at a much higher rate than another category" (Hallgren, 2012, p. 6). In other words, if the marginal

totals are considerably imbalanced, Kappa estimates may be unrepresentatively lowered, thus, overcorrecting for guessing in some circumstances (Feinstein & Cicchetti, 1990; Wongpakaran et al., 2013; Viera & Garrett, 2005).

Due to these issues, Gwet (2014) proposed a new agreement coefficient called the "first-order agreement coefficient" or the AC_1 statistic (Wongpakaran et al., 2013, p. 2). The AC_1 can be utilized with any number of raters and a simple, categorical rating system. Wongpakaran et al. (2013) stated:

> The AC_1 statistic adjusts the overall probability based on the chance that raters may agree on a rating, despite the fact that one or all of the raters may have given a random value. Gwet (2014) adjusted for chance agreement by using the AC_1 tool, such that the AC_1 between two or multiple raters is defined as the conditional probability that two randomly selected raters will agree, given that no agreement will occur by chance. Gwet (2014) found that Kappa gives a slightly higher value than other coefficients when there is a high level of agreement; however, in the paradoxical situation in which Kappa is low despite a high level of agreement, Gwet proposed using AC_1 as a "paradox resistant" alternative to the unstable Kappa coefficient. (p. 2)

Rater Uncertainty

As previously mentioned, modern agreement coefficients are designed to take into account chance agreement, i.e., the probability that raters may guess when classifying subjects into different categories. Schuster and Smith (2002) stated subjects can be classified as either "obvious," "approximable," or "ambiguous" (p. 385). Obvious subjects are

associated with true category membership. Andreasen, McDonald-Scott, Keller, and Shapiro (1981) referred to obvious subjects as "textbook" or the cases that can be assigned with little or no error (p. 411). Subjects defined as ambiguous involve random guessing during the categorization process. These subjects would be associated with "true" random guessing. Subjects defined as approximable are neither obvious (i.e., textbook) nor ambiguous. For instance, raters may find that some cases can belong to one or more categories within a range of categories. That is, diagnostic procedures associated with psychiatric disorders may involve unclear or overlapping boundaries due to comorbidity, e.g., a patient can present with both anxiety and depression.

The primary difference between Cohen's Kappa and Gwet's AC_1 is in their calculation of chance agreement. The AC_1 coefficient is based on the assumption that only a portion of the ratings will lead to agreement by chance. However, Grove et al. (1981) acknowledged that if textbook subjects could be identified, one could treat those subjects separately. Further, the authors theorized that if an IRR study used predominately textbook cases, then the agreement coefficients will be higher compared to using predominately ambiguous cases (Grove et al., 1981). However, determining which subjects could be classified as textbook and ambiguous remains difficult. Therefore, one objective of this study is to demonstrate a methodology for identifying textbook and ambiguous subjects.

Purpose of the Study

In this study, a newly developed classification system based on the motivations of mass shooters was analyzed. An IRR analysis was conducted to test the consistency of the classification ratings across raters to identify the most appropriate method(s) for obtaining

IRR estimates and to evaluate whether the extent of agreement among mental health professionals is high enough to reliably classify mass shooters according to motive. The study analyzed ratings obtained from forensic psychiatrists using percent agreement, Cohen's Kappa, and Gwet's AC_1 to obtain estimates of inter-rater reliability. Data was examined for the extent to which it showed high prevalence—that is, the extent to which one or two categories predominated the ratings, suggesting one or two motivations tend to be perceived as principal in mass shootings. Further, individual cases were classified by the degree of ambiguity judged to be present based on raters' self-reported certainty. Observed discrepancies between Cohen's Kappa and Gwet's AC_1 estimates of IRR were compared between sets of ambiguous and textbook cases.

Research Questions

As previously mentioned, it is common to use classification systems to analyze offender behavior in the fields of forensic psychiatry and criminology. Reliability and validity studies are pertinent to assessing the credibility of such categorizations. The purpose of the present study was to conduct the first IRR analysis on a newly developed classification system based on the motivations of mass shooters. To my knowledge, percent agreement, Cohen's Kappa, and Gwet's AC_1 has never been tested with an IRR analysis of mass shooters. The aim of this study was two-fold: (1) to determine the extent of agreement between raters when classifying mass shooting incidents according to motives and (2) to investigate what factors may influence observed discrepancies between the agreement coefficients.

The following research questions were addressed:

1. Is there a statistically significant mean difference between percent agreement, Cohen's Kappa, Gwet's AC_1?
2. What factors account for any observed discrepancies between Cohen's Kappa and Gwet's AC_1?
    1. Specifically, are there observable discrepancies between Cohen's Kappa and Gwet's AC_1 in the presence of high prevalence rates?
    2. Are there observable discrepancies between Cohen's Kappa and Gwet's AC_1 for cases that are classified as textbook compared to cases that are classified as ambiguous?

The hypotheses of this paper are:

1. There will be a statistically significant mean difference between percent agreement, Cohen's Kappa and Gwet's AC_1. Specifically, it is hypothesized that percent agreement will be the largest and Gwet's AC_1 will be the smallest.
2. Cohen's Kappa is expected to overcorrect for chance agreement in the presence of high prevalence rates.
3. Although Gwet's AC_1 should, in general, be greater than Cohen's Kappa, it is hypothesized that the discrepancy will be greater for textbook cases compared to ambiguous cases.

Nature of the Study

This study used existing data collected from five forensic psychiatrists to examine a newly developed classification system based on the motivations of mass shooters. Category motivations for the classification system are as follows: Mental Illness, Anger, Collateral

Damage, Commission of a Crime, and Lone Actor. Table 5 describes in detail the motivational descriptions associated with each category. A collection of mass shooting incidents was obtained from the Standford Mass Shootings of America (MSA) data project ("Mass Shootings in America," n.d.) and the U.S. Department of Justice's Study of Active Shooter Incidents in the United States Between 2000 and 2013 (Blair & Schwieit, 2014) to be used during the process.

All possible pairings of the five raters resulted in ten rater pairs and the extent of agreement between the raters in each pairing were calculated using percent agreement, Cohen's Kappa, and Gwet's AC_1. Mean differences between percent agreement, Cohen's Kappa, and Gwet's AC_1 were examined. Discrepancies between Cohen's Kappa and Gwet's AC_1 were analyzed according to a prevalence index based on prevalence rates calculated for each rater pair. Prevalence rates were calculated as a percentage based on the number of agreed upon mass shooting incidents per category (i.e., as judged by both raters) then divided by the total number of mass shooting incidents. Finally, textbook and ambiguous cases were determined by examining response strings provided by each rater per case and discrepancies between Cohen's Kappa and Gwet's AC_1 were studied among textbook cases and ambiguous cases.

Significance of the Study

A novel contribution this study makes to the field is by examining the data using percent agreement, Cohen's Kappa, and Gwet's AC_1 to compare their levels of reliability in relation to mass shooting indexes. The agreement coefficients will be reviewed and suggestions regarding their use in forensic psychiatry research will be discussed. Further,

this study adds to the literature by developing a methodology for identifying textbook and ambiguous subjects. As noted by Grove et al (1981), determining which subjects in an IRR analysis are textbook and which subjects are ambiguous is a difficult task and has never been done successfully.

CHAPTER II

REVIEW OF THE LITERATURE

It is common in many psychiatric and medical research studies to conduct IRR experiments because they are fundamental to the design and evaluation of diagnostic instruments. The objective is to evaluate the extent of agreement among two or more raters to ensure the raters can be used interchangeably. However, there are many existing agreement coefficients which can lead to confusion regarding their appropriate use (Gisev, Bell, & Chen, 2013). For instance, IRR ratings can be identified as nominal, ordinal, interval, or ratio. Categories are considered nominal when no meaningful ordering of items or categories are present while subjects classified as "Certain," "Probable," "Possible," or "Doubtful" (for e.g.,) are said to be rated on an ordinal scale. Weighted versions, such as Weighted Kappa and Gwet's AC_2, have been developed for use with ordinal, interval, and ratio data. Cohen's Kappa and Gwet's AC_1 are used with nominal ratings and will be the focus of this study.

Biostatisticians have found Gwet's AC_1 to be superior to Cohen's Kappa under certain conditions (Chan, 2003), however, few researchers have adopted Gwet's AC_1 as a statistical tool in the medical field (Wongpakaran et al., 2013). Past reviews (Gisev et al.,

2013) of IRR methods have discussed Cohen's Kappa but have failed to mention Gwet's AC_1 (Wongpakaran et al., 2013). Making researchers and practitioners more aware of the limitations and benefits to certain agreement coefficients could prove helpful. There are a number of chance-corrected agreement coefficients that can be used when analyzing nominal ratings. In order to narrow the focus of the present study, Cohen's Kappa and Gwet's AC_1 will be reviewed. Additionally, these chance-corrected agreement coefficients were selected because Cohen's Kappa is frequently used in lieu of Gwet's AC_1 despite its limitations (Wongpakaran et al., 2013). The following sections will discuss Cohen's Kappa and Gwet's AC_1 in more detail.

Cohen's Kappa

Percent agreement is often seen as the most intuitive approach to an agreement coefficient. However, one of its primary flaws is that is does not take into account chance agreement among raters. For example, suppose two raters are asked to assign subjects to a two-category IRR experiment (e.g., a patient is diagnosed with clinical depression or a patient is not diagnosed with clinical depression). If one or both raters guess about a subject's category selection due to uncertainty, it's still probable they may agree given the limited number of categories. Agreement by chance may indicate the raters have not mastered the rating process and percent agreement may overestimate the "true" extent of rater agreement. This problem has led to chance-corrected agreement coefficients, such as Cohen's Kappa (Gwet, 2014).

In 1960, Jacob Cohen proposed a coefficient that would determine the level of agreement between raters in nominal scales, provide a basis for testing hypotheses, and

set confidence intervals for the coefficient (Cohen, 1960). Cohen (1960) suggested two relevant quantities for nominal scale agreement between two judges. These quantities are:

p_o = The proportion of units to which judges agree

p_e = the proportion of units to which agreement is expected by chance

The denominator of Kappa (κ) is expressed as $1 - p_e$ and represents the "test of agreement for which the hypothesis of no association would predict disagreement between the judges" (Cohen, 1960, p. 39). The numerator of κ would suggest that nonchance factors are operating in the direction of agreement and is expressed as $p_o - p_e$, respectively. Therefore, κ represents the proportion of agreement after chance agreement is corrected and is defined as:

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

If κ is equal to 0, then the obtained agreement would equal chance agreement. A $\kappa > 0$ (i.e., positive values of κ) indicates greater than chance agreement. Negative values of κ would indicate less than chance agreement. The upper limit of κ is + 1.00 and suggests perfect agreement between judges. The standard error of κ is regarded as an approximation and is given by:

$$\sigma_\kappa = \sqrt{\frac{p_o\,(1 - p_o)}{N\,(1 - p_e)^2}}$$

The formula is regarded as an "approximation" because it treats p_e as a constant and p_o as the population value (Cohen, 1960, p. 43). However, Cohen (1960) stated that it should be adequate, particularity with a large N ($N \geq 100$), because p_e will not vary greatly

relative to κ. Additionally, the sampling distribution of κ will approximate reality with a large *N* and confidence intervals can be established:

$$95\% \text{ confidence interval} = \kappa \pm 1.96\ \sigma_{\kappa}$$

$$99\% \text{ confidence interval} = \kappa \pm 2.58\ \sigma_{\kappa}$$

Limitations of Cohen's Kappa

In the literature, Cohen's Kappa has been praised for addressing chance agreement; however, it has also been known to suffer from certain limitations (e.g., Byrt, Bishop, & Carlin, 1993; Feinstein & Cicchetti, 1990; Gwet, 2002, 2008, 2014; Zec, Soriani, Comoretto, & Baldi, 2017). For example, the Kappa paradox occurs when low Kappa values are seen in the presence of high agreement. Further, the Kappa statistic is affected by high prevalence rates (i.e., trait prevalence) or a substantial discrepancy in classification probabilities (i.e., marginal homogeneity) (Zec et al., 2017). Prevalence occurs when subjects are assigned more often to one of the possible outcomes. This may occur under two conditions: (1) the nature of the outcome itself may involve high prevalence or (2) one or more raters assign subjects to one specific outcome more often (Zec et al., 2017).

In relation to marginal homogeneity and trait prevalence, Gwet (2002) demonstrated that Cohen's Kappa can be "unstable and difficult to interpret" (p. 2). This observation was shown when conducting four 2X2 IRR hypothetical experiments where two raters (i.e., rater A and rater B) were asked to classify subjects into two categories (i.e., the subjects carried a specific trait or did not carry a specific trait). As shown in Table 1, both rater A and rater B tended to classify subjects into the positive category for experiment 1. However, in experiment 2, rater B was more likely to classify a subject into

the positive category and rater A was more likely to classify a subject into the negative category (Table 2). The percent agreement for both experiments was .60; however, the differences in marginal probabilities yielded a Kappa statistic of .13 for experiment 1 and .26 for experiment 2 (Gwet, 2002). The results were considered contrary to expectation since most researchers would expect a higher agreement coefficient for experiment 1 than experiment 2.

Table 1

Hypothetical experiment 1 from Gwet (2002) demonstrating the distribution of 100 subjects by rater and category response

Rater B	Rater A		Total
	+	-	
+	45	15	60
-	25	15	40
Total	70	30	100

Table 2

Hypothetical experiment 2 from Gwet (2002) demonstrating the distribution of 100 subjects by rater and category response

Rater B	Rater A		Total
	+	-	
+	25	35	60
-	5	35	40
Total	30	70	100

During experiment 3 (Table 3) both raters classified subjects into the positive and negative categories 50% of the time; in experiment 4 (Table 4), both rater A and rater B classified subjects in to the positive category 80% of the time and subjects into the

negative category 20% of time. The percent agreement for both experiments was 80%. Although raters demonstrated the same marginal probabilities, experiment 3 yielded a Kappa statistic of .60 and experiment 4 yielded a Kappa statistic of .38. Here, Gwet (2002) demonstrated that Kappa was more affected by the propensity to classify subjects into a positive category as opposed to differences in marginal probabilities. Additionally, in a study conducted by Zec et al. (2017), the authors found that the paradox starts to occur for prevalence rates higher than 60%.

Table 3

Hypothetical experiment 3 from Gwet (2002) demonstrating the distribution of 100 subjects by rater and category response

Rater B	Rater A		Total
	+	-	
+	40	10	50
-	10	40	50
Total	50	50	100

Table 4

Hypothetical experiment 4 from Gwet (2002) demonstrating the distribution of 100 subjects by rater and category response

Rater B	Rater A		Total
	+	-	
+	70	10	80
-	10	10	20
Total	80	20	100

Gwet's AC_1

Gwet's AC_1 statistic has been shown to be a paradox-resistant alternative to Cohen's Kappa (e.g., Gwet, 2014; Zec et al., 2017). According to Gwet (2014), Kappa does not adequately evaluate percent chance agreement. In an IRR analysis based on a *q*-level nominal measurement scale, Gwet's (2014) formula is defined as follows:

$$\kappa_G = \frac{p_a - p_e}{1 - p_e}, \text{with } p_a = \sum_{k=1}^{q} p_{kk}, p_e = \frac{1}{q-1} \sum_{k=1}^{q} \pi_k (1 - \pi_k)$$

where $\pi_k = (p_{k+} + p_{+k})/2$ and the symbol p_{kk} represents "the relative number of subjects classified into category k by both raters (Gwet, 2014, p. 105). Further, π_k "represents the probability for a randomly selected rater to classify a randomly selected subject into category *k* (Gwet, 2014, p. 105). Gwet (2014) stated p_e is the product of the two following quantities:

> The probability that two raters agree given that the subject being rated is nontextbook and was therefore assigned a nondeterministic rating. This conditional probability is $1/q$ since nondeterministic rating are considered random with equal chance for all *q* categories. And the propensity for a rater to assign a nondeterministic rating, which is estimated by the ratio:
>
> $$\sum_{k=1}^{q} \pi_k (1 - \pi_k)/(1 - \frac{1}{q}).$$
>
> What is important to retain from this expression is that a distribution of subjects that is skewed towards a few categories will lower the nondeterministic rating propensity. (p. 105)

In relation to the calculation of percent chance agreement associated with Cohen's Kappa, ratings are assumed to be independent prior to the experiment being carried out (Gwet, 2014). This is seen as an improbable assumption (Gwet, 2014). Specifically, Gwet (2014) argued:

> To justify the two expressions used to evaluate the chance-agreement probabilities of Kappa, the reasoning was that if the processes by which two raters classify a subject are statistically independent, then the probability that they agree is the product of the individual probabilities of classification into the category of agreement. However, raters often rate the same subjects, and are therefore expected to produce ratings that are dependent with possibly a few exceptions when they are in doubt. (p. 103)

The AC_1 statistic operates under the assumption that only a portion of ratings will lead to chance agreement and that independence occurs in the presence of a nondeterministic rating (i.e., the process of rating a subject has no apparent connection with the subject's characteristics) (Gwet, 2014). Conceptually, the AC_1 statistic represents a "trimmed population of subjects where agreement by chance would be impossible" because "all subjects classified into identical categories by pure chance" are first "removed from the population of subjects" (Gwet, 2014, p.104).

The limitations associated with Cohen's Kappa, such as marginal homogeneity and trait prevalence, have been discussed. In comparison to Kappa, the AC_1 statistic has been shown to be a more stable chance-corrected agreement coefficient (Gwet 2002, 2008, 2014; Zec et al., 2017). In a series of analyses, Gwet (2002) demonstrated that the

AC_1 coefficient had more robust properties than the Kappa statistic in the presence of trait prevalence. For example, an acceptable agreement coefficient should include the following properties: (1) if sensitivity (i.e., the propensity of a rater to detect positive cases) and specificity (i.e., the propensity of a rater to detect negative cases) are all equal and high among raters, then IRR should be high even in the presence of high or low trait prevalence, (2) if sensitivity is smaller than specificity among raters, then IRR should be higher in the presence of lower trait prevalence, and (3) if specificity is smaller than sensitivity among raters, then IRR should be higher in the presence of higher trait prevalence (Gwet, 2002). These properties indicate that a combination of high sensitivity and high prevalence would lead to higher IRR. However, the Kappa statistic does not demonstrate such properties. Specifically, with a prevalence of 100% and a constant value of .90 set to each rater's sensitivity and specificity, the Kappa statistic produced an IRR estimate of 0, whereas, the AC_1 statistic produced an IRR estimate of .78 (Gwet, 2002).

Rater Uncertainty

Gwet (2014) recognized that "the notion of chance agreement is pivotal in the study of chance-corrected agreement coefficients;" however, the definition of what constitutes chance agreement can be considered controversial (p. 102). In relation to medical diagnostics, chance agreement would imply that practitioners assign diagnoses to subjects at random. Grove et al. (1981) acknowledged that healthcare professionals, or experts, do not function this way. The authors stated:

> The flaw in this scheme is that it assumes a rating process model that does not depend on the rater's observed behavior (i.e., diagnostic base rates). It suggests that when in doubt the rater mentally flips a coin to make the diagnosis. We hope nobody really does this. (p. 411)

Grove et al. (1981) argued that "Kappa on the other hand can be visualized as embodying the following model of chance agreement: when in doubt on a nontextbook case, each rater mentally flips a biased coin, with the probability of getting "heads" (giving the diagnosis) equal to his own base rate" (p. 411). However, Gwet (2014) argued that this definition of Kappa is too generous because Kappa does not incorporate an estimate of nontextbook (i.e., uncertain) cases.

In their general formulation, both Cohen's Kappa and Gwet's AC_1, can be illustrated as a single quantity:

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

As previously mentioned p_a represents observed agreement and p_e represents the probability that the raters agreed by chance. The primary difference between Cohen's Kappa and Gwet's AC_1 is how chance agreement is calculated. Cohen's Kappa relies on the obtained distributions of two raters in order to correct for chance agreement. In other words, the chance corrected calculation for Cohen's Kappa is dependent on marginal frequencies (i.e., the row and column totals in a given contingency table) (Xu & Lorber, 2014). Due to this dependency, the Kappa statistic is sensitive to base rates and varying levels of skew in the data. This is a common phenomenon in the behavioral sciences

because clinicians often encounter more prevalent disorders (e.g., depression) opposed to less prevalent disorders (e.g., schizophrenia). Skew in one's data can lead to unbalanced marginals and increase the estimate of chance agreement in the Kappa statistic (Xu & Lorber, 2014). Gwet's AC_1, on the hand, has been shown to be less sensitive to base rates (Gwet, 2002, 2008). Chance agreement of the AC_1 statistic is defined as "chance agreement only under the circumstance that two raters agree; however, at least one of them has performed a random rating" (Xu & Lorber, 2014, p. 1220).

Rater bias. There are apparent computational and theoretical differences in how Cohen's Kappa and Gwet's AC_1 correct for chance agreement. However, it is important to note that additional factors may influence base rates. As previously mentioned by Zec et al. (2017), prevalence, i.e., skew in the data, occurs when subjects are assigned more often to one of the many possible outcomes and can occur under two conditions: (1) the nature of the outcome itself may involve high prevalence or (2) one or more raters assign subjects to one specific outcome more often. Additionally, a "gold standard" approach to IRR analysis occurs when raters are unbiased in their assessment of subjects; in other words, raters should essentially be interchangeable (Lorber, 2006). In behavior observation and clinical diagnosis, human judgement can be seen as both a strength and weakness. Xu and Lorber (2014) stated:

> People are capable of integrating a complex set of cues to arrive at psychologically informed judgements. At the same time, these judgements are imperfect; they are influenced by the characteristics of the raters themselves (e.g., experience, conscientiousness) as well as random error. Careful training can reduce but not eliminate differences among raters. (p. 1219)

The impact of rater bias may serve as an additional source of skew in observational data and IRR analysis (Xu & Lorber, 2017). Rater bias, or disagreements among raters, may occur under two conditions: (1) the rater's interpretation of the rating scale may differ, and (2) the rater's perceptions of individual subjects may differ (Hoyt, 2000). For example, a rater may specialize in a specific psychiatric disorder, such as depression, and be more prone to recognize behavioral characteristics and symptoms associated with that disorder. Xu and Lorber (2017) acknowledged that researchers cannot be sure if the statistics or the raters are to blame with both skewed data and low IRR coefficients in the presence of high observed agreement. Additionally, Cronbach (1955) stated that in order to achieve effective ratings, researchers must understand the biases and assumptions through which raters filter information. In the psychometric literature, measurement error can be random or systematic. Both random and systematic error have nothing to do with the construct of interest. The primary difference between the two types of error is that systematic error affects measurement in a consistent or repeatable manner (Raykov & Marcoulides 2011). Random error, on the other hand, is transient and due to pure chance. Hoyt (2000) referred to rater bias as method variance that "contributes to systematic variance in observed scores that is not due to the target" (p. 65). Additionally, the author acknowledged that method variance and rater bias contribute to measurement error because, often times, it is of no substantive interest to investigations (Hoyt, 2000). Hoyt and Kernis (1999) found that two features of a rating system (i.e., attribute type and rater training) contributed to the level of bias in ratings among raters. Attribute type was defined as the degree of inference raters must use to assign ratings. Attribute type was further defined as explicit, inferential, or mixed. Explicit links were seen as readily

observable behaviors, such as the frequency of head nods, and accounted for fewer disagreements among raters. Ratings of global traits, such as personality types or job performance, were classified as inferential and accounted for more disagreements between raters due to the complexity of judgements being made about the subjects. Rating systems that combined both explicit and inferential features were defined as mixed. Rater training refers to the amount of training that raters receive with the rating scale prior to providing data for a particular study (Hoyt & Kernis, 1999). Specifically, the authors defined rater training as the number of hours spent learning the rating system along with an expert or the number of hours spent utilizing the rating system in pilot studies or nonresearch contexts. In sum, the authors found that raters that received little to no training (≤ 5 hours) and provided inferential ratings contributed the largest proportion of bias variance. Kimberlin and Winterstein (2008) acknowledged that IRR is optimized when variables of interest involve precise operational definitions and raters are appropriately trained. Specifically, "the more that individual judgement is involved in a rating, the more crucial it is that independent observers agree when applying the scoring criteria" (Kimberlin & Winterstein, 2008, p. 2278). Hill, O'Grady, and Price (1988) examined ratings as a function of rater characteristics in the field of counseling. The authors found little evidence for rater bias in the study and contributed the findings to the psychometrically sound instruments that were used. For example, the authors noted that "raters told them that most items on the scales were relatively easy to rate because they were highly operationalized, concrete, and specific" (Hill et al., 1988, p. 349). Furthermore, interviews with the raters were conducted for heuristic purpose due to the complex nature of rater bias. Hill et al. (1988) found that the following qualitative factors

may have contributed to rater bias as reported by the raters themselves: (1) fatigue, (2) waning effort and sensitivity to the subjects, (3) changes in the rating processes, and (5) length of the measures. Additionally, Hill et al. (1988) acknowledged that awareness of the possibility of bias may have balanced potential bias. In other words, educating raters about rater bias might reduce rater bias during the rating process. In relation to Cohen's Kappa and Gwet's AC_1, Xu and Lorber (2014) conducted a Monte Carlo evaluation of various IRR coefficients under a combination of conditions commonly encountered in clinical research; these conditions included observed agreement, rater bias, base rate, and sample size. The authors found that the AC_1 statistic was not affected by rater bias and that Cohen's Kappa was slightly sensitive to rater bias under various simulation conditions (Xu and Lorber, 2014). However, the authors did not evaluate specific rater characteristics that may contribute to rater bias. In sum, depending on the type of rating scale that is being utilized (i.e., nominal, ordinal, interval, or ratio) certain factors that may contribute to low IRR coefficients or rater bias may include restricted range, scales that contain poor psychometric properties, poorly trained raters, and trouble with observing the construct of interest (Hallgren, 2012).

Conclusion

The literature has identified several factors that may influence the rater to perceive subjects as either textbook, approximable, or ambiguous. Characteristics associated with the subjects themselves, such as the occurrence of more prevalent disorders (e.g., anxiety) opposed to less prevalent disorders (e.g., schizophrenia), can lead to skewed data and unbalanced marginals. As previously mentioned, this can increase the estimate of chance agreement in the Kappa statistic (Xu & Lorber, 2014). Further,

characteristics associated with the raters, such as their interpretation of the rating scale and/or their perceptions of the individual subjects, may influence the rating process and contribute to skewed data (Hoyt, 2000). Both Cohen's Kappa and Gwet's AC_1 attempt to computationally correct for chance agreement. Specifically, Cohen's Kappa is dependent on the obtained distributions of the raters to correct for chance agreement (Xu & Lorber, 2014). In contrast, chance agreement of the AC_1 statistic is defined as "chance agreement only under the circumstance that two raters agree; however, at least one of them has performed a random rating" (Xu & Lorber, 2014, p. 1220). Gwet (2002) stated that if sensitivity and specificity are all equal and high among raters then IRR should be high even in the presence of high or low prevalence. Further, Grove et al. (1981) noted that if a study uses primarily textbook cases one would, of' course, expect IRR estimates to be higher than if a study primarily used ambiguous cases. The present study aims to examine how Cohen's Kappa and Gwet's AC_1 function among both textbook and ambiguous cases using a real dataset.

CHAPTER III

METHODOLOGY

This study was a secondary analysis, IRR experiment using data collected from five raters to compare the magnitude of agreement coefficients (i.e., Cohen's Kappa and Gwet's AC_1) in the presence of varying levels of prevalence and rater uncertainty. A mass shooting classification system was created to examine motives among offenders and to assist forensic specialists in the identification and prevention of tragic events. Calculation of IRR indices was conducted with AgreeStat 2015.6.1 and SPSS. AgreeStat is a statistical program developed by Kilem L. Gwet and is embedded in a stand-alone Excel Workbook. The program is used to perform statistical analysis on the extent of agreement among multiple raters. SPSS was used for analyses of the IRR indices. The Oklahoma State University Institutional Review Board (IRB) determined that the present study did not qualify as human subjects research. The IRB form can be found in Appendix A.

Raters

A total of five raters used for this IRR experiment were forensic psychiatrists located throughout various regions of the United States. A fully crossed design was

utilized where all raters were asked to rate each mass shooting incident according to five categories based on the motivation of the offender. The raters were comprised of four males and one female. Four out of five raters identified as Caucasian, whereas, one rater identified as African American. Age ranged from 35 to 47 (M = 42.75, SD = 5.32). All raters held medical degrees (i.e., an M.D.) and had obtained their degree, on average, 13.75 years ago (SD = 5.06, range = 8 to 18). Three out of the five raters defined their area of specialization as Forensic Psychiatry, one defined it as Forensic Psychiatry and the severely mentally ill, and one defined it as Psychiatry. Further, four out of five raters had experience with mass shooters such as: (1) post-conviction psychiatric treatment, (2) conducting forensic evaluations, and (3) working on cases that involved more than one shooting (i.e., school shootings and mass murder). A complete list of questions can be found in Appendix B.

Classification System

The mass shooting classification system has been termed the Agoracide classification system and consists of five categories and motive descriptions. The categories are labeled: Mental Illness, Collateral Damage, Anger, Commission of a Crime, and Lone Actor. The descriptions associated with each category can be seen in Table 5. In order to provide content validity, it should be noted that the classification system was developed by a Forensic Psychiatrist.

Methods

Dataset. I was a part of a six-member research team that developed the dataset consisting of mass shooting incidents. Data for this study were derived from the

Table 5

Category description for the Agoracide mass shooting classification system

Category	Description
Mental Illness	Motive is one in which an individual is acting because of psychosis (delusions, hallucinations, disorganized thoughts or behavior). Such individuals could be suffering from schizophrenia, bipolar disorder, severe depression, among other reasons. Substance induced psychosis would also be included in this motive. The hallmark of this motive is that ***but for*** the mental illness, the shooting would not have occurred.
Collateral Damage	Motive is one in which the individual intended to harm a specific individual. However, after accomplishing this goal, the individual continues to harm others, including individuals that he or she has never met and may not be related to the subject of their anger at all. For example, an individual who kills his wife at work but then kills other co-workers, customers or other strangers. A Collateral Damage Motive will usually stem from another type of mass shooting outside of stranger mass shooting, such as domestic or workplace violence.
Anger	Motive is one in which the individual is angry at a specific entity or the world at large. This anger is usually the result of narcissism or other personality traits, not delusional thinking. This anger is not limited to the other types of mass shootings, such as domestic or workplace violence. Often referred to as an injustice collector, this individual often feels as though he or she has been unjustly targeted or persecuted, is hypersensitive and easily offended. This motive would also include revenge.
Commission of a Crime	Motive is one in which the individual's main motive is some other illegal act besides murder. This usually includes theft. For example, if an individual commits a robbery and then murders multiple witnesses.
Lone Actor	Motive is one in which the shooting is performed in order to achieve a desired political objective. These individuals usually have contact or loose affiliations with an organized terror or hate group but have not been subject to formal training. In contrast to a terror attack, lone wolf terrorists usually are radicalized in their place of origin and commit their act in the same or similar place.

following databases: The Standford Mass Shootings of America (MSA) data project ("Mass Shootings in America," n.d.) and the U.S. Department of Justice's Study of Active Shooter Incidents in the United States Between 2000 and 2013 (Blair & Schwieit, 2014). Taken together, the two databases contained mass shootings that occurred between the years 1966-2016. The databases were chosen because they contained uniform cases and were publically available. Originally, a total of 308 incidents were collected. In order to construct a valid list of mass shooting incidents based on pre-established inclusion criteria, the 308 cases were assigned to four research team members to determine if each case meet inclusion criteria. Mass shooting incidents were included for the following reasons: (1) the case involved one shooter, (2) the shooting was classified as non-school/work or the student was not an employee or student where the shooting occurred, (3) a type of firearm was used, (4) the shooting was classified as non-gang affiliated, (5) the shooting(s) occurred within a 24-hour period, and (6) multiple rounds of ammunition were fired. A flowchart depicting this process can be seen in Figure 1. The four research team members independently screened the descriptions of each mass shooting incident and applied the eligibility criteria. Disagreements were resolved by analyzing the description of the incident in question during regularly scheduled meetings until consensus was reached. Thus, a total of 219 mass shooting incidents were retained for expert review.

Once the dataset was finalized, the 219 incidents were divided among three groups of six research team members. In other words, each group (i.e., containing two members) was responsible for conducting further research on 73 mass shooting incidents. Each team member was responsible for obtaining relevant newspaper or research articles for each of

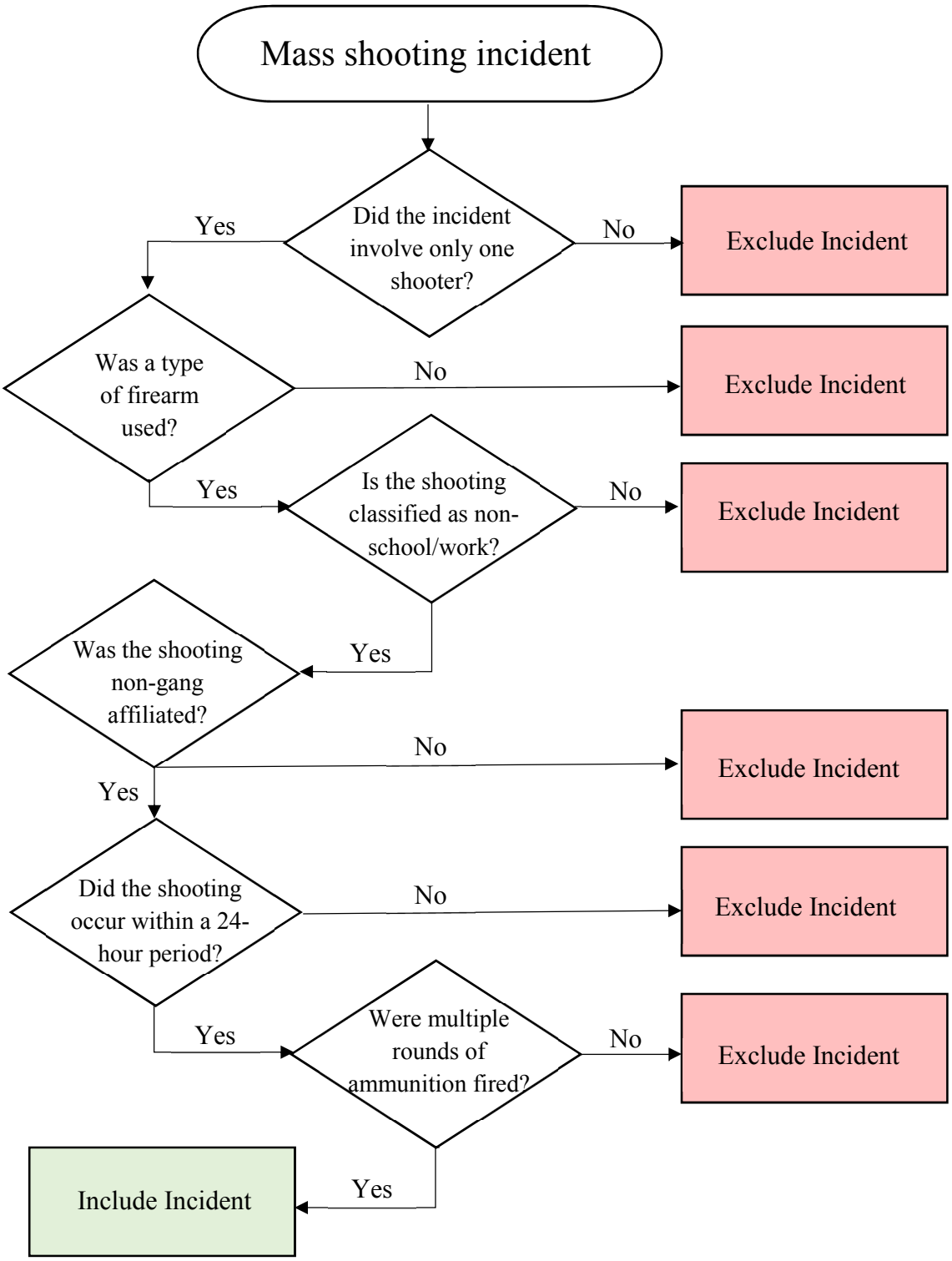

Figure 1. Flow chart depicting the inclusion/exclusion decision making process per mass shooting incident.

their assigned incidents. In order to accomplish this task, a systematic search procedure was executed by all research team members. Specifically, three searches were executed: (1) the first search involved googling the shooter's name followed by the phase "shooting," (2) the second search involved googling the shooter's name followed by the phase "psych," and (3) the third search involved googling the shooter's name followed by the phase "mental illness." The research groups were instructed to open every link on the first two pages per search, while saving the articles and their links, and acquiring relevant information for each mass shooting incident. All searches were performed in Google Incognito in order to achieve consistent search results that were not based on internet history. After relevant newspaper or research articles were collected per mass shooting incident, the 219 cases were divided among 10 volunteer medical students. The medical students were responsible for reading the relevant articles that were associated with a single incident and summarizing the information in a one to three-page document called a "case synopsis." An example case synopsis can be found in Appendix C.

Procedure. Once each case synopsis was finalized per mass shooting incident, I gave the five reviewers access to the 219 mass shooting incidents, their relevant articles, and each case synopsis via a separate link on Dropbox. Additionally, they were provided with a Reviewer Excel Sheet that allowed them to rate 'how much' the case falls into each category using a Likert-type scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Each reviewer was also asked to indicate which Agoracide category is best associated with each case. Complete reviewer instructions can be seen in Appendix D.

Data analyses

This IRR experiment involved five raters, and five possible motive categories into which mass shooting incidents could be classified. The rating scale was considered nominal because the five categories could not be ranked by order, importance, severity, or any other attribute. A total of 10 rater pairs with the five raters were created and analyzed. Table 6 displays the 10 pairs of raters that participated in this reliability experiment including the total number of subjects that each pair rated after excluding missing data. The number of subjects per pair differ because some raters did not complete all ratings. Specifically, the following research questions were addressed:

1. Is there a statistically significant mean difference between percent agreement, Cohen's Kappa, and Gwet's AC_1?
2. Are there observable discrepancies between Cohen's Kappa and Gwet's AC_1 in the presence of high prevalence rates?
3. Are there observable discrepancies between Cohen's Kappa and Gwet's AC_1 for cases that are classified as textbook compared to cases that are classified as ambiguous?

In order to address the first research question, agreement coefficients associated with percent agreement, Cohen's Kappa and Gwet's AC_1 were obtained for each rater pair. A one-factor repeated measures ANOVA was conducted to determine if there was a statistically significant difference between the three agreement coefficients. For completeness, the variance of the coefficients across the 10 rater pairs was also obtained and compared between Cohen's Kappa and Gwet's AC_1 to examine the assumption of

interchangeability of raters. Smaller variance was an indication of greater interchangeability.

Table 6

Rater pairs and the number of subjects rated per pair

Pair Number	1	2	3	4	5	6	7	8	9	10
Rater Names	A	A	A	A	B	B	B	C	C	D
	B	C	D	E	C	D	E	D	E	E
No. of Subjects	199	199	189	199	219	204	219	204	219	204

In order to address the second research question, prevalence rates were obtained for each rater pair. Prevalence refers to skewness in the data. In other words, are the majority of the mass shooting incidents classified to one category more frequently compared to other categories? Prevalence rates were first calculated as a percentage based on the number of agreed upon mass shooting incidents per category (i.e., as judged by both raters) then divided by the total number of mass shooting incidents. For example, if raters A and E agree that 54 out of 219 mass shooting incidents can be classified as motivated by "Anger," then the "Anger" category would receive a category agreement rate of 25% (Wongpakaran et al., 2013). Thus, these two raters assigned 25% of the cases to the Anger category. Variance in these category agreement rates among the categories was then calculated for each rater pair. Larger variances indicated a greater tendency for cases to be assigned to a single category resulting in more skewed data, and thus greater prevalence for that rater pair, whereas smaller variances suggested a more even distribution of cases across the categories. This variance is the Prevalence Index for each rater pair. The discrepancy

between Gwet's AC_1 and Cohen's Kappa (Gwet's AC_1 minus Cohen's Kappa) was also calculated for each pair of raters. For instance, if raters A and B obtained a chance-corrected agreement coefficient of .84 for Gwet's AC_1 and a coefficient of .72 for Cohen's Kappa the discrepancy would be: .84 - .72 = .12. Finally, a Pearson product-moment correlation coefficient was conducted to assess the relationship between the prevalence index and the discrepancy between Gwet's AC_1 and Cohen's Kappa. A positive relationship would indicate that as prevalence increases, the discrepancy between Gwet's AC_1 and Cohen's Kappa also increases. This correlation between the prevalence index and the discrepancy would address the research question regarding how prevalence affects the discrepancy between Cohen's Kappa and Gwet's AC_1.

In order to address the third research question and determine how Cohen's Kappa and Gwet's AC_1 function with respect to rater uncertainty, textbook and ambiguous cases were analyzed. Textbook cases involve "obvious' subjects (i.e., mass shooting incidents) that are associated with a 'true' category membership, whereas, ambiguous cases involve subjects that require 'random' guessing concerning category membership. Specifically, using the Likert-type scale assigned to each category per case, a case was defined as a textbook case for a rater if it received a score of 4 or 5 (*agree* or *strongly agree* the case belongs in this category) to only one category by the rater and a score of 1 or 2 to all other categories (*strongly disagree* or *disagree*). A case was defined as ambiguous for a rater if it received a score of 4 or 5 assigned to two or more categories or a score of 3 or less to all categories by that rater. Table 7 displays examples of cases that would be defined as either textbook or ambiguous by using the rater's response stings. Each rater provided a set of textbook cases. Thus, each case received a textbook "score" between 1 to 5 indicating how

many raters defined that case as textbook. Cases that received a score of ≥ 4 were selected as textbook cases. A total of 19 cases were identified as textbook. Additionally, each rater provided a set of ambiguous cases. Ambiguous cases also received an ambiguous "score" between 1 to 5 indicating how many raters defined that case as ambiguous. Cases that received a score of ≥ 4 were selected as ambiguous cases. A total of 22 cases were defined as ambiguous. The remaining cases were classified as approximable (i.e., neither textbook or ambiguous). Agreement coefficients associated with percent agreement, Cohen's Kappa and Gwet's AC_1 were obtained and one-way repeated measures ANOVAs were conducted for both textbook and ambiguous cases. The interchangeability of raters and prevalence rates were also analyzed among textbook and ambiguous cases as previously described. Finally, the discrepancies between Gwet's AC_1 and Cohen's Kappa were calculated and compared in the two conditions.

Table 7

Example response strings that would be classified as textbook or ambiguous based on the Likert-type responses per case and per rater

Category Description					Certainty Classification	
Mental Illness	Anger	Collateral Damage	Commission of a Crime	Lone Actor	Textbook	Ambiguous
5	1	1	1	1	1	0
1	4	1	1	1	1	0
4	1	1	2	1	1	0
4	4	1	1	3	0	1
5	5	5	5	5	0	1
1	2	1	3	2	0	1

Additionally, in order to communicate the results of this reliability study to a larger audience, benchmarking guidelines were also provided. In the literature, three benchmarking guidelines (Altman, 1991; Fleiss, 1981; Landis & Koch, 1977) have been proposed and are displayed in Table 8. In practice, the models are used with Cohen's Kappa and other agreement coefficients such as Gwet's AC_1 (Gwet, 2014).

Table 8

Proposed benchmarking guidelines for Cohen's Kappa and other agreement coefficients

	Kappa Statistic	Criteria
Landis & Koch (1977)	< 0.0	Poor
	0.0 to 0.20	Slight
	0.21 to 0.40	Fair
	0.41 to 0.60	Moderate
	0.61 to 0.80	Substantial
	0.81 to 1.00	Almost Perfect
Fleiss (1981)	< 0.40	Poor
	0.40 to 0.75	Intermediate to good
	More than 0.75	Excellent
Altman (1991)	< 0.20	Poor
	0.21 to 0.40	Fair
	0.41 to 0.60	Moderate
	0.61 to 0.80	Good
	0.81 – 1.00	Very Good

CHAPTER IV

FINDINGS

This study was a secondary analysis, IRR experiment using data collected from five raters to compare the magnitude of agreement coefficients (i.e., percent agreement, Cohen's Kappa and Gwet's AC_1) against prevalence rates and rater uncertainty. A total of ten rater pairs were created among the five raters. Further, a total of 219 mass shooting incidents were retained for expert review. However, once missing data was excluded the number of mass shooting incidents ranged from 189 to 219 among the 10 rater pairs. In order to study rater uncertainty, mass shooting incidents were classified as textbook or ambiguous. A total of 19 cases were identified as textbook, whereas, a total of 22 cases were identified as ambiguous. No missing values were identified for the textbook cases, however, once missing data for both the Likert-type responses and nominal classifications were excluded for ambiguous cases the number of mass shooting incidents ranged from 20 to 22 among the 10 rater pairs. In other words, one rater completed all the Likert-type responses for each ambiguous mass shooting incident but did not classify two of those cases into the nominal category. The datasets were used to study the observable discrepancies between percent agreement, Cohen's Kappa, and Gwet's AC_1 and to satisfy three objectives: (1) determine if there was a statistically significant mean difference

between percent agreement, Cohen's Kappa, and Gwet's AC_1, (2) to examine how the conditions of the coefficients are affected by prevalence rates, and (3) to study whether there are observable discrepancies between Cohen's Kappa and Gwet's AC_1 for cases that are classified as textbook compared to cases that are classified as ambiguous.

Base Rates

Individually the raters classified the mass shooting incidents into the Anger category with the highest frequency. As shown in Table 9, classification percentages for the Anger category ranged from 36.5% to 71.2% among the five raters. Classification percentages for the Collateral category ranged from 1.4% to 27.9% across all five raters and demonstrated the second highest variability. The motivations of mass shooting incidents were less likely to be categorized as Commission of a Crime and demonstrated the least amount of variability; classification percentages for this category ranged from 1.8% to 3.2% across of all five raters. Finally, classification percentages associated with Mental Illness ranged from 13.2% to 24.7% and classifications associated with Lone Actor ranged from 5.5% to 9.1%. The descriptive statistics for each category across all five raters can be seen in Table 9 including the number and percentage of cases that each rater classified as textbook and ambiguous.

Inter-Rater Reliability Analysis

The following section addresses the first research question: Is there a statistically significant mean difference between percent agreement, Cohen's Kappa, and Gwet's AC_1? As shown in Table 10, statistics associated with percent agreement, Cohen's Kappa, and Gwet's AC_1 were computed across all 10 rater pairs. A one-factor repeated measures

Table 9

The number and percentage of selected categories by each rater including the number of cases each rater classified as textbook and ambiguous

Raters	A	B	C	D	E
Category	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)
Anger	91 (41.6)	156 (71.2)	109 (49.8)	153 (69.9)	80 (36.5)
Collateral	52 (23.7)	3 (1.4)	50 (22.8)	3 (1.4)	61 (27.9)
Commission of a Crime	7 (3.2)	4 (1.8)	6 (2.7)	4 (1.8)	6 (2.7)
Mental Illness	29 (13.2)	41 (18.7)	42 (19.2)	30 (13.7)	54 (24.7)
Lone Actor	20 (9.1)	15 (6.8)	12 (5.5)	14 (6.4)	18 (8.2)
Missing	20 (9.1)	0.0 (0.0)	0.0 (0.0)	15 (6.8)	0.0 (0.0)
Total	219 (100)	219 (100)	219 (100)	219 (100)	219 (100)
Ambiguity Classification	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)
Textbook	86 (39.3)	21 (9.6)	62 (28.3)	120 (54.8)	66 (30.1)
Ambiguous	85 (38.8)	121 (55.3)	88 (40.2)	46 (54.8)	62 (28.3)
Approximable	42 (19.2)	77 (35.2)	69 (31.5)	53 (24.2)	91 (42.6)
Missing	6 (2.7)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Total	219	219	219	219	219

Note: The first section of the table represents the nominal classifications and the second half of the table represents the ambiguity classifications based on the Likert-type responses. Total sample sizes between nominal classifications and ambiguity classifications per rater may vary because some raters completed all the Likert-type responses but did complete the nominal classifications and vice versa.

Table 10

Percent agreement, Cohen's Kappa, and Gwet's AC_1 for all 10 rater pairs including the respective standard errors and confidence intervals associated with each agreement coefficient

	Rater Pairs									
	A & B	A & C	A & D	A & E	B & C	B & D	B & E	C & D	C & E	D & E
PA	.59	.71	.57	.64	.62	.82	.56	.60	.67	.50
SE	.03	.03	.04	.03	.03	.03	.03	.03	.03	.04
C.I.	[.52, .66]	[.65, .78]	[.50, .62]	[.57, .71]	[.55, .68]	[.77, .88]	[.50, .63]	[.54, .67]	[.60, .73]	[.43, .57]
Kappa	.36	.58	.32	.50	.36	.60	.36	.34	.53	.28
SE	.05	.05	.05	.05	.05	.06	.04	.05	.05	.04
C.I.	[.26, .45]	[.49, .67]	[.22, .42]	[.40, .59]	[.26, .46]	[.48, .71]	[.28, .44]	[.24, .44]	[.44, .62]	[.20, .37]
AC_1	.52	.65	.49	.56	.55	.80	.48	.54	.60	.41
SE	.04	.04	.04	.04	.04	.03	.04	.04	.04	.04
C.I.	[.44, .61]	[.58, .73]	[.40, .58]	[.49, .64]	[.47, .63]	[.74, .86]	[.40, .56]	[.46, .62]	[.52, .67]	[.32, .49]
N	199	199	189	199	219	204	219	204	219	204

Note: PA = percent agreement; *SE* = standard error; C.I. = confidence interval; AC_1 = Gwet's first order agreement coefficient; N = sample size

ANOVA was conducted to determine if there was a mean difference among the three agreement coefficients. Mauchly's Test of Sphericity indicated that the assumption of sphericity had been violated, $\chi^2(2) = 9.17$, p = .01. Lomax and Hahs-Vaughn (2012) recommend using multivariate results and a different set of univariate results when the sphercity assumption has not been met. Specifically, the authors suggest reporting Greenhouse-Geisser results when epsilon is ≤ .75. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = .59). Multivariate tests revealed a statistically significant multivariate mean difference, Λ = .03, $F(2, 8) = 134.92$, $p = .00$. Further, univariate results indicated there was a significant difference among percent agreement, Cohen's Kappa, and Gwet's AC_1 across the 10 rater pairs, $F(1.19, 10.70) = 108.59$, $p = .00$. The effect size and observed power were as follows: partial eta squared = .92, observed power = 1.00. Bonferroni multiple comparison procedures (MCPs) revealed statistically significant differences among all pairs of agreement coefficients. The means and standard deviations for the three agreement coefficients were as follows: M = .63 (SD = .09) for percent agreement, M = .42 (SD = .12) for Cohen's Kappa, and M = .56 (SD = .11) for Gwet's AC_1.

Interchangeability of raters. The descriptive statistics associated with the three agreement coefficients across all 10 rater pairs are displayed in Table 11. For completeness and to examine the assumption of interchangeability of raters, the variability of the coefficients for percent agreement, Cohen's Kappa, and Gwet's AC_1 were examined. Standard deviations ranged from .09 to .12 with percent agreement demonstrating the least amount of variability and Cohen's Kappa demonstrating the greatest amount of variability.

This suggests that raters are the least interchangeable with Cohen's Kappa and the raters are slightly more interchangeable with Gwet's AC_1.

Table 11

The descriptive statistics associated with each agreement coefficient across all 10 rater pairs

Agreement Coefficient	*M*	*SD*	*Var*	N
PA	.63	.09	.008	10
Kappa	.42	.12	.014	10
AC_1	.56	.11	.011	10

Note: *M* = mean; *SD* = standard deviation; *Var* = Variance; N = sample size; PA = percent agreement; AC_1 = Gwet's AC_1

Prevalence rates. The following section addresses the second research question: Are there observable discrepancies between Cohen's Kappa and Gwet's AC_1 in the presence of high prevalence rates? Both Table 12 and Figure 2 display the prevalence rates for each category per rater pair. For all 10 rater pairs, Gwet's AC_1 more closely approximated percent agreement. In other words, there was less discrepancy between Gwet's AC_1 and percent agreement than Cohen's Kappa and percent agreement. The variance across the agreed upon categories for each rater pair (i.e., prevalence index) and the discrepancies between Gwet's AC_1 and Cohen's Kappa is depicted in Table 13. The discrepancies between the two chance-corrected agreement coefficients were the greatest when the data was more highly skewed. For example, the variance among categories for raters B and D was calculated as 3144.00 and the discrepancy between Cohen's Kappa and Gwet's AC_1 was calculated as 20.00. Likewise, when the data was less skewed, the discrepancies between Gwet's AC_1 and Cohen's Kappa was smaller. For example, the

Table 12

Prevalence rates per category, agreement coefficients, and sample size for each rater pair

	Category								
	Anger	Collateral	Commission of a Crime	Lone Actor	Mental Illness	PA	Kappa	AC_1	N
Rater Pairs									
A & B	82 (41.2)	1 (0.5)	3 (1.5)	12 (6.0)	20 (10.1)	.59	.36	.52	199
A & C	69 (34.7)	36 (18.1)	4 (2.0)	24 (12.1)	9 (4.5)	.71	.58	.65	199
A & D	75 (39.7)	1 (0.5)	4 (2.1)	15 (7.9)	12 (6.3)	.57	.32	.49	189
A & E	52 (26.1)	33 (16.6)	4 (2.0)	24 (12.1)	14 (7.1)	.64	.50	.56	199
B & C	92 (42.0)	1 (0.5)	3 (1.4)	28 (12.8)	11 (5.0)	.62	.36	.55	219
B & D	133 (65.2)	1 (0.5)	3 (1.5)	20 (9.8)	11 (5.5)	.82	.60	.80	204
B & E	75 (34.2)	0 (0.0)	4 (1.8)	31 (14.2)	13 (5.9)	.56	.36	.48	219
C & D	87 (42.6)	1 (0.5)	4 (2.0)	22 (10.8)	9 (4.4)	.60	.34	.54	204
C & E	62 (28.3)	34 (15.5)	4 (1.8)	37 (16.9)	9 (4.1)	.67	.53	.60	219
D & E	64 (31.3)	1 (0.5)	4 (2.0)	22 (10.8)	11 (5.4)	.50	.28	.41	204

Note: The percentage of agreed upon cases per category is displayed in parentheses. PA = percent agreement; N = sample size.

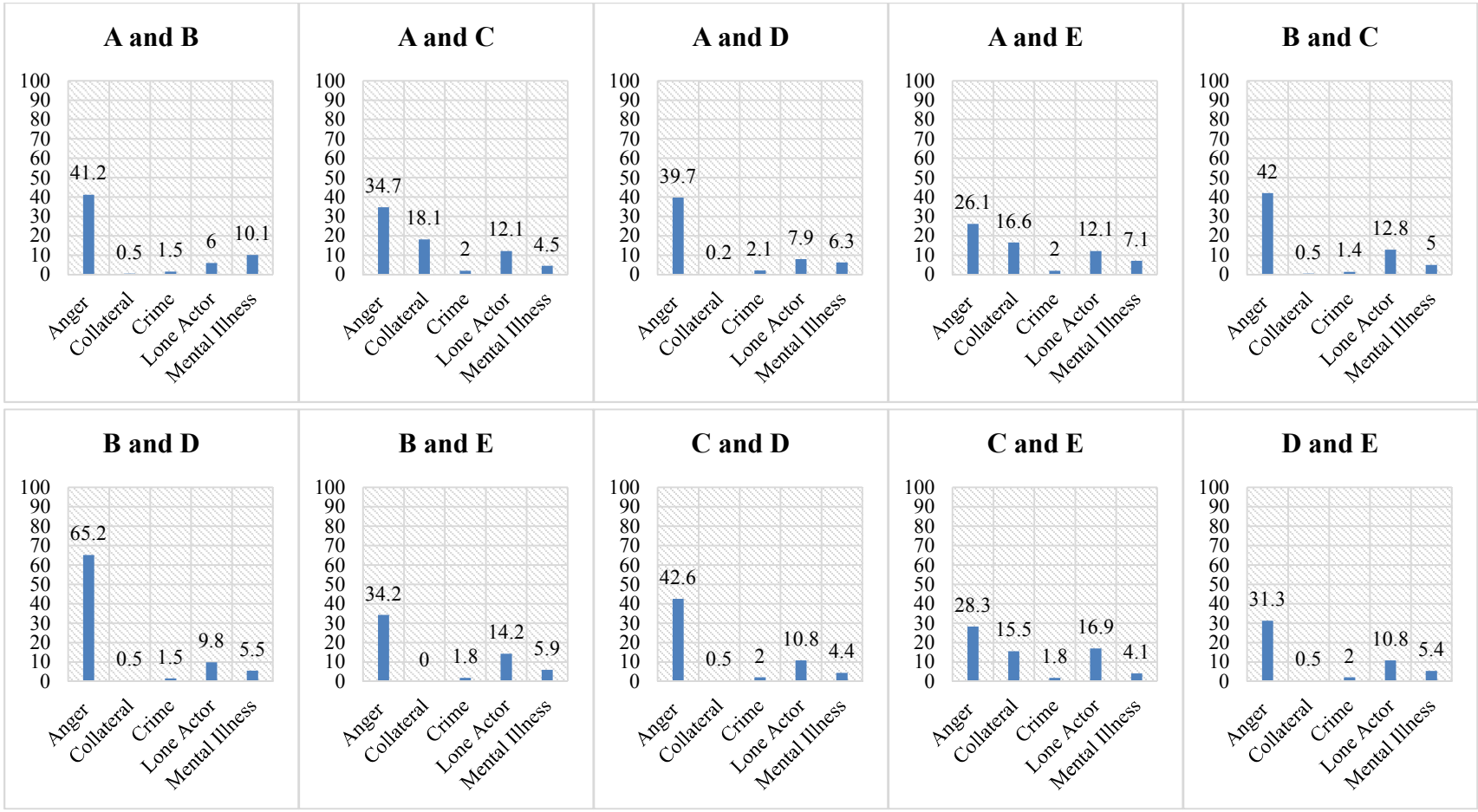

Figure 2. The percentage of agreed upon cases for each category per rater pair.

variance between raters A and E was calculated as 339.00 and the discrepancy between the chance-corrected agreement coefficients was 6.00. A Pearson product-moment correlation coefficient was computed to assess the relationship between the prevalence index and the discrepancies between Gwet's AC_1 and Cohen's Kappa for each rater pair. There was a strong, positive correlation between the two variables, $r = .71$, $p = .02$ indicating that as the prevalence index increases the discrepancy between Gwet's AC_1 and Cohen's Kappa also increases. The discrepancy reflects how much larger Gwet's AC_1 is compared to Kappa. Thus, this result indicates that the AC_1 advantage over Kappa increases as prevalence increases or at least that Kappa's relatively larger correction for chance agreement increases as prevalence increases.

Table 13

Prevalence Index (variance across agreed upon categories) for each rater pair and the discrepancy between Gwet's AC_1 and Cohen's Kappa

Rater Pairs	Prevalence Index	Discrepancy
A & B	1123.00	16.00
A & C	674.00	7.00
A & D	930.00	17.00
A & E	339.00	6.00
B & C	1434.00	19.00
B & D	3144.00	20.00
B & E	936.00	12.00
C & D	1281.00	20.00
C & E	551.00	7.00
D & E	549.00	13.00

The next sections will display and report the results disaggregated by textbook versus ambiguous cases. Textbook cases were defined as cases that involved "obvious' subjects that are associated with a 'true' category membership, whereas, ambiguous cases were defined as cases that involved subjects that required 'random' guessing concerning category membership. A total of 19 textbook cases were identified. The number of mass shooting incidents for the ambiguous cases ranged from 20 to 22 among the five raters once missing data were excluded. Cases considered approximable (i.e., not textbook and not ambiguous) were not included in these analyses.

The following research questions will be addressed first for textbook cases and then for ambiguous cases: (1) Is there a statistically significant mean difference between percent agreement, Cohen's Kappa, and Gwet's AC_1? and (2) Are there observable discrepancies between Cohen's Kappa and Gwet's AC_1 in the presence of high prevalence rates? The final section will include a discussion of the observable discrepancies between Cohen's Kappa and Gwet's AC_1 in order to address the final research question: Are there observable discrepancies between Cohen's Kappa and Gwet's AC_1 for cases that are classified as textbook compared to cases that are classified as ambiguous?

Textbook cases

As shown in Table 14, statistics associated with percent agreement, Cohen's Kappa, and Gwet's AC_1 were computed across all 10 rater pairs among mass shooting incidents classified as textbook cases. Coefficients were relatively high except for raters D and E. Further, Gwet's AC_1 was uniformly higher compared to Cohen's Kappa unless there was perfect agreement between the raters. A one-factor repeated measures ANOVA was conducted to determine if there was a statistically significant mean difference between the

Table 14

Percent agreement, Cohen's Kappa, and Gwet's AC_1 among textbook cases for all 10 rater pairs including the respective standard errors and confidence intervals associated with each agreement coefficient

	Rater Pairs									
	A & B	A & C	A & D	A & E	B & C	B & D	B & E	C & D	C & E	D & E
PA	.95	.95	.84	.79	1.00	.89	.84	.89	.84	.74
SE	.05	.05	.09	.10	.00	.07	.09	.07	.09	.10
C.I.	[.84, 1.00]	[.84, 1.00]	[.66, 1.00]	[.59, .99]	[1.00, 1.00]	[.74, 1.00]	[.66, 1.00]	[.74, 1.00]	[.66, 1.00]	[.52, .96]
Kappa	.90	.90	.69	.65	1.00	.77	.73	.77	.73	.54
SE	.10	.10	.16	.15	.00	.15	.14	.15	.14	.16
C.I.	[.70, 1.00]	[.70, 1.00]	[.34, 1.00]	[.34, .96]	[1.00, 1.00]	[.45, 1.00]	[.44, 1.00]	[.45, 1.00]	[.44, 1.00]	[.19, .88]
AC_1	.93	.93	.79	.74	1.00	.86	.80	.86	.80	.68
SE	.07	.07	.12	.12	.00	.10	.11	.10	.11	.13
C.I.	[.78, 1.00]	[.78, 1.00]	[.55, 1.00]	[.48, .99]	[1.00, 1.00]	[.66, 1.00]	[.58, 1.00]	[.66, 1.00]	[.58, 1.00]	[.41, .95]
N	19	19	19	19	19	19	19	19	19	19

Note: PA = percent agreement; *SE* = standard error; C.I. = confidence interval; AC_1 = Gwet's first order agreement coefficient; *N* = sample size

three agreement coefficients among textbook cases. Mauchly's Test of Sphericity indicated that the assumption of sphericity had been violated, $\chi^2(2) = 26.05$, p = .00. Multivariate tests revealed a statistically significant multivariate mean difference, $\Lambda = .20$, $F(2, 8) = 16.28$, $p = .00$. Further, univariate results indicated there was a significant difference between percent agreement, Cohen's Kappa, and Gwet's AC_1 across the 10 rater pairs, $F(1.02, 9.18) = 32.44$, $p = .00$. The effect size and observed power were as follows: partial eta squared = .78, observed power = 1.00. Bonferroni MCPs revealed statistically significant differences among all pairs of agreement coefficients. The means and standard deviations for the three agreement coefficients were as follows: M = .87 (SD = .08) for percent agreement, M = .77 (SD = .13) for Cohen's Kappa, and M = .84 (SD = .10) for Gwet's AC_1.

Interchangeability of raters. For the textbook cases, descriptive statistics associated with the three agreement coefficients across all 10 rater pairs are displayed in Table 15. In order to examine the assumption of interchangeability of raters, the variance of the coefficients between percent agreement, Cohen's Kappa, and Gwet's AC_1 were examined. Standard deviations ranged from .08 to .13. Again, Cohen's Kappa demonstrated the greatest variability across the 10 raters, whereas, percent agreement demonstrated the least amount of variability. This suggests that raters are the least interchangeable with Cohen's Kappa and the raters are slightly more interchangeable with Gwet's AC_1.

Table 15

The descriptive statistics associated with each agreement coefficient across all 10 rater pairs among cases classified as textbook

Agreement Coefficient	*M*	*SD*	*Var*	N
PA	.87	.08	.006	10
Kappa	.77	.13	.018	10
AC_1	.84	.10	.009	10

Note: *M* = mean; *SD* = standard deviation; *Var* = variance; N = sample size; PA = percent agreement; AC_1 = Gwet's AC_1

Prevalence rates. Both Table 16 and Figure 3 display the prevalence rates for each category that were calculated for each rater pair among textbook cases. For all 10 rater pairs, Gwet's AC_1 more closely approximated percent agreement compared to Cohen's Kappa. In other words, there was less discrepancy between Gwet's AC_1 and percent agreement compared to Cohen's Kappa and percent agreement. The variance across the agreed upon categories (i.e., prevalence index), the discrepancies between Gwet's AC_1 and Cohen's Kappa, and the number of disagreements for each rater pair is depicted in Table 17. In relation to prevalence rates, a Pearson product-moment correlation coefficient indicated that the relationship between the prevalence index and the discrepancy between Gwet's AC_1 and Cohen's Kappa was not statistically significant, $r = -.54$, $p = .11$. For textbook cases, the discrepancies between Gwet's AC_1 and Cohen's Kappa were the greatest when more disagreements could be observed between two raters. For example, raters D and E disagreed when classifying five out of the 19 cases and the discrepancy between Gwet's AC_1 and Cohen's Kappa was calculated as 14.00. Likewise, when perfect agreement was found between two raters (i.e., raters B and C) the discrepancy between

Table 16

The prevalence rates per category, agreement coefficients, and sample size for each rater pair among textbook cases

	Category								
	Anger	Collateral	Commission of a Crime	Lone Actor	Mental Illness	PA	Kappa	AC_1	N
Rater Pairs									
A & B	11 (57.9)	0 (0.0)	1 (5.3)	0 (0.0)	6 (31.6)	.95	.90	.93	19
A & C	11 (57.9)	0 (0.0)	1 (5.3)	0 (0.0)	6 (31.6)	.95	.90	.93	19
A & D	11 (57.9)	0 (0.0)	1 (5.3)	0 (0.0)	4 (21.1)	.85	.69	.79	19
A & E	8 (42.1)	0 (0.0)	1 (5.3)	0 (0.0)	6 (31.6)	.79	.65	.74	19
B & C	12 (63.2)	0 (0.0)	1 (5.3)	0 (0.0)	6 (31.6)	1.00	1.00	1.00	19
B & D	12 (63.2)	0 (0.0)	1 (5.3)	0 (0.0)	4 (21.1)	.89	.77	.86	19
B & E	9 (47.4)	0 (.00)	1 (5.3)	0 (0.0)	6 (31.6)	.84	.73	.80	19
C & D	12 (63.2)	0 (0.0)	1 (5.3)	0 (0.0)	4 (21.1)	.89	.77	.86	19
C & E	9 (47.4)	0 (0.0)	1 (5.3)	0 (0.0)	6 (31.6)	.84	.73	.80	19
D & E	9 (47.4)	0 (0.0)	1 (5.3)	0 (0.0)	4 (21.1)	.74	.54	.68	19

Note: The percentage of agreed upon cases per category is displayed in parentheses. PA = percent agreement; N = sample size.

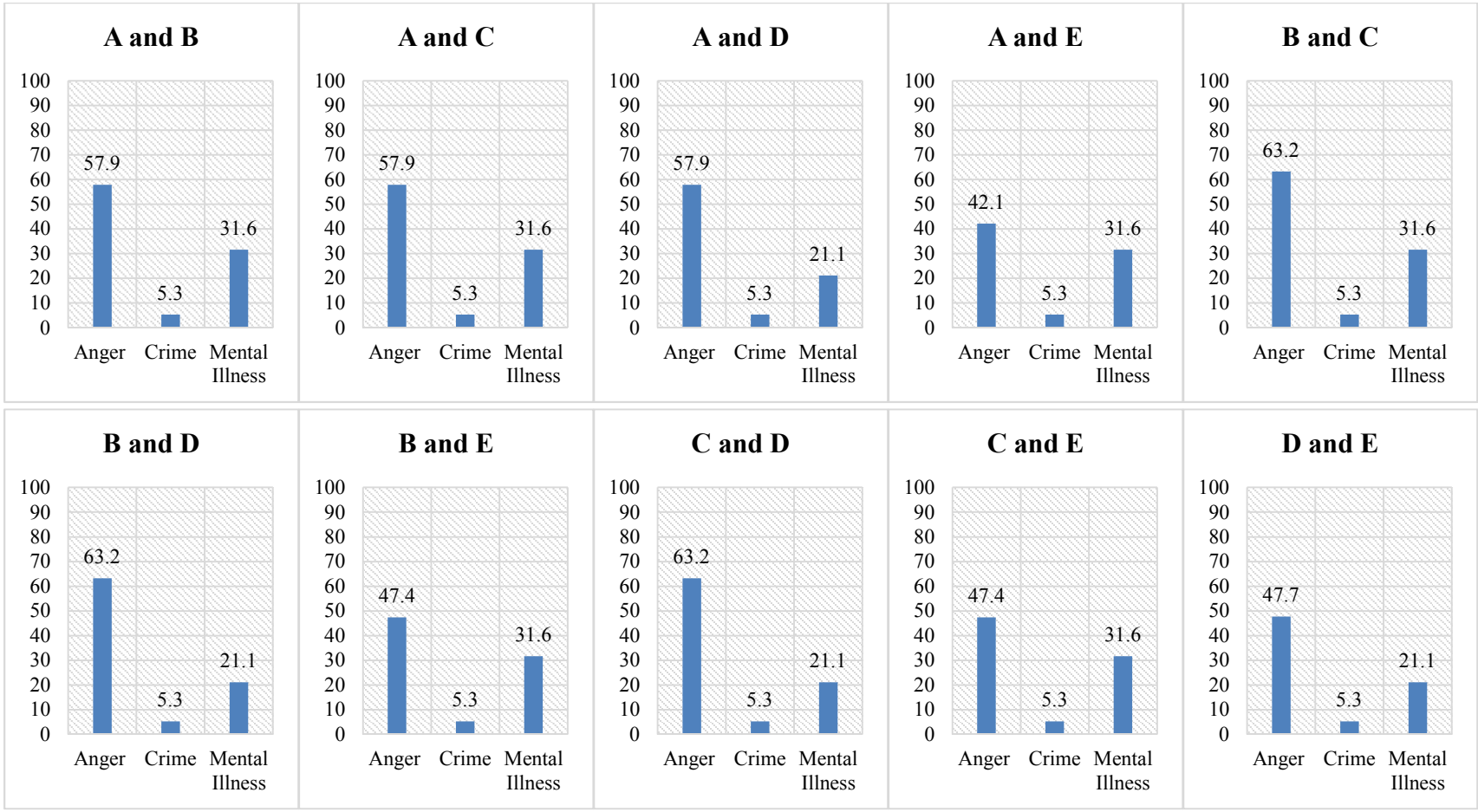

Figure 3. The prevalence rates for each category per rater pair among cases classified as textbook

Gwet's AC_1 and Cohen's Kappa was 0. In other words, the less disagreements between the raters the less discrepancy between Gwet's AC_1 and Cohen's Kappa. A Pearson product-moment correlation coefficient was computed to assess the relationship between the number of disagreements between raters and the discrepancies between Gwet's AC_1 and Cohen's Kappa for each rater pair. There was a strong, positive correlation between the two variables, $r = .89$, $p = .00$ indicating that as the number of disagreements between rater pairs increase the discrepancy between Gwet's AC_1 and Cohen's Kappa increases.

Table 17

Prevalence Index (variance across agreed upon categories) for each rater pair, the discrepancy between Gwet's AC_1 and Cohen's Kappa, and the number of disagreements between the rater pairs among textbook cases

Rater Pairs	Prevalence Index	Discrepancy	Disagreements
A & B	23.30	3.00	1
A & C	23.30	3.00	1
A & D	21.70	10.00	3
A & E	14.00	9.00	4
B & C	27.20	0.00	0
B & D	25.80	9.00	2
B & E	16.70	7.00	3
C & D	25.80	9.00	2
C & E	16.70	7.00	3
D & E	14.70	14.00	5

Ambiguous Cases

Table 18 depicts the statistics associated with percent agreement, Cohen's Kappa, and Gwet's AC_1 computed across all 10 rater pairs among mass shooting incidents

Table 18

Percent agreement, Cohen's Kappa, and Gwet's AC_1 among ambiguous cases for all 10 rater pairs including the respective standard errors and confidence intervals associated with each agreement coefficient

	Rater Pairs									
	A & B	A & C	A & D	A & E	B & C	B & D	B & E	C & D	C & E	D & E
PA	.50	.65	.45	.70	.55	.68	.45	.64	.68	.45
SE	.12	.12	.12	.12	.11	.10	.11	.10	.10	.11
C.I.	[.25, .57]	[.41, .90]	[.21, .70]	[.46, .94]	[.32, .77]	[.47, .89]	[.30, .68]	[.42, .86]	[.47, .89]	[.30, .68]
Kappa	.35	.54	.28	.60	.36	.46	.28	.49	.57	.28
SE	.13	.14	.13	.14	.14	.16	.12	.14	.13	.12
C.I.	[.07, .62]	[.25, .84]	[.00, .55]	[.31, .89]	[.08, .64]	[.12, .80]	[.03, .54]	[.21, .78]	[.30, .84]	[.02, .54]
AC_1	.35	.57	.28	.60	.45	.60	.29	.56	.61	.28
SE	.15	.14	.15	.15	.13	.13	.14	.13	.13	.14
C.I.	[.04, .67]	[.28, .86]	[-.04, .60]	[.30, .90]	[.18, .73]	[.33, .88]	[-.01, .59]	[.29, .83]	[.35, .87]	[-.02, .59]
N	20	20	20	20	22	22	22	22	22	22

Note: PA = percent agreement; *SE* = standard error; C.I. = confidence interval; AC_1 = Gwet's first order agreement coefficient; N = sample size

classified as ambiguous cases. A one-factor repeated measures ANOVA was conducted to determine if there was a mean difference between percent agreement, Cohen's Kappa, and Gwet's AC_1 across the 10 rater pairs among cases classified as ambiguous. Mauchly's Test of Sphericity indicated that the assumption of sphericity had been met, $\chi^2(2) = .61$, $p = .74$. Tests of within-subjects effects indicated there was a significant difference between percent agreement, Cohen's Kappa, and Gwet's AC_1 across the 10 rater pairs, $F(2, 18) = 70.30$, $p = .00$. The effect size and observed power were as follows: partial eta squared = .89, observed power = 1.00. Bonferroni MCPs revealed statistically significant differences among all pairs of agreement coefficients except for Cohen's Kappa and Gwet's AC_1. The means and standard deviations for the three agreement coefficients were as follows: M = .58 (SD = .11) for percent agreement, M = .42 (SD = .13) for Cohen's Kappa, and M = .46 (SD = .15) for Gwet's AC_1.

Interchangeability of raters. For the ambiguous cases, descriptive statistics associated with the three agreement coefficients across all 10 rater pairs are displayed in Table 19. In order to examine the assumption of interchangeability of raters, the variance of the coefficients between percent agreement, Cohen's Kappa, and Gwet's AC_1 were examined. Standard deviations ranged from .11 to .15 with percent agreement demonstrating the least amount of variability and Gwet's AC_1 demonstrating the greatest amount of variability. This suggests that raters are the least interchangeable with Gwet's AC_1 and the raters are slightly more interchangeable with Cohen's Kappa among ambiguous cases.

Table 19

The descriptive statistics associated with each agreement coefficient across all 10 rater pairs among cases classified as ambiguous

Agreement Coefficient	*M*	*SD*	*Var*	N
PA	.58	.11	.011	10
Kappa	.42	.13	.016	10
AC_1	.46	.15	.021	10

Note: *M* = mean; *SD* = standard deviation; *Var* = variance; N = sample size; PA = percent agreement; AC_1 = Gwet's AC_1

Prevalence rates. Both Table 20 and Figure 4 display the prevalence rates for each category that were calculated for each rater pair among ambiguous cases. For 6 of the rater pairs, Gwet's AC_1 more closely approximated percent agreement compared to Cohen's Kappa. The variance across the agreed upon categories (i.e., prevalence index) for each rater pair and the discrepancies between Gwet's AC_1 and Cohen's Kappa is depicted in Table 21. The discrepancies between the two chance-corrected agreement coefficients were the greatest when the data was more highly skewed. For example, the variance between the categories for raters B and D was calculated as 17.00 and the discrepancy between Cohen's Kappa and Gwet's AC_1 was calculated as 14.00. For 4 of the rater pairs, the discrepancy between Gwet's AC_1 and Cohen's Kappa was calculated as 0.00; under such conditions, the prevalence index ranged from 2.30 to 4.70. In other words, when the data was less skewed, the discrepancies between Gwet's AC_1 and Cohen's Kappa was smaller. A Pearson product-moment correlation coefficient was computed to assess the relationship between the prevalence index and the discrepancies between Gwet's AC_1 and Cohen's Kappa for each rater pair. There was a strong, positive correlation between the two

Table 20

The number and percentage of agreed upon cases per category, agreement coefficients, and sample size for each rater pair among cases classified as ambiguous

Category									
	Anger	Collateral	Commission of a Crime	Lone Actor	Mental Illness	PA	Kappa	AC_1	N
Rater Pairs									
A & B	4 (20.0)	0 (0.0)	0 (0.0)	3 (15.0)	3 (15.0)	.50	.35	.35	20
A & C	3 (15.0)	4 (20.0)	0 (0.0)	2 (10.0)	4 (20.0)	.65	.54	.57	20
A & D	4 (20.0)	0 (0.0)	0 (0.0)	2 (10.0)	3 (15.0)	.45	.28	.28	20
A & E	2 (10.0)	6 (30.0)	0 (0.0)	3 (15.0)	3 (15.0)	.70	.60	.60	20
B & C	6 (27.3)	0 (0.0)	0 (0.0)	2 (9.1)	4 (18.2)	.55	.36	.45	22
B & D	10 (45.5)	0 (0.0)	0 (0.0)	2 (9.1)	3 (13.6)	.68	.46	.60	22
B & E	4 (18.2)	0 (0.0)	0 (0.0)	2 (9.1)	4 (18.2)	.45	.28	.29	22
C & D	6 (27.3)	1 (4.5)	0 (0.0)	2 (9.1)	5 (22.7)	.64	.49	.56	22
C & E	4 (18.2)	5 (22.7)	0 (0.0)	1 (4.5)	5 (22.7)	.68	.57	.61	22
D & E	4 (18.2)	1 (4.5)	0 (0.0)	1 (4.5)	4 (18.2)	.45	.28	.28	22

Note: The percentage of agreed upon cases per category is displayed in parentheses. PA = percent agreement; N = sample size.

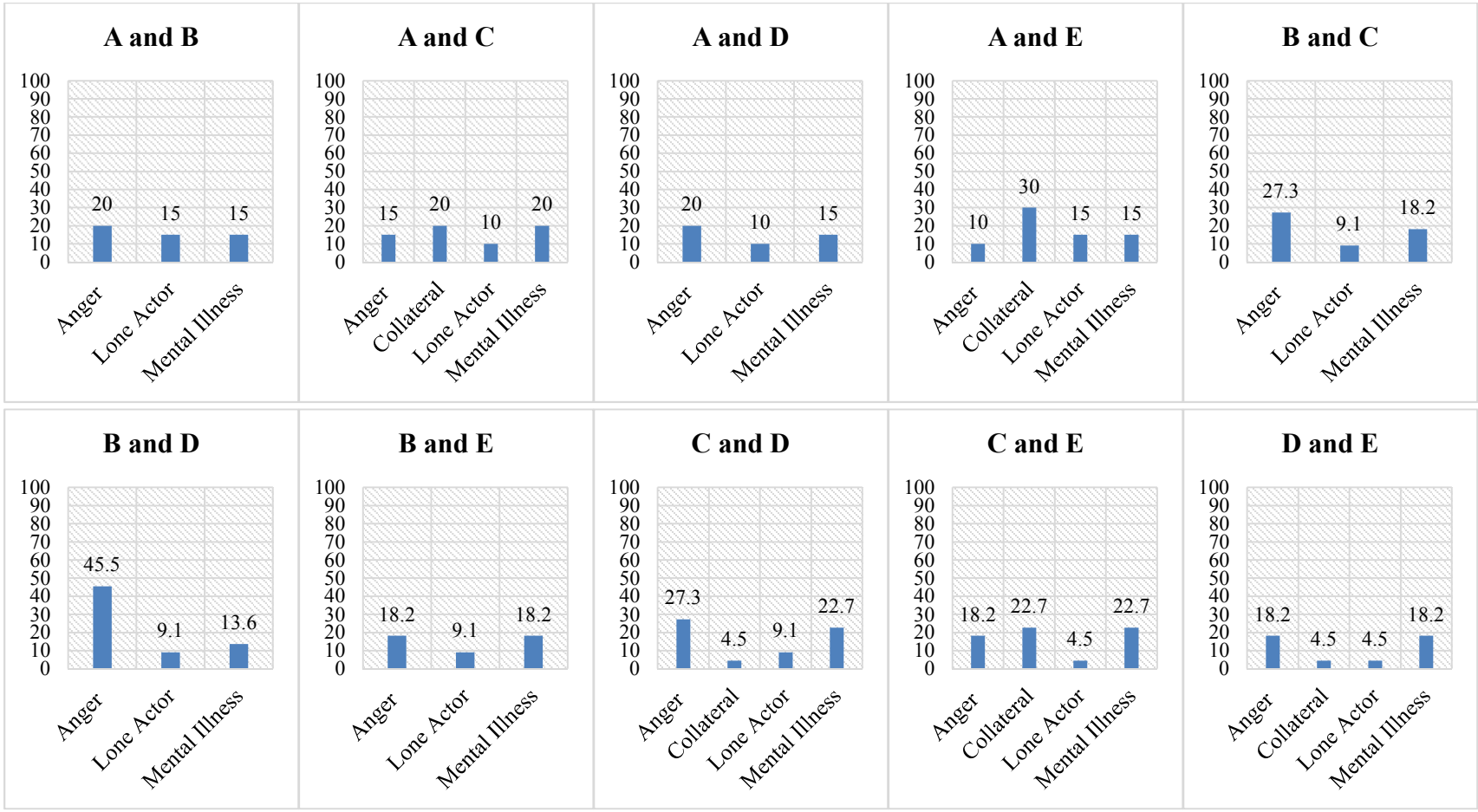

Figure 4. The percentage of agreed upon cases for each category per rater pair among cases classified as ambiguous.

variables, $r = .89$, $p = .00$ indicating that as the prevalence index increases the discrepancy between Gwet's AC_1 and Cohen's Kappa increases.

Table 21

Prevalence Index (variance across agreed upon categories) for each rater pair and the discrepancy between Gwet's AC_1 and Cohen's Kappa among ambiguous cases

Rater Pairs	Prevalence Index	Discrepancy
A & B	3.50	0.00
A & C	2.80	3.00
A & D	3.20	0.00
A & E	4.70	0.00
B & C	6.80	9.00
B & D	17.00	14.00
B & E	4.00	1.00
C & D	6.70	7.00
C & E	5.50	4.00
D & E	3.50	0.00

Discrepancies Between Cohen's Kappa and Gwet's AC_1

The concluding sections address the final research question: Are there observable discrepancies between Cohen's Kappa and Gwet's AC_1 for cases that are classified as textbook compared to cases that are classified as ambiguous? Table 22 displays a side-by-side comparison of the agreement coefficients associated with percent agreement, Cohen's Kappa, and Gwet's AC_1. Additionally, Table 23 depicts the discrepancies between the agreement coefficients for the overall, textbook, and ambiguous analyses.

On average, during both the overall and ambiguous analyses, Gwet's AC_1 (overall: M = 6.80, SD = 1.93; ambiguous: 11.60, SD = 4.14) more closely approximated percent agreement compared to Cohen's Kappa (overall: M = 20.50, SD = 5.08; ambiguous: 15.40,

Table 22

Agreement coefficients associated with percent agreement, Cohen's Kappa, and Gwet's AC_1 for the overall, textbook, and ambiguous analyses

Rater Pairs	Overall Analysis			Textbook Cases			Ambiguous Cases		
	PA	K	AC_1	PA	K	AC_1	PA	K	AC_1
A & B	.59	.36	.52	.95	.90	.93	.50	.35	.35
A & C	.71	.58	.65	.95	.90	.93	.65	.54	.57
A & D	.57	.32	.49	.85	.69	.79	.45	.28	.28
A & E	.64	.50	.56	.79	.65	.74	.70	.60	.60
B & C	.62	.36	.55	1.00	1.00	1.00	.55	.36	.45
B & D	.82	.60	.80	.89	.77	.86	.68	.46	.60
B & E	.56	.36	.48	.84	.73	.80	.45	.28	.29
C & D	.60	.34	.54	.89	.77	.86	.64	.49	.56
C & E	.67	.53	.60	.84	.73	.80	.68	.57	.61
D & E	.50	.28	.41	.74	.54	.68	.45	.28	.28

Note: PA = percent agreement; K = Cohen's Kappa; AC_1 = Gwet's AC_1

Table 23

Discrepancies between the agreement coefficients for the overall, textbook, and ambiguous analyses

Rater Pairs	Overall Discrepancies			Textbook Discrepancies			Ambiguous Discrepancies		
	PA – K	PA – AC_1	AC_1 – K	PA – K	PA – AC_1	AC_1 – K	PA – K	PA – AC_1	AC_1 – K
A & B	23.00	7.00	16.00	5.00	2.00	3.00	15.00	15.00	0.00
A & C	13.00	6.00	7.00	5.00	2.00	3.00	11.00	8.00	3.00
A & D	25.00	8.00	17.00	16.00	6.00	10.00	17.00	17.00	0.00
A & E	14.00	8.00	6.00	14.00	5.00	9.00	10.00	10.00	0.00
B & C	26.00	7.00	19.00	0.00	0.00	0.00	19.00	10.00	9.00
B & D	22.00	2.00	20.00	12.00	3.00	9.00	22.00	8.00	14.00
B & E	20.00	8.00	12.00	11.00	4.00	7.00	17.00	16.00	1.00
C & D	26.00	6.00	20.00	12.00	3.00	9.00	15.00	8.00	7.00
C & E	14.00	7.00	7.00	11.00	4.00	7.00	11.00	7.00	4.00
D & E	22.00	9.00	13.00	2.00	6.00	14.00	17.00	17.00	0.00
Average	20.50	6.80	13.70	8.80	3.50	7.10	15.40	11.60	3.80
SD	5.08	1.93	5.39	5.39	1.90	4.09	3.84	4.14	4.80

Note: PA = percent agreement; K = Cohen's Kappa; AC_1 = Gwet's AC_1; SD = standard deviation

SD = 3.84). Additionally, during the textbook analysis, Gwet's AC_1 (M = 3.50, SD = 1.90) more closely approximated percent agreement, on average, compared to Cohen's Kappa (M = 8.80, SD = 5.39).

Discrepancies between Gwet's AC_1 and Cohen's Kappa ranged from 0.00 to 14 for both textbook and ambiguous cases. On average, the largest discrepancy between Gwet's AC_1 and Cohen's Kappa can be seen from the overall analysis (M = 13.70, SD = 5.39) and the smallest discrepancy can be seen from the ambiguous analysis (M = 3.80, SD = 4.80). The average discrepancy between Gwet's AC_1 and Cohen's Kappa under the textbook condition was 7.10 (SD = 4.09). In 7 of the 10 rater pairs, discrepancies were larger in textbook cases; for two pairs, discrepancies were larger in ambiguous cases; and for one pair the discrepancies were equal between the two types of cases. In other words, the discrepancy between the two chance-corrected agreement coefficients was larger among textbook cases compared to ambiguous.

CHAPTER V

CONCLUSION

This study compared the magnitude of three agreement coefficients against prevalence rates and rater uncertainty among five raters using a real dataset containing mass shooting incidents. Specifically, the study explored the observable discrepancies between percent agreement, Cohen's Kappa, and Gwet's AC_1 under different conditions of trait prevalence (i.e., skewness in the data) and rater uncertainty (i.e., textbook versus ambiguous cases). Further, as a novel contribution to the literature, the research demonstrated a new methodology for determining which mass shooting incidents could be classified as textbook or ambiguous based on rater responses.

Hypotheses of the Study Revisited

The present study examined the following hypotheses: (1) a statistically significant mean difference would be seen between percent agreement, Cohen's Kappa, and Gwet's AC_1, (2) Cohen's Kappa was expected to overcorrect for chance agreement in the presence of high prevalence rates, and (3) a greater discrepancy between Gwet's AC_1 and Cohen's Kappa would be seen for cases classified as textbook compared to cases that were classified as ambiguous.

The first hypothesis set for this study was supported in that significant differences were found between percent agreement, Cohen's Kappa, and Gwet's AC_1 under all three conditions. Specifically, for the overall analysis (i.e., the dataset containing all mass shooting incidents) and the analysis concerning only textbook cases, significant differences could be observed between all pairwise comparisons. In both conditions, Cohen's Kappa demonstrated significantly lower agreement coefficients across the 10 rater pairs compared to percent agreement and Gwet's AC_1. Additionally, Gwet's AC_1 demonstrated significantly lower agreement coefficients across the 10 rater pairs compared to percent agreement. The findings were consistent with the literature in that percent agreement may overestimate the extent of true agreement among raters because it does not take into account chance-agreement; Likewise, Cohen's Kappa may underestimate the extent of true agreement between raters due to the statistic overcorrecting for chance-agreement (Hripcsak & Heitjan, 2002). Concerning the analysis containing only ambiguous cases, results revealed similar findings in that significant differences were found between percent agreement and both Cohen's Kappa and Gwet's AC_1, however, no significant difference was found between Cohen's Kappa and Gwet's AC_1. Still, Cohen's Kappa demonstrated significantly lower agreement coefficients across the 10 rater pairs compared to percent agreement and Gwet's AC_1.

The second hypothesis was partially held and stated that Kappa was expected to overcorrect for chance agreement in the presence of high prevalence rates. In both the overall and ambiguous analyses, the discrepancies between Gwet's AC_1 and Cohen's Kappa were the greatest when the data was highly skewed. There was a strong, positive correlation between the calculated variance across the agreed upon categories and the

discrepancy between the chance-corrected agreement coefficients under these two conditions. Higher variability was observed across the agreed upon categories among textbook classifications compared to ambiguous classifications; however, the relationship between category variability and the discrepancy between Gwet's AC_1 and Cohen's Kappa was negative and nonsignificant. This could be due to the number of disagreements observed between the raters acting as a moderating variable. For instance, as the number of disagreements between rater pairs increased the discrepancy between Gwet's AC_1 and Cohen's Kappa increased. Future research should investigate whether there is an interaction effect between the number of disagreements observed between rater pairs and the calculated variance across the agreed upon categories with a larger sample size.

The third hypothesis was also supported. Though the mean difference was not significant, a larger discrepancy in favor of Gwet's AC1 was seen between Gwet's AC_1 and Cohen's Kappa among cases classified as textbook compared to cases classified as ambiguous. On average, there was a discrepancy of 3.80 between Gwet's AC_1 and Cohen's Kappa among ambiguous cases, whereas, an average discrepancy of 7.10 was observed between Gwet's AC_1 and Cohen's Kappa among textbook cases. Again, these findings were consistent with the literature in that Cohen's Kappa may underestimate the extent of true agreement between raters due to the statistic overcorrecting for chance-agreement (Hripcsak & Heitjan, 2002). Further, the findings support Gwet's (2014) contention that the Kappa statistic overcorrects for chance-agreement in textbook situations when random guessing is less likely to be a factor. This could be due to

additional factors affecting the Kappa statistic such as prevalence rates and the actual number of disagreements seen between rater pairs.

Implications for IRR Theory

Findings suggested there was a lot of uncertainty during the classification process. For instance, the Likert-ratings indicated that a majority of the mass shooting incidents were classified as approximable or ambiguous across the five raters. Specifically, rater B classified 121 of the cases as ambiguous and 77 of the cases as approximable. Further, textbook classifications across the five raters ranged from 21 to 120 indicating there was large variation in the number of cases classified as textbook. The findings highlighted the importance of using chance-corrected agreement coefficients when conducting IRR experiments.

Among textbook cases, Cohen's Kappa and Gwet's AC_1 were relatively high across the rater pairs except for rater D and E. However, Gwet's AC_1 was uniformly higher compared to Cohen's Kappa unless there was perfect agreement between the raters. The results indicated that raters do not randomly guess when cases are classified as textbook and that the Kappa statistic overcorrects for guessing on textbook cases. Interestingly, when raters were more certain of their classifications, they tended not to use all of the categories. Specifically, the only categories that were utilized among textbook cases were the Anger, In the Commission of a Crime, and Mental Illness categories. Finally, the results showed that the discrepancy between Gwet's AC_1 and Cohen's Kappa was generally larger under conditions of more certainty and that the Kappa statistic modelled actual rater behavior more poorly compared to Gwet's AC_1.

Recommendations for Future Research

Future research should assess how percent agreement, Cohen's Kappa, and Gwet's AC_1 function with respect to varying rater populations. Moreover, future studies can examine whether additional factors may have influenced the observable discrepancies between percent agreement, Cohen's Kappa, and Gwet's AC_1. For example, it may be of interest to determine which cases are considered low and high profile cases and incorporate those differences into the design of the study. Further, rater characteristics should be taken into consideration to determine whether rater bias may have contributed to skewness in the data. Rater base rates from the present research indicated that, on average, raters were slightly more likely to classify cases as ambiguous compared to classifying cases as textbook. However, it is unclear what guided each rater's decision-making process; therefore, rater characteristics such as the amount of their professional experience, area of specialization, and their understanding of the construct of interest and rating scale should be examined.

The present research demonstrated a new methodology for determining which mass shooting incidents could be classified as textbook or ambiguous based on rater responses. However, additional validity evidence should be provided for this new methodology. Replication studies are needed to determine if the same mass shooting incidents can be classified as textbook or ambiguous among additional rater populations. Further, it should be noted whether similar levels of certainty are seen across additional rater populations.

Recommendations for Practice

The coefficients associated with Cohen's Kappa ranged from .28 to .60 among the 10 rater pairs for the overall analysis. According to benchmarking guidelines provided in the literature, the extent of agreement between the raters can be regarded as slight to moderate (Landis & Koch, 1977), poor to intermediate to good (Fleiss, 1981), and fair to moderate (Altman, 1991). However, Hripcsak & Heitjan, (2002) acknowledged that Kappa values between 0 and 1 cannot be interpreted consistently and, therefore, do not recommend the use of such guidelines. This is due to the interpretation of the guidelines relying on additional factors, such as the number of categories, the purpose of the measurement, and the definition of chance-agreement (Hripcsak & Heitjan, 2002). For example, an IRR experiment that contains more levels on its scale will most likely generate a lower Kappa coefficient compared to an IRR experiment that contains less levels. Instead, Hripcsak and Heitjan (2002) stated that the goal of the experiment should be heavily considered in order to determine a level of Kappa that represents acceptable reliability. For instance, in a situation where disagreements between experts about patient diagnoses could have dire consequences for those patients, a higher Kappa coefficient would be more appropriate.

The Likert-type responses provided by each rater per category for each mass shooting incident was used to determine certainty classifications (i.e., textbook versus ambiguous) for the present research. However, it may also be useful to use the Likert-type responses to determine which mass shooting incidents involve overlapping boundaries. That is, diagnostic procedures may be associated with psychiatric disorders that include overlapping boundaries due to comorbidity, e.g., a patient can present with

both anxiety and depression. Although raters were asked to classify mass shooting incidents into a nominal category based on the motivation of the offender, it is probable that many of the shooters could have been driven by more than one motivation.

Limitations

A potential limitation of this study was the homogenous nature of the rater population. The five raters demonstrated similar backgrounds and had extensive experience in the field of Forensic Psychiatry. In terms of generalizability, the research findings and conclusions from this present study cannot be applied to other rater populations (e.g., Neuropsychologists, Criminologists, etc.) at this time. Future studies may find that other rater populations utilize or interpret the classification system and/or mass shooting incidents differently. Further, a relatively small number of textbook (N = 19) and ambiguous cases (N = 22) were identified. A larger sample size may have revealed more observable discrepancies between Cohen's Kappa and Gwet's AC_1 among both textbook and ambiguous cases. Therefore, caution should be used when generalizing these results to other IRR experiments where subjects could be classified as textbook or ambiguous.

Additional factors that were not analyzed could have contributed to the findings. Although the dataset used in this study cannot be considered a true population or random sample of mass shooting incidents, a large number of mass shooting incidents were collected and were intended to constitute a representative sample. However, the dataset contained incidents that were not considered high-profile cases and were less likely to gain media coverage. Therefore, raters may have been less certain when classifying the

low-profile cases due to these cases containing less information about the motivation(s) of the shooter. Future research should also disaggregate the dataset according to low or high-profile cases and examine how Cohen's Kappa and Gwet's AC_1 function within these subsets. Further, rater bias was not assessed. Lorber (2006) stated that raters should remain unbiased in their assessment of subjects, thus, ensuring that raters can be used interchangeably when conducting IRR experiments. In relation to the present study, the assumption of interchangeability of raters was examined by observing the variance of the coefficients across the 10 rater pairs. Results indicated the following: (1) in all three conditions (i.e., overall, textbook, and ambiguous analyses) percent agreement demonstrated the least amount of variability across the 10 rater pairs, (2) Gwet's AC_1 demonstrated less variability across the 10 rater pairs compared to Cohen's Kappa when examining all mass shooting incidents (i.e., overall analysis) and textbook cases, and (3) Cohen's Kappa revealed less variability across the 10 rater pairs compared to Gwet's AC_1 among ambiguous cases. However, rater bias may have served as an additional source of skew in the observational data and IRR analyses (Xu & Lorber, 2017).

Further, the missing data across the raters for both the nominal classifications and Likert-type responses were identified at random. However, the study utilized a fully crossed design and some of the raters may have chosen not to classify some of the mass shooting incidents due to fatigue. In other words, the large number (N = 219) of mass shooting incidents that the raters were asked to classify may have contributed to the observed attrition rates. Future studies may be interested in utilizing other study designs to reduce fatigue such as incomplete block designs, subjects nested within raters, raters nested within subjects, or raters joint with subjects (Hoyt, 2000).

Conclusion

This study was conducted to evaluate the magnitude of percent agreement, Cohen's Kappa, and Gwet's AC_1 against prevalence rates and rater uncertainty using a newly developed mass shooting classification index based on the motivation(s) of the offender. The observable discrepancies between percent agreement, Cohen's Kappa, and Gwet's AC_1 were examined under different conditions of trait prevalence (i.e., skewness in the data). Further, cases were classified as textbook or ambiguous in order to examine how the agreement coefficients function in respect to rater uncertainty.

Results of this study indicated that observable discrepancies between the three agreement coefficients could be seen in all the conditions. Specifically, during all three analyses (i.e., overall, textbook, and ambiguous) percent agreement was likely to overestimate the extent of true agreement among raters and Cohen's Kappa was likely to underestimate the extent of true agreement among raters. The overall and ambiguous analyses revealed larger discrepancies between Gwet's AC_1 and Cohen's Kappa in the presence of highly skewed data, however, discrepancies between Gwet's AC_1 and Cohen's Kappa appeared to be more dependent of the number of observable disagreements between raters during the textbook analysis. Despite the previously discussed limitations, to my knowledge, this study was the first to classify subjects as textbook or ambiguous using a real dataset and to examine the magnitude of agreement coefficients in respect to rater uncertainty.

REFERENCES

Altman, D. G. (1991). *Practical statistics for medical research*. Chapman and Hall.

Blair, J. P., & Schwieit, K. W. (2014). A Study of Active Shooter Incidents in the United States between 2000 and 2013. *US Department of Justice*.

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

Chan, Y. H. (2003). Biostatistics 104: correlational analysis. *Singapore Med J*, *44*(12), 614-9.

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.

Cronbach, L. J. (1955). Processes affecting scores on" understanding of others" and" assumed similarity.". *Psychological bulletin*, *52*(3), 177.

Douglas, J. E., Burgess, A. W., Burgess, A. G., & Ressler, R. K. (2013). *Crime classification manual: A standard system for investigating and classifying violent crime*. John Wiley & Sons.

Eugenio, B. D., & Glass, M. (2004). The kappa statistic: A second look. *Computational linguistics*, *30*(1), 95-101.

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

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

Gisev, N., Bell, J. S., & Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330-338.

Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B., & Shapiro, R. W. (1981). Reliability studies of psychiatric diagnosis: Theory and practice. *Archives of General Psychiatry*, *38*(4), 408-413.

Gwet, K. (2002). Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment Series*, *2*(1), 9.

Gwet, K. L. (2008). Computing inter□rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29-48.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, *8*(1), 23.

Hill, C. E., O'Grady, K. E., & Price, P. (1988). A method for investigating sources of rater bias. *Journal of Counseling Psychology*, *35*(3), 346.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological methods*, *5*(1), 64.

Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, *4*(4), 403.

Hripcsak, G., & Heitjan, D. F. (2002). Measuring agreement in medical informatics reliability studies. *Journal of biomedical informatics*, *35*(2), 99-110.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American journal of health-system pharmacy*, *65*(23), 2276-2284.

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

Lomax, R. G. & Hahs-Vaughn, D. L. (2012). *An introduction to statistical concepts*. Routledge.

Mass Shootings in America. (n.d.). Retrieved from https://library.stanford.edu/projects/mass-shootings-america

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

Meindl, J. N., & Ivy, J. W. (2017). Mass shootings: The role of the media in promoting generalized imitation. *American journal of public health*, *107*(3), 368-370.

Petee, T. A., Padgett, K. G., & York, T. S. (1997). Debunking the stereotype: An examination of mass murder in public places. *Homicide Studies*, *1*(4), 317-337.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge.

Schuster, C., & Smith, D. A. (2002). Indexing systematic rater agreement with a latent-class model. *Psychological Methods*, *7*(3), 384.

Towers, S., Gomez-Lievano, A., Khan, M., Mubayi, A., & Castillo-Chavez, C. (2015). Contagion in mass killings and school shootings. *PLoS one*, *10*(7), e0117259.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Fam med*, *37*(5), 360-363.

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*, *13*(1), 61.

Xu, S., & Lorber, M. F. (2014). Interrater agreement statistics with skewed data: Evaluation of alternatives to Cohen's kappa. *Journal of Consulting and Clinical Psychology*, *82*(6), 1219.

Zec, S., Soriana, N., Comoretto, R., & Baldi, I. (2017). High agreement and high prevalence: The paradox of Cohen's Kappa. *The Open Nursing Journal, 11*, 211-218.

APPENDICES

APPENDIX A: IRB Form

Date:	11/12/2019
Application Number:	ED-19-147
Proposal Title:	A comparison of Cohen's Kappa and Gwet's AC1 with a mass shooting classification index: A study of rater uncertainty
Principal Investigator:	
Ashley Keener Co-Investigator(s):	
Faculty Adviser:	Laura Barnes Project Coordinator:
Research Assistant(s):	
Processed as:	Not Human Subjects Research

Status Recommended by Reviewer(s): Closed

Based on the information provided in this application, the OSU-Stillwater IRB has determined that your project does not qualify as human subject research as defined in 45 CFR 46.102 (d) and (f) and is not subject to oversight by the OSU IRB. Should you have any questions or concerns, please do not hesitate to contact the IRB office at 405-744-3377 or irb@okstate.edu.

Sincerely,
Oklahoma State University IRB

APPENDIX B: Demographic Information

Please Provide the following information:

1. What is your gender identification?

☐ Male ☐ Female ☐ Prefer not to answer

2. What is your ethnicity? Please check all that apply.

☐ Hispanic or Latino
☐ American Indian or Alaska Native
☐ Asian
☐ Black or African American
☐ Native Hawaiian or other Pacific Islander
☐ White/Caucasian
☐ Other
☐ Prefer not to answer

3. What year were you born? Click or tap here to enter text.

4. What is the highest level educational degree you have obtained?

☐ Bachelor's Degree
☐ M.D.
☐ Ph.D.
☐ J.D.
☐ Master's Degree
☐ D.O.
☐ Psy.D.
☐ Prefer not to answer

5. What year did you obtain the highest level educational degree? Click or tap here to enter text.

6. What is your professional area of specialization? Click or tap here to enter text.

7. Have you had any experience with mass shooters? If yes, please provide an explanation.

☐ Yes ☐ No

Explanation (if applicable): Click or tap here to enter text.

APPENDIX C: Example Case Synopsis

Synopsis - 20. McDonald's Restaurant in San Ysidro

On July 18, 1984, James Oliver Huberty, 41, shot and killed 21 people and injured 19 others at a San Diego, California McDonald's before being fatally shot by a SWAT team. Prior to the attack he took his family to a different McDonald's and then for a trip to the San Diego zoo. Looking at the caged animals, Huberty told his wife, 'Society had their chance…', referring to the mental health clinic's failure to return his phone call the previous day. Back at home, he changed into combat gear and told his wife, "I want to kiss you goodbye," and that he was going "to hunt humans".

Around 4pm, Huberty arrived at the McDonalds wearing camouflage trousers and a black T-shirt. He was armed with a semi-automatic rifle, a shotgun and a pistol. He ordered those in the restaurant to lie prone. When an employee picked up a telephone to call the police, the gunman began firing at those on the floor. If anybody moved, he shot them. Later, he fired indiscriminately at adults and children outside the restaurant.

Some of the dead and wounded were children in a McDonald's playground next to the restaurant. Seventeen of the bodies, including the assailant's, were inside the restaurant and four were outside. The windows were riddled with bullets. Victims ranged in age from eight months to 74 years old. The attack ended after an hour and ten minutes when police snipers fired from the roof of an adjacent building, killing Huberty.

Huberty was born in Canton, Ohio, in 1942. He was raised by his grandmother after his parents divorced. At age 3, he contracted polio, which left him with leg paralysis and needing braces.

He had an obsession with guns, shooting the heads off cabbages and running into the woods at night for target practice. He once shot a neighbor's cat. When he visited his father and step mom, whom he didn't get along with, he'd get out of his car with a gun and fire a round of shots to signal his arrival.

He was married and had two daughters. The family lived in middle-class suburb Massillon, in Ohio. Huberty jumped from one job to the next. He even trained as a funeral director and embalmer. His funeral-parlour boss remembered him as a 'loner', with a 'short, quick, temper'. Huberty then found work at a steel plant but, when it shut in 1981, and he lost his job, he ranted to colleagues of his despair. In January 1984, the family moved from Ohio to San Diego. They rented a tiny apartment, and Huberty found work as a security guard. He eventually lost that job, just a few days before the massacre.

He talked obsessively of war, even walked up to a policeman one day and announced he was a 'war criminal', despite having never served in the Forces. His wife suspected he was having a breakdown. The day before the massacre, she apparently urged him to call a mental-health clinic. After the massacre, his wife claimed the man she'd loved 'would never have done this…if he had been in his right mind.' An autopsy confirmed that Huberty wasn't under the influence of alcohol or drugs.

Huberty's only prior run-in with the police was for being drunk and disorderly at a gas station, for which he was fined and paid court costs.

APPENDIX D: Reviewer Instructions

Instructions:

You have been provided with an excel sheet entitled "Reviewer Excel Sheet." It contains the case number associated with each mass shooting incident, a brief description of each case, the five categories related to the Agoracide classification system (i.e., Mental Illness, Collateral Damage, Anger, Commission of a Crime, and Lone Wolf Terrorism), and a "primary category" column.

In each of the Agoracide category columns you have been provided with a dropdown bar that ranges from 1-5. Please review each case and rate 'how much' the case falls into each category using a Likert-type scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). The full Likert-type scale is structured as follows:

1	2	3	4	5
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

There is also a column labeled "Primary Category." Please indicate which Agoracide category is best associated with each case. Again, you have been provided with a dropdown bar that lists each Agoracide category.

If you have any questions, please feel free to email ashley.keener@okstate.edu or jason.beaman@okstate.edu.

Thank you for your time!

VITA

Ashley Keener

Candidate for the Degree of

Doctor of Philosophy

Thesis: A COMPARISON OF COHEN'S KAPPA AND GWET'S AC1 WITH A MASS SHOOTING CLASSIFICATION INDEX: A STUDY OF RATER UNCERTAINTY

Major Field: Educational Psychology

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Educational Psychology: Research, Evaluation, Measurement and Statistics at Oklahoma State University, Stillwater, Oklahoma in May, 2020.

Completed the requirements for the Master of Science in Educational Psychology: Research, Evaluation, Measurement and Statistics at Oklahoma State University, Stillwater, Oklahoma in 2015.

Completed the requirements for the Bachelor of Arts in Psychology at Northeastern State University, Broken Arrow, Oklahoma in 2011.

Experience:

Sr. Research Assistant; Oklahoma State University. Tulsa, OK. (2015 – Present). Center for Health Sciences, Department of Psychiatry and Behavioral Sciences.

Graduate Research Assistant; Oklahoma State University. Tulsa OK. (2014 – 2017). Department of Research, Evaluation, Measurement and Statistics.

Rehabilitation Specialist/Case Manager II; Improving Lives Counseling Services, Tulsa, OK. (2011 – 2015).