Hindawi
Human Behavior and Emerging Technologies
Volume 2022, Article ID 4579178, 18 pages
https://doi.org/10.1155/2022/4579178

Research Article

A Three-Step Reliability Strategy Applied to Police-Worn Body Camera Footage

Daryl G. Kroner[1] **and Joseph A. Schafer**[2]

[1]*Southern Illinois University Carbondale, 1000 Faner Drive, Carbondale IL 62901, USA*
[2]*Saint Louis University, Tegeler Hall, 3550 Lindell Blvd., 319, St. Louis, MO 63103, USA*

Correspondence should be addressed to Daryl G. Kroner; dkroner@siu.edu

Received 11 May 2022; Accepted 16 June 2022; Published 29 July 2022

Academic Editor: Zheng Yan

Strong inferences are drawn from police-worn body camera (BWC) footage, frequently without an assessment of reliability. Unique characteristics of BWC footage (i.e., capturing trends and less frequent behavior, a focal actor (officer) is absent) suggest a specific reliability strategy. A three-step strategy of selecting appropriate reliability indexes, providing salient reliability categories, and ranking the reliability categories was applied to BWC footage. Five interrater agreement (AC_1, $\widehat{\alpha}_K$, $B-P$ coefficient, r_{wg}, ad.m), two interrater consistency ($ICC(1,1), ICC(2,1)$), and three internal consistency (ω_t, ω_h, GLB) indexes were applied to police BWC footage. A focus was to ascertain the upper limits of reliability for BWC footage. Item development and rater training were conducted to optimize rating reliability. Using a within design and confidence intervals, the relatively stronger and weaker reliabilities across the six domains of video completeness, respect (passive, active, discourse), threats, and behavioral stance were assessed. Applied to the admissibility of court evidence, central aspects of video completeness have relatively stronger reliabilities. For research, lower reliabilities have a cost of limited generalizability and ecological validity. Policy recommendations include the usage of a standardized scale with multiple ratings to determine what information should be used in high-stake decision-making based on BWC footage. The three-step strategy integrated the reliability indexes into a single figure to reflect a reliability summary of each component of BWC footage. Weighted rankings found the Overall Audio Quality (-4.9) and Empathy (-4.9) items to have the weakest reliabilities and the Clarification (5.1) and Physical Resistance (4.9) items to have the strongest reliabilities.

1. Introduction

Video recordings of public events are easy to gather and share, resulting in a convenient record. As a record, the visual components of video footage make the information easier to process for the observer (no reading) and for the researcher (no physical observation). For police BWC footage, researchers can study inherently complex interactions that include rare and difficult-too-observe aspects of police events and decision-making [1]. Inferences from police BWC footage, whether general observation or research-based, often assume the trustworthiness of video footage content. Before we can formulate inferences, engage in subsequent analyses, draw conclusions, and make decisions, some level of video footage trustworthiness is necessary. This paper outlines and applies a strategy to evaluate the trustworthiness (i.e., reliability) of BWC footage.

The reliability of BWC footage can be evaluated via different methods. Viewing an event that is similar to past correct inferences, one can logically infer some level of reliability to a current event. Ontologically, a trustworthy person or system (i.e., media) can say that inferences are warranted. Within an epistemology framework, statistical evidence can be used to examine evidence (focus of the present study). Statistical reliability is a tool of social science. One drawback of statistical reliability is that it cannot directly evaluate the footage, requiring inferences from the content to estimate reliability. For this method of reliability to have benefit, the inferred content needs to have importance for making video inferences. The present study's scales

cover officer-citizen interaction areas relevant for legal/court requirements and contributors to a potential officer-citizen conflict (see Supplementary Materials Part A for an overview and description of scales (available here)). Statistical reliability places constraints on levels of valid inferences [2]. To better isolate which BWC content areas could be used for valid inferences, the current study was designed for the possibility of maximizing reliabilities.

1.1. Uniqueness of BWC Footage. Prior to the common usage of BWC footage, the gathering of officer-citizen interactions was a challenging endeavor. Researchers typically had to rely on resource-intensive observation strategies, such as systematic observation [3, 4]. These approaches constrained sampling methods [5] and created reactivity concerns [6, 7]. An observer would have to be present for many extraneous interactions, especially if the target observation was a less frequent behavior. BWC footage has the methodological advantage to observe many interactions, capturing trends and less frequent behavior without requiring a direct observer. Fixed-placement cameras also have these advantages [8], but BWC footage has the unique characteristic of an absence of solely outside actors (an officer is wearing the camera), which changes observational perceptions of emotionally charged events. These unique characteristics result in greater difficulties in formulating inferences, namely, sufficient objectiveness to assess intentionality [9]. These unique characteristics of BWC footage [10] warrant a reliability review, focused on understanding reliabilities associated with BWC footage content.

Inferences related to officer-citizen interactions can be informed by one's prior internal experiences [11, 12], which in and of itself, has a degree of internal and experiential reliability [13, 14]. BWC footage, by definition, lacks this internalness component of reliability. The external nature (we can observe) of BWC footage affords the opportunity to estimate the trustworthiness of BWC footage via reliability statistics and designs [15]. The presence of reliable data, though, does not automatically ensure the measurement validity [16, 17], but greater agreement can result in stronger validities [18, 19]. Given the intersect of a general under representation reliability index coverage [11, 20–23] and the strong inferences drawn from BWC footage, the present paper proposes a strategy of integrating a representation of reliability indexes. Interrater and internal consistency reliability indexes are applied to the structured content rating *data* of BWC footage [24]. The uniqueness of BWC data and its structure guide the process for which reliability indexes are used and which indexes have greater value.

1.2. Reliability Indexes. Indexes of reliability can fall into two broad methodological categories with their respective subcategories: interrater and internal consistency. The interrater approach uses multiple raters using the same method to rate the same observations [17, 19, 25]. The emphases with interrater statistics are agreement and consistency among the raters. The internal consistency approach focuses on grouped observations (i.e., items), with each observation having multiple data points. The emphasis with internal consistency is the assessment of scale total variance and is not directly related to the raters.

Interrater designs fall into two basic categories: interrater agreement and interrater reliability [26, 27]. Conceptually, interrater agreement assesses the degree of agreement or consensus among raters, indicating the degree that raters are interchangeable. Interrater reliability focuses on the consistency and stability (i.e., similarity in rank and profile similarity) among raters (labeled "inter-rater consistency" for the current paper). The comparison of interrater agreement and interrater consistency indexes highlights four possible high/low combinations [28]. Using a graph with rater A's ratings on the y axis and rater B's ratings on the x axis, the data points falling on the diagonal line would indicate a perfect relationship (see Figure 1). So, ratings close to the diagonal line would indicate high interrater agreement and high interrater consistency. Practically, this would involve raters having the same or very similar ratings in most cases. Ratings close to a straight line, but not near the diagonal would indicate low interrater agreement, but high interrater consistency. Practically, this would involve raters having similar patterns of ratings, but no or few ratings with the same assigned rated number. Ratings clustered at the center of the graph would indicate high interrater agreement but low interrater consistency. This would involve most cases having similar assigned rater numbers, indicating a restriction of range. Ratings scattered throughout the graph would indicate both low interrater agreement and low interrater consistency. Practically, this would involve a dissimilar pattern of ratings and dissimilar assigned rated numbers.

Interrater agreement is the ratio of the difference between chance and obtained agreement to the maximum nonchance agreement. In terms of application, interrater agreement indexes attempt to adjust for chance or expected agreement. Five agreement indexes are used in the present study (Gwet's AC_1, Krippendorff's α, Brennan-Prediger's ($B-P$) coefficient, James et al.'s within-group agreement (r_{wg}), Burke et al.'s average deviation for mean (ad.m)). Interrater agreement indexes are unable to assess the degree of alignment (i.e., consistency) among the raters, which is accomplished with interrater consistency indexes.

Interrater consistency is the ratio of the difference between the data variance that is due to the rated targets and the data variance that is due to the raters (i.e., signal to noise ratio). As such, interrater consistency indexes require that the variable be in a continuous format. Different than the agreement indexes, interrater consistency measures cannot be computed for prefect agreement. The application of interrater reliability involves comparing the portion of variance that is associated with the target to the portion of variance that is associated with the raters. Greater rater variance than target variance would indicate low reliability. Intraclass correlation is calculated for the current study, with intraclass referring to an inability to distinguish among the class of raters [19]. Interrater reliability indexes have the benefit of reflecting different sources of variation with data than is able to be done with agreement indexes [2]. The ICC(2, 1) formula was used for the item data and the ICC(

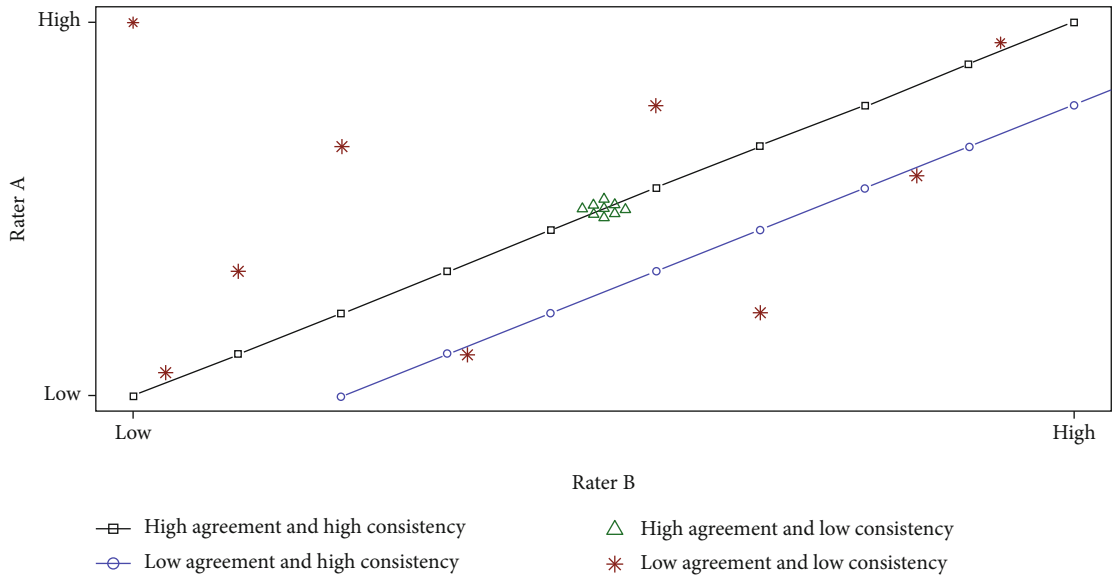

Figure 1: Four possible high/low interrater combinations.

1, 1) formula for the item-deletion approach (described in Method). [29] demonstrated that the ICC(2, 1) formula could be used with nominal data.

Internal consistency measures the ratio of observed scores to the associated true scores, which assesses the error measurement in a scale score. True scores refer to a scale score in a perfect world, or administrating a test under all possible circumstances, obtaining an unbiased mean of all observed scores [16, 30]. Through the use of sample standard deviations, a true score is estimated. The level of trustworthiness for internal consistency is the proportion of true-score variance explained by the observed scores. In terms of application, internal consistency determines the degree that the measures will allow for similar results, given the multiple opportunities for variation to occur. This provides a framework for test score generalization to occur. Three internal consistency indexes are used for the present study (omega total, $[\omega_t]$, omega Hierarchical, $[\omega_h]$, greatest lower bound [GLB]). Similar to the interrater reliability indexes, internal consistency indexes focus is on scale total variance. To facilitate an item focus, the scale index calculation is made with scale items sequentially removed (i.e., increased or lowered internal consistency based upon one scale item removed).

The above logical and computational differences among the three reliability categories are reinforced by simulation studies. Under various levels of rating score agreement, interrater agreement indexes are more strongly correlated with each other, and indexes of interrater consistency are more strongly correlated with each other [2]. In addition, the consistency indexes were more sensitive to the magnitude of score differences, which would be conceptually expected.

1.3. Three-Step Strategy for Choosing and Applying Reliability Estimates. Multiple indexes of reliability are necessary to present a comprehensive picture of rated contents' trustworthiness [19, 31]. Given that many reliability indexes are available, a three-step analytic strategy to apportion reliability indexes to an overall evaluative framework is proposed. First is the exclusion of reliable indexes, although common, that have demonstrated statistical difficulties. Second is the assignment of categorical importance for agreement, interrater consistency, and internal consistency reliability indexes. Third is the assignment of weights to each of the three categories of reliability indexes. The use of a three-step methodological process has also been applied to medical imagery (PET, fMRI, CT) [32].

1.3.1. Excluded Reliability Indexes (First Step). Percent agreement, kappa coefficient, and Cronbach's alpha have been much used indexes of statistical reliability. Yet, each of these have statistical difficulties for which other indexes have demonstrated improvements. Percent agreement does not account for the possibility of chance agreement, resulting in an overly liberal index [2, 22]. Kappa coefficient is dependent on marginal distributions, which introduces "bias" of a rater [33]. Cronbach's alpha makes assumptions of tau equivalence and uncorrelated errors that are rarely met [34].

1.3.2. Ranking of Importance of the Three Categories of Reliability Indexes (Second Step). The second step ranks the importance of the three categories of reliability indexes. The interrater agreement indexes are ranked first, followed by interrater consistency, and then internal consistency indexes. There are four reasons for this order. First, the areas rated in the BWC footage cover a wide breadth of information, including environmental conditions, quality of video, presenting issues, reactive behaviors, affect, and the presence of objects. This breadth goes beyond solely examining a narrowly defined construct (i.e., rage). Second, the level of inference needed to make the ratings varies. Some content involves minimal inference (i.e., presence of light) and some content involves greater inference to make a rating (i.e., empathy), for which interrater agreement and interrater consistency indexes are better resourced. Thus, evaluating

individual items will better reflect the basic differences among the items. Third, agreement coefficients, which reflect an absolute measurement rating among these areas, will have more salience (i.e., content present or not) in evaluating the trustworthiness of targeted content. This better mirrors the value of consensus before making inferences and drawing conclusions of BWC footage, which is contrasted to rank similarity or pattern similarity of the rated content. Similarly, researchers examining communication content analyzes place a high premium on agreement, emphasizing this type of reliability as a precondition for valid inferences [22]. Fourth, agreement coefficients emphasize the similarity amongraters, as compared to the consistency of the content. Knowing the interchangeability of raters (i.e., independent replications) of the judgments on BWC footage has stronger importance for drawing BWC footage conclusions than the consistency of the construct being rated. The greater the commonality in the absolute measurement rating of key components contributing to BWC footage ratings will suggest that these detected differences are possible and likely to be found by others. This occurrence of independent replications in the ratings increases the trust that these data can have similarity among information stakeholders in an applied setting [15].

The greater emphasis on interrater agreement and interrater consistency will not preclude incorporating internal consistency indexes, as interrater reliability indexes will confound measurement error with other sources of variability koo 2016, [17]. Drawing upon the idea of evidence synthesis, which allows for various kinds of evidence to be integrated within a higher order structure [2], internal consistency will be weighted less in summarizing the reliability results.

1.3.3. Assignment of Weights for the Three Types of Reliability Indexes (Third Step). In order to more finely reflect the importance of the three types of reliability indexes and to allow for multiple reliability indexes within each type of reliability, weights are assigned for each of the three types (categories) of reliability indexes. The agreement index category is weighted 3 : 1 against the internal consistency index category. The agreement index category is weighted 2 : 1 against the interrater consistency index category. The interrater consistency index category is weighted 2 : 1 against the internal consistency index category. Viewed in terms of points, the agreement index category coefficients have the potential of 3 points, the interrater consistency index category has the potential of 1.5 points, and the internal consistency index category has 1 point. From the weights, an overall summary of the index coefficients can be obtained.

1.4. Present Study. The current paper has three goals. First, a three-step strategy is proposed to assess the reliability of BWC footage. Second, in the context of measuring the complexities of officer-citizen encounters, we seek to ascertain the BWC footage content that are capable of the upper limits of reliability. This involved the development of an optimal rating measure (with detailed guide) and ratings completed by adequately trained raters. Technology-based data with unstructured content will have limited utility [35]. Basic reliabilities across key content areas will help inform which of these components ought to be used to formulate inferences, engage in subsequent analyses, draw conclusions, and make decisions upon BWC footage. Of note, reliability issues are absent in critical BWC footage reviews [36–39]. The third goal focused on an application, assessing differences between reliabilities (via confidence intervals), and not according to benchmark criteria (i.e., .70). Thus, our overall goal was to understand the levels of reliability associated with this data set [24].

2. Method

Attempting to maximize reliability was accomplished through the two tasks of item development strategies and rater training. In contrast to measuring general interpersonal interactions (i.e., "courteous"), the current measure was patterned after the direct behavior rating approach, which is strongly context-specific [40].

2.1. Item Development. The development of the items considered the type of scaling of the items and the use of direct observation rating strategies. The determination of the response scale for each item took into account the content being evaluated in BWC footage (See Supplementary Materials, Part A for scale content rationale). For example, the use of a 1-5 rating scale was tied to the item content (i.e., tone: 1 = hostile, 5 = warm) and promotes less error variance in targeted ratings [41]. Also, each rating level was viewed appropriate if each of the levels had an opportunity to be endorsed. If the responses would be bounded (i.e., limited chance of a normal distribution or difference between levels not equal), a more limited rating scale was chosen. In addition to distribution concerns, this strategy helps to reduce measurement error [42]. If rating responses would not approximate equal intervals between the scale values, then a categorical approach was used. For example, measuring video obstruction did not lend itself to a numeric rating scale. Either the BWC was properly situated on the officer or not. Thus, this item was rated with two levels of "No" or "Yes." Given that meeting an item threshold (i.e., interagreement) was of greater importance than construct measurement, the majority of the items (notably outward behaviors, see Supplementary Materials, Table A.1) incorporated a nominal level of measurement.

The focus of direct observation ratings involved three guidelines in the development of the items [40]. First, items reflected a current, specific time-frame (present study used 30 min intervals of the same event), minimizing retrospective judgments. Second, items had straightforward descriptions with the anticipation that field supervisors would be rating these items. Third, a substantial number of items were physical and social indicators, reducing the number of inferential judgments [43].

2.1.1. Item Development Process. The development of the items occurred throughout biweekly meetings over 12 months (3 semesters) among the two authors and two graduate students [44]. Goals included addressing: (a) scaling raw BWC footage data [45] and (b) development of items

that would allow for optimal reliability by minimally trained raters. Much of the discussion centered on resolving disagreement through item content clarifications and scoring revisions. Four footage cases were used to develop and pilot the items. These developmental and pilot footage cases were not included in the current study. Also, these two graduate students were not a part of the study's raters.

2.1.2. Calculation of Scale Scores. The scaling of each item for the scale scores varied and was measured according to nominal, ordinal, and interval levels of measurement. In addition, the interval items had different number of response levels (e.g., 3 to 6 levels). The use of different item response levels and the associated distributional qualities (i.e., skewness) differentially impact internal consistency [46]. In order to reduce the impact of these issues and to facilitate using items to calculate a scale score, items were converted to a binary response. [47] presented arguments of how binary items can appropriately indicate the internal consistency of scale content. For the current study's interval items, frequency mid-points were used to create the binary item. For the nominal and ordinal items, approximately 90% of the responses fell into two response levels. The 10% typically involved a "Do not Know" response, which was coded as a "No" or "Not Present" response.

2.2. Rater Training, Cases, and Study Design. Four raters were graduate students (different than for scale development) who received the same training protocol. The training, including practice ratings, was 30 hours. Thus, from a reliability design perspective, the raters, because of similarity in training, can be assumed to function as parallel measures. All the cases with interrater ratings ($n = 13$) were rated by either two ($n = 8$) or three ($n = 5$) raters. The interrated cases were sequentially chosen (every 7th case) from the total dataset ($n = 99$). Excluding the training cases (not included in the rated dataset), raters were blind to each others' ratings.

The footage cases were chosen according to the potential of the footage case containing behavioral health issues. The participating agency uses a common police Computer Aided Dispatch (CAD) system to manage citizen calls for service and officer activities. The CAD system tracks officer-citizen encounters at the event level [3], with classifications assigned by police communications personnel based on input from officers after a call for service is resolved. This includes classifying a officer-citizen encounter as primarily concerning a behavioral health issue. Encounters recorded in the CAD system from (June, 2015 to May, 2017) were selected for inclusion in this project. These encounters were further filtered based on whether body camera footage was available, as the agency was in the process of a multiyear staggered deployment of BWC units. Some encounters did not involve officers equipped with BWC units, resulting in a sample of 99 encounters.

2.3. Analytic Plan

2.3.1. Interrater and Internal Consistency Indexes. For the following five agreement indexes, interrater agreement is the ratio of the difference between chance and obtained agreement to the maximum nonchance agreement, with the main difference among the indexes being how chance-agreement is calculated. Except for the ad.m index (reversed), a coefficient close to 1 reflects near-perfect agreement, and a coefficient near 0 reflects agreement that can be expected by chance (not necessarily no agreement). Gwet's AC_1 (B.1, Supplementary Materials), Krippendorff's alpha ($\widehat{\alpha}_K$; B.2), and Brennan-Prediger ($B-P$; B.2) coefficients can be calculated on ordinal data with multiple raters. For items with multiple response levels, James et al.'s within-group agreement statistic (r_{wg}; B.4) uses a rectangular distribution (i.e., symmetrical probability distribution between certain parameters) to handle chance agreement [48]. The average deviation statistic (ad.m; B.5) assesses agreement through the average of the absolute differences between each score and the overall mean.

Interrater consistency indexes measure profile similarity, which is different than the extent to which raters give interchangeable ratings. In terms of application, the interrater consistency index provides a consistency, conformity, or repeatability of a measure. The $\mathrm{ICC}(2,1)$ formula (B.6, Supplementary Materials) is applied to nominal/ordinal data [29] and the $\mathrm{ICC}(1,1)$ (B.7) to continuous data.

Omega indexes require a scale of items and are based on a factor analytic model. Omegas are assessed by a ratio of item variability that is explained by the total variance of all the items. For omega total (ω_t) [49], the factor model is transformed by a Schmid-Leiman rotation [50], which rotates the factor solution according to a bifactor model with one general factor and several smaller factors (B.8, Supplementary Materials). Omega hierarchical (ω_h; B.9) is similar to ω_t, but it only includes the contributions of the general factor, with group and item loadings being omitted. Greatest lower bound (GLB; B.10), developed within classical test theory, incorporates the two components of the following: (a) the sum of the interitem covariance matrix for true scores and (b) the sum of the interitem covariance matrix for the error term [51].

Calculations for the AC_1, $\widehat{\alpha}_K$, $B-P$, r_{wg}, ad.m, and $\mathrm{ICC}(1,1)$ indexes were computed in the **R** package multilevel [52]. The $\mathrm{ICC}(2,1)$, ω_t, ω_h, and GLB indexes were computed in the **R** package psych [49]. Reliability formulas with descriptions are presented in Supplementary Materials, Part B.

2.3.2. Item Deletion Approach. To assist in assessing the individual items, an item deletion approach was used with the interrater correlation statistics, as the r_{wg}, ad.m, and $\mathrm{ICC}(1,1)$ indexes can be used with both single items and scale scores. This approach compared the individual rated items to the reliability of the items comprising the total scale score. If a single item is removed, then the total scale coefficient is assessed to determine if there was an increase or decrease with the item removed [53]. A decrease would indicate that the item is contributing to the scale's reliability, and an increase coefficient would indicate that the item is reducing the scale's reliability. Given that this approach assesses the items' performance through the total score (crossitem and

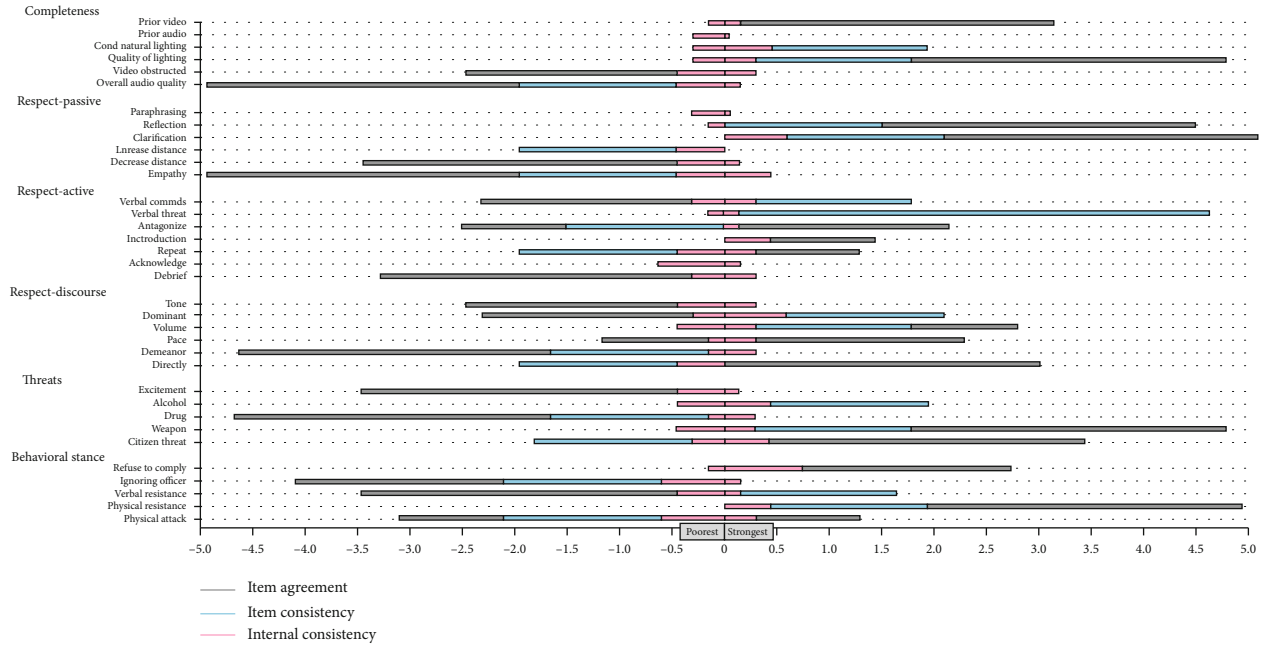

Figure 2:

crossrater variance assumptions are made), these results are deemed to be of less importance in assessing the overall trustworthiness of the items.

2.3.3. Calculations of Assigned Weights for the Three Types of Reliability Indexes. To accommodate the 3 : 1 and 2 : 1 weights, each of the interrater agreement coefficients were weighted "1," given that there were three coefficients used in the current study. In addition to using the coefficient confidence intervals to assess if the coefficients within a scale are different (i.e., higher or lower), to facilitate a visual representation of the within scale differences (Figure 2), the two higher interrater agreement coefficients received a weight of "1," and the two lower coefficients received a weight of "-1." The interrater consistency (ICC(1, 1)) coefficients were weighted "1.5," given that there was one coefficient for the consistency category used in the current study. The coefficients for the item deletion approach were weighted "0.15," given that there were six coefficients for the consistency category used in the current study.

3. Results

3.1. Interrater Agreement. Table 1 reports three interrater agreement indexes (Gwet's AC_1, $\widehat{\alpha}_K$, $B - P$ coefficient) on the individual rated items. For the Completeness scale, the Prior Video and Quality of Lighting items had the strongest agreement. The level of agreement for video obstruction was moderate, neither the strongest nor the weakest, suggesting some relative agreement if video interference was occurring. Overall audio quality had the poorest agreement, suggesting that a minimum standard for understanding audio may be difficult to assess. For the Respect-Passive scale, the Reflection and Clarification items had the strongest agreement. The Decreased Distance and Empathy items had the poorest agreement. For the Respect-Active scale, the Verbal Threat and Antagonize items had the strongest agreement. The Verbal Commands and Debrief items had the poorest agreement. For the Respect-Discourse scale, the Directly and Volume items had the strongest agreement. The Tone and Dominant items had the poorest agreement. For the Threats scale, the Weapons and Citizen Threat items had the strongest agreement. The Excitement and Drug items had the lowest agreement. For the Behavioral Stance scale, the Physical Resistance and Physical Attack items had the strongest agreement. The Ignoring Officer and Verbal Resistance items had the poorest agreement.

Of note, the strongest and poorest coefficients were outside the 95% confidence intervals for the three indexes (Gwet's AC_1, $\widehat{\alpha}_K$, $B - P$ coefficient) across five scales (Completeness, Respect-Passive, Respect-Active, Respect-Discourse, and Threats). An exception were the Gwet's AC_1 and $B - P$ coefficient indexes for the Behavioral Stance scale, for which the item coefficients were within each other's 95% confidence intervals.

3.2. Interrater Consistency. The ICC(2, 1) consistency results are presented in Table 2. For the Completeness scale, the Quality of Lighting and Natural Lighting items had the strongest consistency. Overall Audio Quality had the poorest consistency, suggesting that the conceptual understanding for audio may be difficult for raters to align their ratings. For the Respect-Passive scale, the Reflection and Clarification items had the strongest consistency. The Empathy and Increase Distance items had the poorest consistency. For the Respect-Active scale, the Verbal Threat and Verbal Commands items had the strongest consistency. The Antagonize and Repeat items had the poorest consistency. For the Respect-Discourse scale, the Volume and

Table 1: Item calculations for Gwet's AC_1, Krippendorff's alpha, and Brennan-Prediger's agreement indexes.

Item	AC_1	95% CI	$\widehat{\alpha}_K$	95% CI	$B-P$	95% CI
Completeness						
Prior Video	0.841	[0.589, 1.000]	0.655	[0.213, 1.000]	0.808	[0.517, 1.000]
Prior Audio	0.768	[0.467, 1.000]	0.189	[-0.157, 0.535]	0.692	[0.517, 1.000]
Cond Natural Lighting	0.711	[0.280, 1.000]	0.490	[-0.033, 1.000]	0.608	[0.088, 1.000]
Quality of Lighting	0.812	[0.559, 1.000]	0.550	[0.154, 0.947]	0.725	[0.434, 1.000]
Video Obstructed	0.635	[0.197, 1.000]	0.348	[-0.098, 0.794]	0.538	[0.092, 0.985]
Overall Audio Quality	0.366	[-0.192, 0.924]	0.022	[-0.533, 0.577]	0.231	[-0.311, 0.772]
Respect-Passive						
Paraphrasing	0.684	[0.420, 0.948]	0.300	[-0.333, 0.934]	0.548	[0.237, 0.860]
Reflection	0.825	[0.626, 1.000]	0.448	[-0.204, 1.000]	0.713	[0.430, 0.995]
Clarification	0.723	[0.460, 0.988]	0.519	[0.095, 0.944]	0.635	[0.369, 0.938]
Increase Distance	0.670	[0.165, 1.000]	0.160	[-0.226, 0.546]	0.556	[0.031, 1.000]
Decrease Distance	0.100	[-0.501, 0.701]	0.114	[-0.452, 0.679]	0.091	[-0.501, 0.683]
Empathy	0.142	[-0.258, 0.542]	0.100	[-0.255, 0.455]	0.115	[-0.278, 0.508]
Respect-Active						
Verbal Commds	0.163	[-0.476, 0.802]	0.182	[-0.365, 0.729]	0.128	[-0.474, 0.730]
Verbal Threat	1.000	[1.000, 1.000]	1.000	[1.000, 1.000]	1.000	[1.000, 1.000]
Antagonize	0.821	[0.584, 1.000]	-0.071	[-0.201, 0.058]	0.692	[0.339, 1.000]
Introduction	0.601	[0.232, 0.970]	0.375	[-0.008, 0.758]	0.542	[0.164, 0.920]
Repeat	0.282	[-0.143, 0.708]	0.083	[-0.261, 0.426]	0.231	[-0.171, 0.633]
Acknowledge	0.360	[-0.074, 0.795]	0.216	[-0.201, 0.633]	0.308	[-0.111, 0.727]
Debrief	-0.026	[-0.346, 0.294]	-0.126	[-0.367, 0.115]	-0.060	[-0.361, 0.241]
Respect-Discourse						
Tone	0.576	[0.294, 0.904]	0.327	[-0.097, 0.752]	0.308	[0.180, 0.836]
Dominant	0.670	[0.346, 0.995]	0.483	[0.086, 0.879]	0.625	[0.305, 0.945]
Volume	0.855	[0.689, 1.000]	0.615	[-0.109, 0.530]	0.795	[0.596, 0.994]
Pace	0.891	[0.739, 1.000]	0.210	[-0.109, 0.530]	0.798	[0.560, 1.000]
Demeanor	0.785	[0.488, 1.000]	-0.071	[-0.205, 0.062]	0.641	[0.218, 1.000]
Directly	0.945	[0.815, 1.000]	0.651	[0.588, 0.714]	0.923	[0.755, 1.000]
Threats						
Excitement	0.405	[0.079, 0.731]	0.259	[-0.017, 0.534]	0.385	[0.063, 0.706]
Alcohol	0.934	[0.776, 1.000]	0.722	[0.259, 1.000]	0.897	[0.674, 1.000]
Drug	0.854	[0.606, 1.000]	-0.034	[-0.143, 0.074]	0.744	[0.357, 1.000]
Weapon	1.000	[1.000, 1.000]	1.000	[1.000, 1.000]	0.897	[0.674, 1.000]
Citizen Threat	1.000	[1.000, 1.000]	1.000	[1.000, 1.000]	1.000	[1.000, 1.000]
Behavioral Stance						
Refuse to Comply	0.828	[0.563, 1.000]	0.764	[0.426, 1.000]	0.808	[0.517, 1.000]
Ignoring Officer	0.754	[0.370, 1.000]	0.748	[0.367, 1.000]	0.744	[0.357, 1.000]
Verbal Resistance	0.821	[0.538, 1.000]	0.727	[0.330, 1.000]	0.795	[0.492, 1.000]
Physical Resistance	0.905	[0.691, 1.000]	0.872	[0.592, 1.000]	0.897	[0.674, 1.000]
Physical Attack	0.886	[0.702, 1.000]	-0.034	[-0.141, 0.072]	0.795	[0.492, 1.000]

Note. AC_1: Gwet's AC_1; $\widehat{\alpha}_K$: Krippendorff's alpha; B – P: Brennan-Prediger's coefficient. Ordinal and interval scales (see Table A.1 in Supplementary Materials) computed with weighted formulas.

Dominant items had the strongest consistency. The Tone and Demeanor and Directly items had the poorest consistency. For the Threats scale, the Weapons and Alcohol items had the strongest consistency. The Drug and Citizen Threat items had the lowest consistency. For the Behavioral Stance scale, the Physical Resistance and Verbal Resistance items had the strongest consistency. The Physical Attack and Ignoring Officer items had the poorest consistency. Of note, the strongest and poorest coefficients were outside the 95% confidence intervals for the ICC(2, 1) index across the six scales (Completeness, Respect-Passive, Respect-Active, Respect-Discourse, Threats, and Behavioral Stance).

Table 2: Item calculations for intraclass correlation index (ICC (2, 1)).

Item	ICC	95% CI
Completeness		
Prior Video	0.770	[0.538, 0.905]
Prior Audio	0.465	[-0.029, 0.774]
Cond Natural Lighting	0.862	[0.725, 0.943]
Quality of Lighting	0.879	[0.757, 0.950]
Video Obstructed	0.722	[0.441, 0.884]
Overall Audio Quality	0.322	[-0.117, 0.679]
Respect-Passive		
Paraphrasing	0.830	[0.623, 0.932]
Reflection	0.899	[0.777, 0.960]
Clarification	0.888	[0.775, 0.954]
Increase Distance	0.480	[-0.009, 0.782]
Decrease Distance	0.631	[0.300, 0.840]
Empathy	0.000	[-0.249, 0.361]
Respect-Active		
Verbal Commds	0.611	[0.267, 0.831]
Verbal Threat	1.000	[1.000, 1.000]
Antagonize	0.000	[-0.351, 0.424]
Introduction	0.016	[-0.311, 0.423]
Repeat	0.000	[-0.206, 0.326]
Acknowledge	0.015	[-0.082, 0.215]
Debrief	0.000	[-0.075, 0.165]
Respect-Discourse		
Tone	0.691	[0.378, 0.873]
Dominant	0.748	[0.494, 0.896]
Volume	0.911	[0.820, 0.963]
Pace	0.560	[0.114, 0.819]
Demeanor	0.000	[-1.010, 0.587]
Directly	0.000	[-0.906, 0.574]
Threats		
Excitement	0.764	[0.413, 0.909]
Alcohol	0.900	[0.800, 0.959]
Drug	0.000	[-0.949, 0.580]
Weapon	1.000	[1.000, 1.000]
Citizen Threat	0.740	[0.477, 0.893]
Behavioral Stance		
Refuse to Comply	0.971	[0.942, 0.988]
Ignoring Officer	0.560	[0.114, 0.819]
Verbal Resistance	0.866	[0.707, 0.946]
Physical Resistance	0.959	[0.917, 0.983]
Physical Attack	0.000	[-1.033, 0.590]

Note: ICC: intraclass correlation (ICC(2, 1)).

3.3. Item Reliability according to Item Deletion Approach

3.3.1. Interrater Agreement. Incorporating the item removal method, the strongest item contributing to the scale's level of agreement is indicated by the lowest coefficient (ad.m is reversed; Table 3). With the $r_{wg(i)}$ agreement index, Prior Audio, Empathy, Debrief, Pace, Alcohol, and Physical Attack items had the strongest contribution to their respective scales' level of agreement. With the $r_{wg(i)}$ agreement index, Cond Natural Lighting, Reflection, Acknowledge, Tone, Excitement, and Ignoring Officer items had the poorest contribution to their respective scales' level of agreement.

With the ad.m agreement index, Overall Audio Quality, Clarification, Debrief, Pace, Alcohol, and Physical Attack (higher CIs) items had the strongest contribution to their respective scales' level of agreement. With the ad.m agreement index, Cond Natural Lighting, Decrease Distance, Acknowledge, Tone, Excitement, and Ignoring Officer items had the poorest contribution to their respective scales' level of agreement.

3.3.2. Interrater Consistency. Table 3 contains the interrater consistency item results. With the ICC consistency index, Cond Natural Lighting, Clarification, Verbal Threat, Dominant, Citizen Threat, and Physical Resistance items had the strongest contribution to their respective scales' level of consistency. With the ICC consistency index, Prior Video, Increase Distance, Repeat, Volume, Alcohol (higher CIs), and Ignoring Officer items had the poorest contribution to their respective scales' level of consistency.

3.3.3. Internal Consistency. Table 4 contains the internal consistency results using the item removal method. As with the interrater agreement and consistency indexes, the strongest item contributing to the scale's level of consistency is indicated by the lowest coefficient. With the ω_t internal consistency index, Prior Audio, Paraphrasing, Acknowledge, Dominant, Drug, and Refuse to Comply items had the strongest contribution to their respective scales' level of internal consistency. With the ω_t internal consistency index, Overall Audio Quality, Increase Distance, Debrief, Directly, Alcohol, and Physical Attack items had the poorest contribution to their respective scales' level of internal consistency.

With the ω_h internal consistency index, Prior Audio, Decrease Distance, Repeat, Dominant, Citizen Threat, and Physical Attack items had the strongest contribution to their respective scales' level of internal consistency. With the ω_h internal consistency index, Overall Audio Quality, Increase Distance, Verbal Threat, Volume, Alcohol, and Verbal Resistance items had the poorest contribution to their respective scales' level of internal consistency.

With the glb internal consistency index, Cond Natural Lighting, Paraphrasing, Repeat, Dominant, Excitement, and Refuse to Comply items had the strongest contribution to their respective scales' level of internal consistency. With the glb internal consistency index, Overall Audio Quality, Empathy, Debrief, Directly, Weapon, and Physical Attack items had the poorest contribution to their respective scales' level of internal consistency. Tables 1–4 were generated in the **R** package xtable [54].

3.4. Summaries of Index Coefficients. Figure 2 provides a visual summary of Tables 1–4 (summarized in Supplementary Materials, Table C.1) using the strongest and poorest reliability coefficients within each scale. Figure 2 provides

Table 3: James et al.'s agreement for multi-item, Burke et al.'s average deviation for mean, and intraclass correlation indexes for total scale (bold) and if scale item is removed.

Item	$r_{wg(i)}$	95% CI	ad.m	95% CI	ICC	95% CI
Completeness						
Scale Coefficient	**0.862**	[0.69, 1.039]	**0.321**	[-0.004, 0.645]	**0.712**	[0.414, 0.989]
Prior Video	0.862	[0.69, 1.039]	0.303	[0.009, 0.598]	0.764	[0.455, 0.989]
Prior Audio	0.853	[0.68, 1.028]	0.342	[0.051, 0.633]	0.751	[0.448, 0.987]
Cond Natural Lighting	0.913	[0.82, 1.008]	0.252	[0.031, 0.473]	0.558	[0.334, 0.971]
Quality of Lighting	0.888	[0.72, 1.052]	0.286	[-0.017, 0.590]	0.634	[0.238, 0.990]
Video Obstructed	0.859	[0.68, 1.035]	0.368	[0.018, 0.717]	0.709	[0.423, 0.979]
Overall Audio Quality	0.865	[0.71, 1.026]	0.393	[0.105, 0.682]	0.686	[0.423, 0.972]
Respect-Passive						
Scale Coefficient	**0.427**	[0.16, 0.693]	**0.868**	[0.503, 1.232]	**0.212**	[-0.006, 0.817]
Paraphrasing	0.563	[0.31, 0.819]	0.709	[0.346, 1.073]	0.165	[-0.005, 0.825]
Reflection	0.565	[0.33, 0.797]	0.726	[0.406, 1.047]	0.185	[-0.014, 0.816]
Clarification	0.504	[0.25, 0.757]	0.799	[0.479, 1.119]	0.139	[-0.023, 0.798]
Increase Distance	0.530	[0.29, 0.769]	0.735	[0.417, 1.053]	0.363	[0.097, 0.852]
Decrease Distance	0.556	[0.31, 0.798]	0.705	[0.358, 1.052]	0.165	[-0.069, 0.831]
Empathy	0.501	[0.26, 0.742]	0.774	[0.485, 1.062]	0.207	[0.019, 0.811]
Respect-Active						
Scale Coefficient	**0.756**	[0.57, 0.945]	**0.376**	[0.119, 0.633]	**0.717**	[0.460, 0.979]
Verbal Commds	0.724	[0.55, 0.902]	0.449	[0.212, 0.685]	0.606	[0.312, 0.948]
Verbal Threat	0.756	[0.57, 0.945]	0.376	[0.119, 0.633]	0.569	[0.236, 0.948]
Antagonize	0.769	[0.59, 0.948]	0.393	[0.150, 0.636]	0.707	[0.435, 0.948]
Introduction	0.724	[0.53, 0.916]	0.415	[0.179, 0.650]	0.575	[0.292, 0.948]
Repeat	0.782	[0.61, 0.951]	0.359	[0.153, 0.565]	0.738	[0.518, 0.948]
Acknowledge	0.788	[0.61, 0.972]	0.329	[0.114, 0.544]	0.734	[0.539, 0.948]
Debrief	0.673	[0.46, 0.890]	0.466	[0.233, 0.699]	0.727	[0.551, 0.944]
Respect-Discourse						
Scale Coefficient	**0.687**	[0.47, 0.905]	**0.338**	[0.118, 0.557]	**0.161**	[-0.007, 0.825]
Tone	0.862	[0.69, 1.032]	0.171	[-0.003, 0.345]	0.021	[-0.209, 0.875]
Dominant	0.769	[0.60, 0.943]	0.286	[0.107, 0.466]	-0.181	[-0.204, 0.765]
Volume	0.703	[0.46, 0.941]	0.333	[0.077, 0.590]	0.452	[0.224, 0.915]
Pace	0.672	[0.44, 0.905]	0.355	[0.122, 0.587]	0.031	[-0.162, 0.828]
Demeanor	0.677	[0.44, 0.910]	0.308	[0.087, 0.528]	0.161	[0.020, 0.862]
Directly	0.728	[0.53, 0.927]	0.321	[0.108, 0.533]	0.203	[0.010, 0.848]
Threats						
Scale Coefficient	**0.711**	[0.50, 0.923]	**0.252**	[0.060, 0.444]	**0.598**	[0.345, 0.950]
Excitement	0.904	[0.69, 1.116]	0.073	[-0.035, 0.180]	0.683	[0.204, 1.000]
Alcohol	0.673	[0.46, 0.885]	0.286	[0.098, 0.475]	0.683	[0.266, 1.000]
Drug	0.692	[0.48, 0.904]	0.248	[0.103, 0.392]	0.609	[0.379, 0.903]
Weapon	0.711	[0.50, 0.923]	0.252	[0.060, 0.444]	0.598	[0.355, 0.951]
Citizen Threat	0.750	[0.54, 0.962]	0.218	[0.025, 0.411]	0.593	[0.255, 0.960]
Behavioral Stance						
Scale Coefficient	**0.826**	[0.60, 1.051]	**0.179**	[-0.039, 0.398]	**0.926**	[0.846, 1.000]
Refuse to Comply	0.872	[0.65, 1.097]	0.141	[-0.031, 0.313]	0.908	[0.791, 1.000]
Ignoring Officer	0.887	[0.66, 1.113]	0.124	[-0.021, 0.269]	0.924	[0.839, 1.000]
Verbal Resistance	0.862	[0.64, 1.087]	0.179	[-0.009, 0.368]	0.911	[0.850, 1.000]
Physical Resistance	0.841	[0.62, 1.066]	0.162	[-0.036, 0.361]	0.889	[0.783, 1.000]
Physical Attack	0.826	[0.60, 1.051]	0.179	[-0.039, 0.398]	0.920	[0.829, 1.000]

Note: r_{wg}: James et al.'s within-group agreement; ad.m: Burke et al.'s average deviation for mean; ICC: intraclass correlation (ICC(1, 1)).

Table 4: Omega total, omega hierarchical, and greatest lower bound indexes for total scale (bold) and if scale item is removed.

Item	ω_t	95% CI	ω_h	95% CI	GLB
Completeness					
Scale Coefficient	**0.870**	[0.787, 0.926]	**0.402**	[0.232, 0.628]	**0.701**
Prior Video	0.737	[0.670, 0.800]	0.407	[0.266, 0.679]	0.543
Prior Audio	0.652	[0.480, 0.848]	0.403	[0.210, 0.595]	0.606
Cond Natural Lighting	0.735	[0.673, 0.794]	0.422	[0.095, 0.678]	0.481
Quality of Lighting	0.760	[0.670, 0.824]	0.439	[0.273, 0.574]	0.529
Video Obstructed	0.817	[0.748, 0.877]	0.445	[0.287, 0.581]	0.696
Overall Audio Quality	0.891	[0.798, 0.933]	0.510	[0.371, 0.685]	0.729
Respect-Passive					
Scale Coefficient	**0.784**	[0.736, 0.824]	**0.444**	[0.347, 0.531]	**0.584**
Paraphrasing	0.583	[0.390, 0.742]	0.407	[0.218, 0.572]	0.311
Reflection	0.757	[0.640, 0.861]	0.432	[0.228, 0.584]	0.537
Clarification	0.750	[0.617, 0.834]	0.438	[0.235, 0.721]	0.512
Increase Distance	0.853	[0.766, 0.922]	0.574	[0.394, 0.728]	0.531
Decrease Distance	0.774	[0.656, 0.852]	0.384	[0.179, 0.535]	0.541
Empathy	0.718	[0.554, 0.875]	0.449	[0.231, 0.598]	0.569
Respect-Active					
Scale Coefficient	**0.760**	[0.678, 0.813]	**0.448**	[0.156, 0.689]	**0.628**
Verbal Commds	0.820	[0.691, 0.894]	0.479	[0.213, 0.676]	0.635
Verbal Threat	0.765	[0.721, 0.829]	0.515	[0.333, 0.692]	0.568
Antagonize	0.760	[0.712, 0.798]	0.448	[0.328, 0.581]	0.527
Introduction	0.756	[0.676, 0.817]	0.425	[0.107, 0.645]	0.505
Repeat	0.798	[0.710, 0.865]	0.421	[0.295, 0.527]	0.495
Acknowledge	0.732	[0.665, 0.797]	0.504	[0.283, 0.617]	0.615
Debrief	0.890	[0.810, 0.940]	0.478	[0.276, 0.787]	0.699
Respect-Discourse					
Scale Coefficient	**0.906**	[0.888, 0.918]	**0.492**	[0.372, 0.621]	**0.693**
Tone	0.904	[0.877, 0.920]	0.524	[0.478, 0.604]	0.263
Dominant	0.781	[0.707, 0.842]	0.377	[0.241, 0.639]	0.247
Volume	0.822	[0.760, 0.858]	0.692	[0.334, 0.790]	0.709
Pace	0.883	[0.844, 0.912]	0.713	[0.454, 0.831]	0.635
Demeanor	0.895	[0.811, 0.925]	0.427	[0.249, 0.603]	0.662
Directly	0.958	[0.907, 0.999]	0.648	[0.128, 0.928]	0.819
Threats					
Scale Coefficient	**0.580**	[0.470, 0.658]	**0.377**	[0.192, 0.582]	**0.361**
Excitement	0.453	[0.162, 0.586]	0.264	[0.122, 0.398]	0.022
Alcohol	0.738	[0.665, 0.784]	0.358	[0.090, 0.732]	0.294
Drug	0.356	[0.209, 0.535]	0.251	[0.152, 0.344]	0.367
Weapon	0.516	[0.343, 0.622]	0.325	[0.213, 0.519]	0.393
Citizen Threat	0.504	[0.216, 0.723]	0.237	[0.147, 0.329]	0.327
Behavioral Stance					
Scale Coefficient	**0.942**	[0.894, 0.982]	**0.745**	[0.531, 0.854]	**0.922**
Refuse to Comply	0.891	[0.854, 0.926]	0.579	[0.348, 0.824]	0.851
Ignoring Officer	0.938	[0.915, 0.954]	0.745	[0.219, 0.906]	0.920
Verbal Resistance	0.948	[0.928, 0.965]	0.766	[0.640, 0.856]	0.892
Physical Resistance	0.915	[0.877, 0.964]	0.627	[0.197, 0.838]	0.920
Physical Attack	0.962	[0.928, 0.986]	0.559	[0.022, 0.934]	0.925

Note: ω_t: omega total; ω_t: omega hierarchical; GLB: greatest lower bound.

additional information, in that the reliability coefficients are weighted. As previously noted, greater interpretative weight is given to the item agreement indexes (AC_1, $\widehat{\alpha}_K$, and $B-P$ applied to individual items). Within each scale (i.e., Completeness and Respect-Passive), a weight of "1" is assigned to each of the top two item agreement coefficients, and a weight of "-1" is assigned to the two lowest item agreement coefficients (from Table 1). The interpretative weight for the item consistency index ($ICC(2,1)$) is less than the item agreement indexes, but the item consistency coefficients receive an assigned weight of "1.5" because only one coefficient represents the item consistency category. Within each scale (i.e., Completeness and Respect-Passive), a weight of "1.5" is assigned to each of the top two item agreement coefficients, and a weight of "-1.5" is assigned to the two lowest item agreement coefficients (from Table 2). The item deletion indexes ($r_{wg(i)}$, ad.m, $ICC(1,1)$, ω_t, ω_h, and glb) are weighted less. Within each scale (i.e., Completeness and Respect-Passive), a weight of "0.15" is assigned to each of the top two item agreement coefficients, and a weight of "-0.15" is assigned to the two lowest item agreement coefficients (from Tables 3 and 4).

In Figure 2, the AC_1, $\widehat{\alpha}_K$, and $B-P$ coefficients are indicated by the dark-gray lines. The item agreement coefficient weights ($ICC(2,1)$) are in are indicated by light-blue lines. The lower weighted coefficients ($r_{wg(i)}$, ad.m, $ICC(1,1)$, ω_t, ω_h, and glb) are represented by the pink lines. The item agreement coefficients (dark-gray), item consistency coefficients (light-blue), and item deletion coefficients (pink) are summed for each item. The top coefficients are placed on the right (positive) side, and the lowest two coefficients are placed on the left (negative) side of the figure.

For the Completeness scale, the Quality of Lighting item had the highest summed coefficients. The Overall Audio Quality item had the poorest summed coefficients. For the Respect-Passive scale, the Clarification item had the highest summed coefficients. The Empathy item had the poorest summed coefficients. For the Respect-Active scale, the Verbal Threat item had the highest summed coefficients. The Debrief item had the poorest summed coefficients. For the Respect-Discourse scale, the Directly item had the highest summed coefficients. The Demeanor item had the poorest summed coefficients. For the Threats scale, the Weapon item had the highest summed coefficients. The Drug item had the poorest summed coefficients. For the Behavioral Stance scale, the Physical Resistance item had the highest summed coefficients. The Ignoring Officer item had the poorest summed coefficients.

4. Discussion

BWC footage is a powerful tool to contest or control a narrative [55]. Associated with BWC footage is the promise for greater objectivity, increasing trust. This expectation is pertinent for courts (i.e., testimony impacted by silence and credibility contests), public discourse (i.e., attributing culpability), and police accountability (i.e., transparency). A three-step reliability strategy, which allowed for multiple reliability indexes integrated into a single evaluative framework, was applied to the complexities captured by police BWC footage. From this strategy, evaluations can be made to determine if key BWC footage components ought to (and ought *not* to) be a part of outcome decisions (i.e., "Was use of force justified?" and "Was the officer's behavior appropriate?") [56].

In conjunction with the three-step reliability strategy, attempts to optimize both items and ratings resulted in obtaining upper level reliabilities of BWC footage. The development of each item received considerable discussion and was then subjected to pilot cases. These items are strongly context-specific, developed solely for rating BWC footage. The raters were trained on these items [3, 57]. Ratings were conducted with a detailed rating guide (40 pages), overcoming definition and conceptual difficulties that often occur in the use of force literature [58]. This approach provided relative upper level interrater agreement, interrater consistency, and internal consistency. Upper level interrater agreement, interrater consistency, and internal consistency occurred for quality of lighting condition, officer using paraphrasing, presence of a weapon, and citizen physical resistance.

The current upper level interrater coefficients are above what are found in applied settings [59]. In addition, these referenced studies by Kraemer et al. occurred in settings with greater structure (i.e., planned interviews with standardized protocol) than the current BWC footage. Although the present methodology was limited to within analyses, the Kraemer comparison suggests that our agreement, consistency, and reliability coefficients have the potential to reflect adequate to strong levels.

Even with a limited within design, there were relatively poor levels of reliability among key components of BWC footage. This is counter to the Voigt et al. study that suggests sufficient agreement across BWC footage domains (SI annotator agreement, Cronbach's alpha index) [45]. The current study found officer decreasing distance, officer empathy, officer debriefing, and citizen excitement having relatively poorer levels of reliability.

Three study design features reduce the likelihood of relative differences (i.e., coefficients outside of confidence intervals) among rated BWC components. First, the focus on relative comparisons within each scale limited the number of potential comparisons. Second, attempts to optimize the upper levels of agreement and interrater consistency through item development, training, and rating adherence should increase the likelihood of overall stronger index coefficients. This study feature should facilitate upper index levels for the majority of items. Third, few number of raters ($n=4$), resulting in poorer distribution qualities, would naturally facilitate larger confidence intervals. These three study design features made it more difficult to have index coefficients outside of each other's confidence interval range, thereby having conservative results. Countering these study design features is the three-step strategy, which would increase the potential differences among the items. This would occur because poorer functioning indexes (i.e., less discrimination) are excluded from use.

Even with these design and strategy considerations, there was a pattern of considerable range of reliabilities, demonstrated by all the scales that have within scale items outside the 95% confidence interval. This coefficient difference provides the bases for relative weaker and stronger coefficient conclusions.

Raters are not typical public evaluators of BWC footage, let alone randomly selected raters. Ideally, reliability indexes are sample-based, estimate statistics, which infer back to the population. In the current study, there are no probability inferences calculated, thereby limiting the generalizability of the current results, but what is lost for generalizability is gained through within-study gold standard comparisons. Consequently, the interpretation of the coefficients is based upon relative differences within each scale.

4.1. Relative Stronger Agreement: Interrater Consistency Items. Upper level agreements and reliabilities were possible, with some having 1.00 coefficients or near perfect coefficients. The obtaining of upper-level coefficients is notable, given that the context is typically unstructured, highly charged, and a wide range of citizen states, often involving liberty decisions. Similarly, rating studies examining violence among mentally ill patients have found key constructs to have adequate levels of agreement [60].

Key components of conflict demonstrated relatively strong agreement. Reflection on what was said, verbal threat, and speaking directly has relatively strong agreement properties. In addition, voice volume has relatively strong interrater consistency. The relative trustworthiness of these areas in assessing officer/citizen interactions was strong.

Contextual characteristic of weapon presence and citizen behavior of degree of threat and physical resistance had strong agreement. These relatively stronger results may allow for responsibility attributions. As the attribution of dangerousness increases, attributions of responsibility can also increase [61]. Insomuch as the current targeted content assesses key components of complex interactions, aspects of officer-citizen interactions can be rated and scored to be usable information.

4.2. Relative Weaker Agreement: Interrater Consistency Items. Key components of conflict demonstrated relatively weak agreement, which occurred in a context of obtaining optimal reliability. Decreasing distance may be an antecedent to a violent interaction [62] and is a violent signal for police officers [63]; yet, footage of decreasing distance had relatively poor agreement. Empathy has been related to dominance and aggressive interactions [64]; yet, it had relatively poor agreement properties. Thus, key components of contributing to and managing conflict, when examined via BWC footage, have relatively weaker levels of agreement. The relative trustworthiness of these areas in assessing officer-citizen interactions is weak.

Areas with relatively weak agreement and consistency are central to making responsibility attributions. The following areas have previously demonstrated for making responsibility attributions: aspects of the footage audio quality [65], social distance [66], an actor's critical stance [67], and an actor's reaction/goal [68]. Each of these areas are represented by current content items of Overall Audio Quality (aspects of the footage audio quality), Decrease Distance (social distance), Tone (an actor's critical stance), and Ignoring Officer (an actor's reaction/goal). In this dataset, inferences in key areas of responsibility attributions maybe justifiably doubted or considered meaningless.

There are three implications for these relatively weak agreement and consistency results. First, as noted, inferences for responsibility attributions maybe unjustified. Second, unduly placing a level of trustworthiness in areas of weak agreement and consistency may prevent areas that have stronger agreement and consistency from being used to making inferences [56]. For example, if ignoring an officer and related tone are used to form strong inferences, then the fact that a verbal threat reliably occurred may not enter into inference making. Third, relatively poorer items may not have the capacity for reliable ratings. [13] highlights the interrater communication of the area description as an indicator of reliability. Poorer items, even with the methodological rigor applied to items, training, and ratings, may be unable to be rated.

4.3. Summary of Agreement, Interrater Consistency, and Internal Consistency Indexes Applied to Legal and Use of Force Issues. Police-worn body camera (BWC) footage is a powerful source of information for multiple contexts. Police use footage to document their actions, support internal investigations, and enhance supervision of officer conduct [69]. Courts use this footage, both as evidence to indict and evidence to assign cause and convict [55]. Media uses BWC footage expecting it to serve as systems of accountability and transparency [70]. Public movements have at times viewed BWC footages as a mechanism to document police abuses and a method to bring about police reform [71].

Court admissibility of BWC footage often centers on the completeness of the footage. The completeness issue is the amount of time that is missing from the total interaction [55]. For this criterion of completeness, the Completeness scale item of Prior Video ("Did the interaction between officer and citizen begin prior to the start of the video") had relatively strong reliability (Figure 2). Based on the current data, and in conjunction with a legal scholar suggestion to exclude partial footage from court testimony [72], this process could be made in a trustworthy manner. There are, though, other criteria of completeness. In a review of video instances where the court granted a summary judgment, 50% of the cases involved only some audio or video evidence [73]. Based on the current data, Video Obstruction and Overall Audio Quality were the two relatively poorest reliability items in the Completeness scale. Thus, adequately assessing the partialness in these two areas in order to make further decisions will be difficult.

An officer's perception of what should guide the degree of force to be used is contingent on the citizen's behavior [74, 75]. At the upper end of nonlethal force is the display of strong nonphysical and physical noncompliance. Based on the current data, items of Refuse to Comply and Physical Resistance, which are associated with the upper end of

nonlethal force, can be reliably rated (Figure 2). Two items, Ignoring Officer and Verbal Resistance, which are associated with more moderate levels of citizen resistance, have relatively poorer reliabilities (Figure 2). This may be due to greater officer discretion. But other factors of large measurement error, which could occur due to lack of training, poor measures, poorly defined target content, or the targeted content is not amenable to measurement, could result in relatively poorer reliabilities.

4.4. Observability. Commenting on the level of observability for rated items is warranted for two reasons. First, the legal system, and its emphasis on evidence law, values evidence that reflects the physical world with deductive reasoning, logical inference, and suppositions [76]. The observable nature of certain BWC footage allows for a broader recitation of the context, which may facilitate the meeting of material evidence as noted in Rule 401 [77]. Second, the level of observability for rated items has been used to explain why certain ratings have greater agreement and reliability [78]. Directly observable actions are amenable to ratings [79].

Comparisons between rating internal and external factors result in greater agreement and reliabilities for external factors [80, 81]. Similarly, comparisons between impulsive aggression and premeditated aggression resulted in greater agreement for aggression [60], but other research suggests that observability may not increase agreement and reliability [79, 82–84]. From Figure 2, relatively strong agreement and reliability occurred for both the Weapon (Threats: more observable) and Volume (Respect-Discourse: less observable) items. In contrast, relatively weak agreement and reliability occurred for both the Ignoring officer (Compliance: more observable) and Empathy (Respect-Passive: less observable) items. Similarly, other research has found both components of compliance and empathy can have poor interrater reliability [43, 85]. Based on these results, the present study concurs with past research suggesting that observability is not central to determining the relative reliabilities of BWC footage.

4.5. Poor Reliability Applied to Research and BWC Inferences. Given that key components of BWC footage can have relatively poor agreement, consistency, and internal consistency, the use of these indexes as a gate keeper of trustworthy data will influence methodological designs and statistical analyses. In a recent study using YouTube BWC footage, slightly over one-half of the selected videos were removed from the analyses because of a lack of rater agreement [9]. On the one hand, reliable data allows for the possibility of finding data relationships, but this methodological consideration may come at a cost of generalization and ecological validity of the results.

In addition to reliability reducing usable data sets, reliability can impede statistical analyses. As noted by Vacha-Hasse and Thompson, there is a General Linear Model-(GLM-) related statistics "assume perfect or at least very good score reliabilities" [159] [23]. This position is substantiated with an empirical evaluation of interrater index's relationship with the validity of outcome measures. With poorer levels of interrater agreement and consistency (varied experimentally), there are lower outcome effect sizes [2]. This emphasis on reliability for basic statistics is of particular importance for the BWC footage literature, as the majority of the empirical analyses use GLM related statistics. Given the importance of research with BWC footage, even studies that use a qualitative methodology should evaluate the content for reliability prior to using detailed content to derive summaries.

The use of BWC footage to examine implicit biases will purposely have ambiguous components of the video footage. This research design makes the assumption that there is an inference that can be had. With the ambiguous components systemic to media [86] and purposely introduced, the apparent content would not guide the inference. Thus, prior to the ambiguous component being framed, a sufficient level of reliability to allow for an inference is necessary. Without a prior level of sufficient reliability, a counterfactual argument (this is what would be, but with ambiguity other factors guide) cannot resource the methodological use of ambiguity for evaluating the presence of implicit bias. Secondly, even if there was sufficient initial reliability prior to the ambiguity frame, without a reliability assessment, the ambiguous frame maybe of such poor reliability that it cannot be used to make any trustworthy inferences.

4.6. Policy Implications and Procedural Implementation. System implementation and application to individual cases (i.e., courtroom testimony and internal police investigations) will differ. A system application would involve completing rating scales on well-sampled (100+) officer-citizen interactions. Approximately 10 to 15% of the interactions would have interrater ratings. Using the three categories of reliability indexes from the three-step reliability procedure, an equivalent Figure 2 could be computed. This would inform what elements should not be considered in routine evaluations of BWC footage within that context. For an individual case, which typically involves high-stakes decision-making, the BWC footage in question should be rated by a minimum of two raters for the purpose of assessing interrater agreement and interrater consistency. Following the three-step reliability procedure, internal consistency would not be assessed. As with the system application, element with poor reliability should not enter into the deliberations. These ratings, both for system and application to individual cases, should be conducted by raters who have a rater certification that emphasizes agreement and consistency. As noted in other areas, rater training can increase agreement and consistency [87, 88]. Such training, though, does not consistently result in strong agreement and consistency [78, 89]. And even with rater certification, agreement may only occur at moderate levels [90], for which prior attitudes are key [91].

An argument could be made that experience in officer-citizen interactions and knowledge of the benefits of best practices will draw valid conclusions without attention to interrater agreement and consistency. This experience may be a necessary condition for valid conclusions but not sufficient. In addition to this expertise, these raters require

similar ratings and consistent ratings with others for valid conclusions.

Within a court context, BWC footage, as compared to other types of evidence, increases the perceived ability to draw conclusions [55]. Compared to officer-citizen reporting modes of text and audio, greater culpability is associated with BWC footage [92]. At a broader level, the reliability challenges preclude BWC footage being a panacea for conflicting discourse. If raters trained to apply the same scale to reliable ends could not do so across all dimensions, then those with conflicting agendas viewing the same footage will be curtailed in inferring the truth-of-the-matter in an officer-citizen encounter.

4.7. Limitations and Future Research. Considerable efforts were made in the development of the items, but this was the first study in which these items were used. The items have not been standardized, and therefore, restriction of range and other statistical biases could be reflected in the structure of the items. Interrater agreement, consistency, and internal consistency were addressed. Stability of the items over time (i.e., test retest reliability) was not assessed.

The combination of methodological (sampling procedure, number of ratings, rater characteristics, endorsement rates) and statistical (reliability a function of data) considerations precludes (in our opinion) using inferential statistics. Thus, no proposed "objective" levels via a reliability benchmark criteria for BWC footage are suggested. Greater rationale and empirical work should conducted prior to applying specific benchmark criteria to BWC footage. Without this additional work, low coefficients may unduly be viewed negatively (see [20] clinical test results for lower back pain). The first step and weighting component in step three of the three-step reliability strategy should assist in developing benchmark criteria. The strongest conclusions of the present study are from the within analyses, determining if an item/scale coefficient falls outside the 95% confidence intervals of another item/scale. Yet, within a legal context, for testimony to be admissible and reliable, a reasonable method is required, as other forms of evidence are incorporated [93].

The efforts to maximize agreement and consistency preclude assessing field reliability. Understanding the upper limits of reliability in field studies (i.e., supervisor ratings) will be useful, as this may lead to the development of benchmark reliability criteria. Just as with court evidence, the current ratings are removed in time and place [55]. BWC footage and subsequent ratings cannot utilize smells, full depth perception, or ask questions; complexities which could be appreciated by a live viewer or participant [94]. Contrasting field reliability with laboratory reliability has produced lower field study reliabilities [95]. This, then, may limit some important contextual information to assess "truer" ratings. Although current field reliability studies suggest lower reliabilities, there may be complexities and content areas that are better assessed by being in the context.

Studies with a broader view of reliability will increase our ability to make inferences (both explanatory and predictive, see [96]) of BWC footage. These studies will include a focus on both internal and external aspects of reliability. The current study focused on methodologically internal aspects, accounting for random/systematic error within the data, which is a function of the process through which the data were collected. An externally focused reliability study will examine the context sensitive reliabilities in terms of their stability for other contexts (i.e., settings, statistical models, and types of validity) [97]. Although the context-specific measure and subsequent ratings can have limited generalizability [40], this context specific measure will assist in examining external focused reliabilities. Similarly, expanding reliability analyses from solely methodological (current study) to address reliability process (epistemology) and reliable reasoning from data (philosophy of science) will provide a fuller picture [32].

5. Conclusions

Trustworthiness of data is essential to assessing terrorism [98], basic demographics [99], COVID-19 measurement [35], and for ensuring management of high risk situations [100]. Even with wide usage and acceptance of medical imagery (PET, fMRI, CT), an "illusion of immediacy" may result in an unfounded level of trustworthiness [32].

Similarly, BWC footage has been viewed as containing "the unmediated truth" [55] or, at minimum, "objective" truth [101]. Components of BWC footage, such as an officer's verbal threat and speaking directly, and citizen's threat and physical resistance, have the potential for agreement and consistency. But even attempting to maximize the agreement and consistency of BWC footage ratings, the current research suggests that the trustworthiness of certain key content areas for drawing inferences can be brought into question. Consideration of BWC footage to draw inferences should not occur without addressing its trustworthiness.

Data Availability

The rated data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Gratefully acknowledged are Joe Pachea and Phil Galli for their efforts in developing the rating manual and Mike Kyle for structuring and facilitating access to the BWC footage.

Supplementary Materials

We have included a supplementary file that contains three sections. Part A contains a description of the scale content and two tables (Tables A.1 and A.2) containing measurement characteristics and content references. Part B contains the reliability formulas used with descriptive comments. Part C contains a summary count of the results in Tables 1–4. These results are used to compile Figure 2. *(Supplementary Materials)*

References

[1] V. Chillar, E. Piza, and V. Sytsma, "Conducting a systematic social observation of body-camera footage: methodological and practical insights," *Journal of Qualitative Criminal Justice & Criminology*, vol. 10, no. 4, 2021.

[2] A. G. Wilhelm, A. G. Rouse, and F. Jones, "Exploring differences in measurement and reporting of classroom observation inter-rater reliability," *Practical Assessment, Research & Evaluation*, vol. 23, no. 4, pp. 1–16, 2018, July 2020, https://login.proxy.lib.siu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=ofs&AN=129299368&site=ehost-live&scope=site.

[3] S. D. Mastrofski, R. B. Parks, A. J. Reiss Jr. et al., *Systematic observation of public police: applying _eld research methods to policy issues*, National Institute of Justice, 1998.

[4] A. J. Reiss, "Systematic Observation of Natural Social Phenomena," *Sociological Methodology*, vol. 3, p. 3, 1971.

[5] S. D. Mastrofski, R. B. Parks, and J. D. McCluskey, "Systematic social observation in criminology," in *Handbook of Quantitative Criminology*, A. R. Piquero and D. Weisburd, Eds., pp. 225–247, Springer, 2010.

[6] R. Spano, "How does reactivity affect police behavior? Describing and quantifying the impact of reactivity as behavioral change in a large-scale observational study of police," *Journal of Criminal Justice*, vol. 35, no. 4, pp. 453–465, 2007.

[7] D. W. Willits and D. A. Makin, "Show me what happened," *Journal of Research in Crime and Delinquency*, vol. 55, no. 1, pp. 51–77, 2018.

[8] V. A. Sytsma and E. L. Piza, "Script analysis of open-air drug Selling," *Journal of Research in Crime and Delinquency*, vol. 55, no. 1, pp. 78–102, 2018.

[9] B. L. Turner, E. M. Caruso, M. A. Dilich, and N. J. Roese, "Body camera footage leads to lower judgments of intent than dash camera footage," *Proceedings of the National Academy of Sciences*, vol. 116, no. 4, pp. 1201–1206, 2019.

[10] C. McKay and M. Lee, "Body-worn images: point-of-view and the new aesthetics of policing," *Culture*, vol. 16, no. 3, pp. 431–450, 2020.

[11] F. Merenda, J. Trent, and C. R. Rinke, "Untangling the role of interactions in police satisfaction: examining direct and indirect contacts with the police," *The Police Journal: Theory, Practice and Principles*, vol. 94, no. 4, pp. 496–514, 2021.

[12] J. A. Schafer, B. M. Huebner, and T. S. Bynum, "Citizen perceptions of police services: race, neighborhood context, and community policing," *Police Quarterly*, vol. 6, no. 4, pp. 440–468, 2003.

[13] C. Cope, "Ensuring validity and reliability in phenomenographic research using the analytical framework of a structure of awareness," *Qualitative Research Journal*, vol. 4, no. 2, pp. 5–18, 2004.

[14] M. Gerken, "Internalism and externalism in the epistemology of testimony," *Philosophy and Phenomenological Research*, vol. 87, no. 3, pp. 532–557, 2013.

[15] K. Krippendorff, "Misunderstanding reliability," *Methodology*, vol. 12, no. 4, pp. 139–144, 2016.

[16] J. D. W. Clifton, "Managing validity versus reliability trade-offs in scale-building decisions," *Psychological Methods*, vol. 25, no. 3, pp. 259–270, 2020.

[17] S. K. Mitchell, "Interobserver agreement, reliability, and generalizability of data collected in observational studies," *Psychological Bulletin*, vol. 86, no. 2, pp. 376–390, 1979.

[18] D. E. McNiel, J. N. Lam, and R. L. Binder, "Relevance of interrater agreement to violence risk assessment," *Journal of Consulting and Clinical Psychology*, vol. 68, no. 6, pp. 1111–1115, 2000.

[19] W. Revelle and D. M. Condon, "Reliability," in *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*, P. Irwing, T. Booth, and D. J. Hughes, Eds., pp. 709–749, John Wiley & Sons, 2018.

[20] L. Denteneer, U. Van Daele, S. Truijen, W. De Hertogh, J. Meirte, and G. Stassijns, "Reliability of physical functioning tests in patients with low back pain: A systematic review," *The Spine Journal*, vol. 18, no. 1, pp. 190–207, 2018.

[21] J. F. Edens and M. T. Boccaccini, "Taking forensic mental health assessment 'out of the lab' and into 'the real world': introduction to the special issue on the field utility of forensic assessment instruments and procedures," *Psychological Assessment*, vol. 29, no. 6, pp. 599–610, 2017.

[22] M. Lombard, J. Snyder-Duch, and C. C. Bracken, "Content analysis in mass communication: assessment and reporting of intercoder reliability," *Human Communication Research*, vol. 28, no. 4, pp. 587–604, 2002.

[23] T. Vacha-Haase and B. Thompson, "Score reliability: a retrospective look back at 12 years of reliability generalization studies," *Measurement and Evaluation in Counseling and Development*, vol. 44, no. 3, pp. 159–168, 2011.

[24] B. Thompson and T. Vacha-Haase, "Psychometrics is datametrics: the test is not reliable," *Educational and Psychological Measurement*, vol. 60, no. 2, pp. 174–195, 2000.

[25] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[26] J. M. LeBreton and J. L. Senter, "Answers to 20 questions about interrater reliability and interrater agreement," *Organizational Research Methods*, vol. 11, no. 4, pp. 815–852, 2008.

[27] D. Liljequist, B. Elfving, and K. S. Roaldsen, "Intraclass correlation–A discussion and demonstration of basic features," *PLoS One*, vol. 14, no. 7, article e0219854, 2019.

[28] S. M. Wagner, C. Rau, and E. Lindemann, "Multiple informant methodology: A critical review and recommendations," *Sociological Methods & Research*, vol. 38, no. 4, pp. 582–618, 2010.

[29] I. Fletcher, M. Mazzi, and M. Nuebling, "When coders are reliable: the application of three measures to assess interrater reliability/agreement with doctor–patient communication data coded with the VR-CoDES," *Patient Education and Counseling*, vol. 82, no. 3, pp. 341–345, 2011.

[30] H. K. Suen, *Principles of test theories*, Lawrence Erlbaum Associates, Inc, 1990.

[31] H. E. Tinsley and D. J. Weiss, "Interrater reliability and agreement of subjective judgments," *Journal of Counseling Psychology*, vol. 22, no. 4, pp. 358–376, 1975.

[32] E. Lalumera, S. Fanti, and G. Boniolo, "Reliability of molecular imaging diagnostics," *Synthese*, vol. 198, no. S23, pp. 5701–5717, 2021.

[33] M. Banerjee, M. Capozzoli, L. McSweeney, and D. Sinha, "Beyond kappa: A review of interrater agreement measures," *The Canadian Journal of Statistics*, vol. 27, no. 1, pp. 3–23, 1999.

[34] D. McNeish, "Thanks coefficient alpha, we'll take it from here," *Psychological Methods*, vol. 23, no. 3, pp. 412–433, 2018.

[35] E. Mbunge, B. Akinnuwesi, S. G. Fashoto, A. S. Metfula, and P. Mashwama, "A critical review of emerging technologies for tackling COVID-19 pandemic," *Human Behavior and Emerging Technologies*, vol. 3, no. 1, pp. 25–39, 2021.

[36] I. Adams and S. Mastracci, "Visibility is a trap: the ethics of police body-worn cameras and control," *Administrative Theory & Praxis*, vol. 39, no. 4, pp. 313–328, 2017.

[37] T. I. Cubitt, R. Lesic, G. L. Myers, and R. Corry, "Body-worn video: a systematic review of literature," *Australian & New Zealand Journal of Criminology*, vol. 50, no. 3, pp. 379–396, 2017.

[38] E. Laming, "Police use of body worn cameras," *Police Practice and Research*, vol. 20, no. 2, pp. 201–216, 2019.

[39] A. Mateescu, A. Rosenblat, and D. Boyd, "Dreams of accountability, guaranteed surveillance: the promises and costs of body-worn cameras," *Surveillance and Society*, vol. 14, no. 1, pp. 122–127, 2016.

[40] S. M. Chafouleas, "Direct behavior rating: a review of the issues and research in its development," *Education and Treatment of Children*, vol. 34, no. 4, pp. 575–591, 2011.

[41] R. J. Volpe and A. M. Briesch, "Dependability of two scaling approaches to direct behavior rating multi-item scales assessing disruptive classroom behavior," *School Psychology Review*, vol. 45, no. 1, pp. 39–52, 2016.

[42] T. M. Liddell and J. K. Kruschke, "Analyzing ordinal data with metric models: what could possibly go wrong?," *Journal of Experimental Social Psychology*, vol. 79, pp. 328–348, 2018.

[43] R. J. Volpe, R. K. Chaffee, T. S. Yeung, and A. M. Briesch, "Initial development of multi-item direct behavior rating measures of academic enablers," *School Mental Health*, vol. 12, no. 1, pp. 77–87, 2020.

[44] M. J. Kyle and D. G. Kroner, *Video coding manual*, Southern Illnois Univeristy Carbondale, Carbondale, Illinois, 2017.

[45] R. Voigt, N. P. Camp, V. Prabhakaran et al., "Language from police body camera footage shows racial disparities in officer respect," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 25, pp. 6521–6526, 2017.

[46] D. V. Cicchetti, D. Shoinralter, and P. J. Tyrer, "The effect of number of rating scale categories on levels of interrater reliability: a Monte Carlo investigation," *Applied Psychological Measurement*, vol. 9, no. 1, pp. 31–36, 1985.

[47] B. D. Zumbo and E. Kroc, "A measurement is a choice and Stevens' scales of measurement do not help make it: A response to Chalmers," *Educational and Psychological Measurement*, vol. 79, no. 6, pp. 1184–1197, 2019.

[48] L. R. James, R. G. Demaree, and G. Wolf, "Estimating within-group interrater reliability with and without response bias," *Journal of Applied Psychology*, vol. 69, no. 1, pp. 85–98, 1984.

[49] W. Revelle, "Psych: Procedures for psychological, psychometric, and personality research (Version 1.9.12.31)," 2020, February 2020, https://CRAN.R-project.org/package=psych.

[50] J. Schmid and J. M. Leiman, "The development of hierarchical factor solutions," *Psychometrika*, vol. 22, no. 1, pp. 53–61, 1957.

[51] J. M. F. Ten Berge and G. Soffcan, "The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality," *Psychometrika*, vol. 69, no. 4, pp. 613–625, 2004.

[52] P. D. Bliese, "Multilevel: Multilevel Functions (Version 2.6)," 2016, https://CRAN.R-project.org/package=multilevel.

[53] P. D. Bliese, M. A. Maltarich, J. L. Hendricks, D. A. Hofmann, and A. B. Adler, "Improving the measurement of group-level constructs by optimizing between-group differentiation," *Journal of Applied Psychology*, vol. 104, no. 2, pp. 293–302, 2019.

[54] D. B. Dahl, M. D. Scott, C. Roosen et al., "Package 'Xtable', R package version 1.8-4," 2019, https://cran.r-project.org/web/packages/xtable/index.html.

[55] M. D. Fan, "Justice visualized: courts and the body camera revolution," *U.C. Davis Law Review*, vol. 50, no. 3, pp. 897–960, 2016, February 2021, https://heinonline.org/HOL/P?h=hein.journals/davlr50&i=917.

[56] B. Holman, "An ethical obligation to ignore the unreliable," *Synthese*, vol. 198, no. S23, article 2483, pp. 5825–5848, 2021.

[57] L. M. Fallon, L. M. H. Sanetti, S. M. Chafouleas, M. N. Faggella-Luby, and A. M. Briesch, "Direct training to increase agreement between teachers' and observers' treatment integrity ratings," *Assessment for Effective Intervention*, vol. 43, no. 4, pp. 196–211, 2018.

[58] C. F. Klahm IV., J. Frank, and J. Liederbach, "Understanding police use of force," *Policing*, vol. 37, no. 3, pp. 558–578, 2014.

[59] H. C. Kraemer, D. J. Kupfer, D. E. Clarke, W. E. Narrow, and D. A. Regier, "DSM-5: how reliable is reliable enough?," *American Journal of Psychiatry*, vol. 169, no. 1, pp. 13–15, 2012.

[60] A. R. Felthous, D. Weaver, R. Evans et al., "Assessment of impulsive aggression in patients with severe mental disorders and demonstrated violence: inter-rater reliability of rating instrument," *Journal of Forensic Sciences*, vol. 54, no. 6, pp. 1470–1474, 2009.

[61] V. L. Quinsey and M. Cyr, "Perceived dangerousness and treatability of Offenders," *Journal of Interpersonal Violence*, vol. 1, no. 4, pp. 458–471, 1986.

[62] J. S. Wormith, "Personal space of incarcerated offenders," *Journal of Clinical Psychology*, vol. 40, no. 3, pp. 815–827, 1984.

[63] R. R. Johnson, "Perceptions of interpersonal social cues predictive of violence among police officers who have been assaulted," *Psychology*, vol. 30, no. 2, pp. 87–93, 2015.

[64] S.-k. T. J. Hudson, M. Cikara, and J. Sidanius, "Preference for hierarchy is associated with reduced empathy and increased counter-empathy towards others, especially out-group targets," *Journal of Experimental Social Psychology*, vol. 85, article 103871, 2019.

[65] N. O. Sidaty, M.-C. Larabi, and A. Saadane, "Inuence of video resolution, viewing device and audio quality on perceived multimedia quality for steaming applications," in *2014 5th European Workshop on Visual Information Processing (EUVIP)*, pp. 1–6, Paris, France, 2014.

[66] O. Ybarra and W. G. Stephan, "Attributional orientations and the prediction of behavior: the attribution–prediction bias," *Journal of Personality and Social Psychology*, vol. 76, no. 5, pp. 718–727, 1999.

[67] S. R. Lopez, K. A. Nelson, K. S. Snyder, and J. Mintz, "Attributions and affective reactions of family members and course of schizophrenia," *Journal of Abnormal Psychology*, vol. 108, no. 2, pp. 307–314, 1999.

[68] E. Stephan, D. Shidlovski, and D. Heller, "Distant determination and near determinism: the role of temporal distance in prospective attributions to will," *Journal of Experimental Social Psychology*, vol. 68, pp. 113–121, 2017.

[69] D. Makin, "Avoiding the Technological Panacea: The Case of the Body-Worn Camera," in *Criminal justice technology in the 21st century*, pp. 86–102, Charles C Thomas Publisher, 2017.

[70] C. Naoroz and H. M. D. Cleary, "News media framing of police body-worn cameras: a content analysis," *Policing: A Journal of Policy and Practice*, vol. 15, no. 1, pp. 540–555, 2021.

[71] J. E. Wright and A. M. Headley, "Can technology work for policing? Citizen perceptions of police-body worn cameras," *The American Review of Public Administration*, vol. 51, no. 1, pp. 17–27, 2021.

[72] M. D. Fan, "Missing police body camera videos: remedies, evidentiary fairness, and automatic activation," *Georgia Law Review*, vol. 52, no. 1, pp. 57–108, 2017, July 2021, https://login.proxy.lib.siu.edu/login?url=https://search.ebscohost.com/login. aspx?direct=true&db=a9h&AN=128737618&site=ehost-live&scope=site.

[73] M. Zamoff, "Assessing the impact of police body camera evidence on the litigation of excessive force cases," *Georgia Law Review*, vol. 54, no. 1, pp. 1–60, 2019, April 2021, https://login.proxy.lib.siu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=141483792&site=ehost-live&scope=site.

[74] E. A. Paoline III. and W. Terrill, "Listen to me! Police officers' views of appropriate use of force," *Journal of Crime and Justice*, vol. 34, no. 3, pp. 178–189, 2011.

[75] E. A. Paoline III., W. Terrill, and L. J. Somers, "Police officer use of force mindset and street-level behavior," *Police Quarterly*, vol. 24, no. 4, pp. 547–577, 2021.

[76] C. P. Nemeth, *Law and Evidence: A Primer for Criminal Justice, Criminology, Law and Legal Studies (Second)*, Jones & Bartlett Publishers, 2010.

[77] "Federal Rules of Evidence," 2019, November 2020, https://www.law.cornell.edu/rules/fre.

[78] P. J. Kennealy, J. L. Skeem, and I. R. Hernandez, "Does staff see what experts see? Accuracy of front line staff in scoring juveniles' risk factors," *Psychological Assessment*, vol. 29, no. 1, pp. 26–34, 2017.

[79] D. J. Pepler and W. M. Craig, "A peek behind the fence: naturalistic observations of aggressive children with remote audiovisual recording," *Developmental Psychology*, vol. 31, no. 4, pp. 548–553, 1995.

[80] J. Duxbury and R. Whittington, "Causes and management of patient aggression and violence: staff and patient perspectives," *Journal of Advanced Nursing*, vol. 50, no. 5, pp. 469–478, 2005.

[81] M. Willoughby, J. Kupersmidt, and D. Bryant, "Overt and covert dimensions of antisocial behavior in early childhood," *Journal of Abnormal Child Psychology*, vol. 29, no. 3, pp. 177–187, 2001.

[82] M. Gebhardt, J. M. DeVries, J. Jungjohann, G. Casale, A. Gegenfurtner, and J.-T. Kuhn, "Measurement invariance of a direct behavior rating multi item scale across occasions," *Social Sciences*, vol. 8, no. 2, p. 46, 2019.

[83] K. G. Lampe, E. A. Mulder, O. F. Colins, and R. R. J. M. Vermeiren, "The inter-rater reliability of observing aggression: A systematic literature review," *Aggression and Violent Behavior*, vol. 37, pp. 12–25, 2017.

[84] S. G. Roch, A. R. Paquin, and T. W. Littlejohn, "Do raters agree more on observable items?," *Human Performance*, vol. 22, no. 5, pp. 391–409, 2009.

[85] S. Brown, M. St. Amand, and E. Zamble, "The dynamic prediction of criminal recidivism: a three-wave prospective study," *Law and Human Behavior*, vol. 33, no. 1, pp. 25–45, 2009.

[86] K. Ishii, M. M. Lyons, and S. A. Carr, "Revisiting media richness theory for today and future," *Human Behavior and Emerging Technologies*, vol. 1, no. 2, pp. 124–131, 2019.

[87] D. A. Cook, D. M. Dupras, T. J. Beckman, K. G. Thomas, and V. S. Pankratz, "Effect of rater training on reliability and accuracy of mini-CEX scores: A randomized, controlled trial," *Journal of General Internal Medicine*, vol. 24, no. 1, p. 74, 2009.

[88] E. S. Holmboe, R. E. Hawkins, and S. J. Huot, "Effects of training in direct observation of medical residents' clinical competence: a randomized trial," *Annals of Internal Medicine*, vol. 140, no. 11, pp. 874–881, 2004.

[89] S. Barrett, "The impact of training on rater variability," *International Education Journal*, vol. 2, no. 1, pp. 49–58, 2001.

[90] Y. Attali, "Rater certification tests: a psychometric approach," *Educational Measurement: Issues and Practice*, vol. 38, no. 2, pp. 6–13, 2019.

[91] H. Bijani, "Investigating the validity of oral assessment rater training program: a mixed-methods study of raters' perceptions and attitudes before and after training," *Cogent Education*, vol. 5, no. 1, article 1460901, 2018.

[92] M. McCamman and S. Culhane, "Police body cameras and us: public perceptions of the justification of the police use of force in the body camera era," *Psychological Science*, vol. 3, no. 2, pp. 167–175, 2017.

[93] E. J. Imwinkelried, "Should courts incorporate a best evidence rule into the standard determining the admissibility of scientific testimony?: enough is enough even when it is not the best," *Case Western Reserve Law Review*, vol. 50, no. 1, pp. 19–51, 1999, July 2021, https://login.proxy.lib.siu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=ofs&AN=502312568&site=ehost-live&scope=site.

[94] J. A. Schafer, J. Hibdon, and M. Kyle, "Studying rare events in policing: the allure and limitations of using body-worn camera video," *Journal of Crime and Justice*, pp. 1–16, 2022.

[95] T. L. F. De Beuf, C. de Ruiter, J. F. Edens, and V. de Vogel, "Taking "the boss" into the real world: field interrater reliability of the Short-Term Assessment of Risk and Treatability: Adolescent Version," *Behavioral Sciences & the Law*, vol. 39, no. 1, pp. 123–144, 2021.

[96] G. Shmueli, "To explain or to predict?," *Statistical Science*, vol. 25, no. 3, pp. 289–310, 2010.

[97] B. Osimani, "Epistemic gains and epistemic games: reliability and higher order evidence in medicine and pharmacology," in *Uncertainty in Pharmacology: Epistemology, Methods, and Decisions*, A. LaCaze and B. Osimani, Eds., pp. 345–372, Springer Nature, 2020.

[98] C. Clemmow, N. Bouhana, and P. Gill, "Analyzing person-exposure patterns in lone-actor terrorism," *Policy*, vol. 19, no. 2, pp. 451–482, 2020.

[99] P. T. Cirino, C. E. Chin, R. A. Sevcik, M. Wolf, M. Lovett, and R. D. Morris, "Measuring socioeconomic status: reliability and preliminary validity for different approaches," *Assessment*, vol. 9, no. 2, pp. 145–155, 2002.

[100] R. Moura, M. Beer, E. Patelli, J. Lewis, and F. Knoll, "Learning from accidents: interactions between human factors, technology and organisations as a central element to validate risk studies," *Safety Science*, vol. 99, pp. 196–214, 2017.

[101] S. W. Phillips, "Eyes are not cameras: the importance of integrating perceptual distortions, misinformation, and false memories into the police body camera debate," *Policing*, vol. 12, no. 1, article paw008, 2016.