AEGIS
Current Criminology
2026 / 3 (1) / 22-39

RESEARCH ARTICLE

Liberal-Conservative Hierarchies of Intercoder Reliability Indices for Criminological Research

Shuhuan Zeng[1]; Dianshi Moses Li[2]; Guangchao Charles Feng[3]; Song Harris Ao[4]; Ming Milano Li[5]; Hui Huang[6]; Ke Deng[7]; Zhujin Zhang[8]; Xinshu Zhao[9]*

[1] Faculty of Science, The University of Melbourne, Melbourne, Australia

[2] Centre for Empirical Legal Study, Faculty of Law, University of Macau, Taipa, Macao

[3] Department of Interactive Media, School of Communication, Hong Kong Baptist University, Kowloon, Hong Kong

[4] School of Journalism and Communication, Sun Yat-sen University, Guangzhou, China

[5] Department of Government and Public Administration, Faculty of Social Sciences, University of Macau, Taipa, Macao

[6] Tenly Inc., Shanghai, China

[7] Department of Statistics & Data Science, Tsinghua University, Beijing, China

[8] Faculty of Law, University of Macau, Macau SAR, China

[9] Professor Emeritus, UNC Hussman School of Journalism and Media, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Corresponding author: Xinshu Zhao (xszhaoum@gmail.com)

Submitted: 2026-04-15 Accepted: 2026-04-28 Published: 2026-04-30

Abstract: Criminological research depends extensively on human coding. Interviews, narratives, case files, visual records, and open-source materials are routinely converted into analyzable data through coding decisions made by researchers and research assistants. Yet criminological studies report intercoder and interrater reliability inconsistently. Even when reliability is reported, it is often assessed using heterogeneous coefficients that are treated as directly comparable, despite resting on different assumptions and lacking interchangeability on a common scale. The present study does not offer a procedural guide or prescribe a universal coefficient for criminological research. Instead, it identifies the liberal–conservative ordering of reliability estimators and examines how that ordering can inform coefficient selection and interpretation in human-coded research. We first extend previous mathematics-based hierarchies to include 23 indices. We then use Monte Carlo simulations to construct six additional hierarchies under varying conditions of category number, sample size and distributional skew. Across eight hierarchies, a consistent pattern emerges, together with a previously undetected paradox in Perreault and Leigh's Ir. The resulting hierarchies provide criminological researchers with a principled basis for choosing among non-equivalent reliability estimators and for interpreting reported coefficients more cautiously.

Keywords: intercoder reliability, interrater reliability, qualitative criminology, criminal justice methods, Cohen's κ, Krippendorff's α

Affiliations

Author Email Addresses

Author	Email
Shuhuan Zeng	shuhuanz@student.unimelb.edu.au
Dianshi Moses Li	yc37228@um.edu.mo
Guangchao Charles Feng	fffchao@gmail.com
Song Harris Ao	aosong3@mail.sysu.edu.cn
Ming Milano Li	yc17316@um.edu.mo
Hui Huang	hh@tenly.com
Ke Deng	kdeng@tsinghua.edu.cn
Zhujin Zhang	yc57208@um.edu.mo
Xinshu Zhao	xszhaoum@gmail.com

Postal Address

Xinshu Zhao: UNC Hussman School of Journalism and Media, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3365, USA

Declarations

Competing interests: The authors declare no competing interests.

Funding This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Introduction

Criminology depends not only on counts, rates and models, but also on interpretation. Researchers in the field routinely turn to qualitative interviews, case files, observational materials, visual records and media texts when they seek to understand lived experience, social meaning, interactional dynamics and the representation of crime and control (Brent et al., 2024; Kort-Butler, 2016). Recent discussions within the discipline likewise point to the continued vitality of qualitative, narrative, visual and culturally oriented criminological research (Copes et al., 2020; Kokkalera et al., 2023).

Where criminological evidence is generated through human coding, intercoder reliability (hereafter including interrater reliability where the latter term is used in criminological practice) becomes part of the methodological foundation of the research. This is true for interview-based thematic analysis, for the unitizing and coding of narratives, for behavioural coding in video-based studies of violence, and for open-source databases that convert dispersed textual and visual traces into structured event data (Adams et al., 2017; Chermak et al., 2025; Ejbye-Ernst, 2023; Weenink et al., 2022).

Yet criminology has paid uneven attention to this issue. A review of 438 qualitative articles published in 17 leading criminology and criminal justice journals between 2010 and 2019 found that 89.7% made no reference to intercoder or interrater reliability, while only 10.3% discussed it in any form (Copes et al., 2020). This is not just a reporting problem. When coding procedures and coding agreement remain underspecified, readers have less leverage for evaluating analytic transparency, comparing studies and judging the evidentiary status of qualitative findings (Copes et al., 2020).

Even when reliability is reported, criminological practice is far from uniform. Some studies rely on Guetzkow's U for unitizing and Cohen's κ for thematic coding (Adams et al., 2017). Other studies use Conger's κ to assess agreement among three coders (Stone, 2016), Cohen's κ for time-sensitive behavioural coding in video data (Ejbye-Ernst, 2023), Krippendorff's α together with intraclass correlation coefficients in visual analyses of violent incidents (Weenink et al., 2022), and Tinsley and Weiss's T in offender risk assessment. This pluralism is not inherently problematic (van der Knaap et al., 2012). However, it does create a practical difficulty: coefficients based on different assumptions are often interpreted as if their numerical values occupied a single common scale (Zhao et al., 2024; Zhao et al., 2025).

The purpose of this study is not to offer a procedural guide to reliability assessment in criminology, nor to prescribe a universal coefficient for all human-coded research. Instead, this study establishes liberal–conservative hierarchies among intercoder reliability estimators. It extends prior mathematics-based hierarchies to 23 indices and develops six simulation-based hierarchies using Monte Carlo simulations that vary the number of categories, sample size, and distributional skewness. Taken together, these eight hierarchies identify a broadly stable ordering

across estimators and surface a previously undetected paradox in Perreault and Leigh's Ir. Such hierarchies provide criminological researchers with a principled basis for selecting among non-equivalent coefficients and for interpreting published reliability estimates more cautiously. In a field where reliability is methodologically consequential yet inconsistently conceptualized, that task is foundational rather than ancillary.

From underreporting to coefficient choice

One response to this situation would be to search for a single coefficient that criminology should adopt across all coding tasks. We do not take that position. Criminological coding problems vary widely across interviews, narratives, observational records, videos and open-source event data. At the same time, available coefficients embody different assumptions about coder behaviour, chance agreement and the meaning of reliability. The more defensible question is therefore not which estimator is universally correct, but how the available estimators are ordered and what that ordering implies for methodological choice.

This is where the broader reliability debate becomes directly relevant to criminology. The long-standing dominance of chance-corrected coefficients, particularly Cohen's κ and Krippendorff's α, has increasingly been challenged by theoretical, mathematical, and experimental work showing that these indices may rely on questionable assumptions about random coding and may behave paradoxically under ordinary empirical conditions (Zhao et al., 2013, 2018, 2022). For criminological researchers, the immediate implication is not that reliability assessment has become less important. It is that coefficient choice itself has become a substantive methodological decision. Before one can decide which coefficient to report, one needs to know how available estimators relate to one another.

A 2022 controlled experiment provides a particularly useful point of departure here, testing seven widely used indices of intercoder reliability against true intercoder reliability.

Among the most disturbing features of the trio was that each one showed questionable validity in four respects:

1) *Chance estimates of κ, α and π are negatively correlated with true chance agreement*. The κ-, α- and π-estimated chance agreements, the unique core of each index, were each negatively correlated with the true chance agreement (dr^2=-.152~-.151). This means that the trio tends to subtract more agreement when chance agreement is lower, and less agreement when chance agreement is higher. In other words, many "chance agreements" removed by the trio are in fact true agreements, while many "true agreements" not removed by the trio are in fact chance agreements. This feature alone should have invalidated the trio, given that accurately removing chance agreements was the main task of any chance-correcting index.

2) *Indices κ, α and π are poorly correlated with empirical validity, and much more poorly than percent agreement*. In the experiment, the true reliability was highly correlated with correct coding, i.e., empirical validity of the measurement, while each of the trio's estimated reliabilities was poorly correlated with true reliability. Each of the trio also significantly underperformed percent agreement while predicting true reliability (dr^2=.841 vs dr^2=.312). Taken together, these findings suggest that each member of the trio is weakly correlated with measurement validity and tends to underperform percent agreement in predicting empirical validity. This feature alone casts serious doubt on the validity of the three indices, since the main purpose of estimating reliability is to

predict validity, and the central justification for chance-correcting indices is that they should predict validity better than percent agreement.

3) *Indices κ, α and π are affected by evenness when they should not be.* The three indices were not only poorly correlated with true reliability (dr^2=.312), but they were correlated nearly as much with the evenness of distribution (dr^2=.292~.293), which means the three reliability indices measure evenness as much as they measure reliability. The experiment also shows that true reliability was not at all affected by distribution (dr^2=.000), which means that reliability indices should not be correlated with evenness at all.

4) *Imposing κ, α* or *π imposes evenness bias, making the world appear flatter.* That *κ, α* and *π* measure evenness has a troublesome implication. Fixed benchmarks, such as $\alpha \geq 0.80$ or $\alpha \geq 0.667$, are often used to evaluate measurement instruments (Krippendorff, 2004). Since these indices are influenced by distribution evenness, more evenly distributed datasets are more likely to meet these benchmarks, skewing the representation of scientific knowledge towards a "flatter" view of the world, a phenomenon referred to as KAP-imposed evenness bias.

These findings align with a growing body of opinions, analyses, and studies that highlight the pitfalls of α and κ, with some even calling for their banishment (de Vet et al., 2006; Delgado & Tibau, 2019; Feng, 2014; Gwet, 2008; Hoehler, 2000; Jakubauskaite, 2021; Jiang et al., 2021; Kraemer, 1979; Krippendorff, 1970; Krippendorff, 2004; Lombard et al., 2002; Riffe et al., 2023; Stütz et al., 2022; Tong et al., 2020; Xu & Lorber, 2014; Zec et al., 2017; Zhao, 2011; Zhao et al., 2018). This supports the "best available for a situation (BAFS)" approach, which is increasingly advocated by experts. Recognizing that no single index is perfect for all situations, the BAFS approach encourages researchers to select two or more indices that exhibit the fewest and least harmful deficiencies for the specific research context (Chermak et al., 2025; Dettori & Norvell, 2020; Gwet, 2008; Han et al., 2023; Hoehler, 2000; Jiang et al., 2021; Ju et al., 2026; Kraemer, 1979; Kraemer, 1992; Li et al., 2018; Li et al., 2025; Liu & Li, 2024; Liu et al., 2025; Nili et al., 2020; ten Hove et al., 2018; Zhao et al., 2022; Zhao et al., 2018; Zhao et al., 2012).

The BAFS approach requires researchers to know the indices' characteristics, including their tendencies to score high or low, namely their positions on a liberal-conservative scale (Krippendorff, 2019). The knowledge also may help readers when interpreting research based in part on the scores produced by an intercoder reliability index, such as Cohen's κ (Cohen, 1960) and Krippendorff's α (Krippendorff, 1980). Since these two seminal publications, numerous studies have cited κ or α to verify or document the empirical basis of their measurement. More studies may be expected built in part on the two indices. Now that the eccentric behavioral assumptions and poor empirical performance of the two indices are better known, systematically investigated liberal-conservative tendencies of the indices may also help in proper interpretation or reinterpretation of the future and past studies.

Two Functions of Reliability Indices and Need for Hierarchy

In research practice, intercoder reliability indices perform two functions.

Function 1, cross-instrument comparison. One function is to evaluate instruments, e.g., diagnoses, coding, and observations, against each other. The instruments with higher scores are considered more reliable than those with lower scores (Krippendorff, 2016; Riffe et al., 2023; Shrout, 1998).

Function 2: benchmark comparison. The second function is to evaluate instruments against fixed benchmarks or benchmark systems. For example, the Landis-Koch benchmark system marks Cohen's $\kappa<0$ as poor, $\kappa=0\sim.2$ as slight, .21~.4 as fair, .61~.80 as substantial, and $\kappa >.81$ as almost perfect (Cohen, 1960; Silveira & Siqueira, 2023). The system is influential across disciplines (Li et al., 2018). Krippendorff (2004) recommends $\alpha \geq .80$ as the standard for drawing reliable conclusions, while allowing $\alpha \geq .667$ to be accepted tentatively in some circumstances. This approach has been particularly influential in communication research (Hayes & Krippendorff, 2007; Shrout, 1998).

For either function, knowledge of the indices' positions on liberal-conservative hierarchies is necessary for proper reading and interpretation of index scores. For cross-instrument comparison, e.g., a Cohen's $\kappa = .81$ for one diagnostician may not necessarily indicate higher reliability than a Krippendorff's $\alpha = .79$ for another diagnostician, especially if κ is known to produce higher values than α when diagnostic categories are numerous and prevalence estimates differ substantially between diagnosticians (Cohen, 1960; Krippendorff, 1970; Krippendorff, 2004; Vach, 2005; von Eye & von Eye, 2008; Zhao et al., 2013).

For benchmark-based comparisons, a reader might classify $\kappa = .81$ in one study as "almost perfect" under the Landis–Koch benchmark, while treating $\alpha = .79$ in another study as only tentatively acceptable under Krippendorff's benchmark (Krippendorff, 2004, 2018; Silveira & Siqueira, 2023). The readers would be less tempted to do so if they understand that, in this situation, the difference in scores may reflect less different qualities of the instruments, but more the discrepant assumptions of the indices.

In criminology, these two uses of reliability estimates are both common and consequential. Researchers compare coefficients across studies in order to judge the apparent quality of coding procedures, and they invoke thresholds or benchmark language to decide whether a given body of coded evidence is methodologically acceptable. The present study is not intended to function as a field manual for all such decisions. Its narrower aim is to identify the liberal–conservative ordering of available estimators, so that criminological researchers can make more informed choices among non-equivalent coefficients and interpret reported values with greater caution. In this sense, the hierarchies developed here are best understood as an interpretive basis for coefficient selection rather than as a substitute for substantive judgement about coding design.

Liberal and Conservative Tendencies of Intercoder Reliability Indices

The concept of liberal versus conservative scales of reliability indices is not new. Lombard et al. (2002) observed that some indices are more "liberal" while others are more "conservative," with which Krippendorff (2004) disagreed. Zhao et al. (2013) opined that information about numerical patterns can be helpful so long as the information is interpreted with a clear understanding of the concepts and assumptions behind the indices.

Based on their analysis of indices' mathematical formulas, Zhao et al. (2013) showed that the numerical values can be dramatically different for different indices, and they produced two liberal-conservative hierarchies for the 22 indices. Relatedly, ten Hove et al. (2018) showed that different intercoder reliability indices may yield substantially different numerical values when applied to the same data, reinforcing earlier concerns about coefficient non-equivalence (Popping, 1988; Zhao et al., 2013).

However, math-based hierarchies alone may not provide the full picture. Mathematical analyses depend on interpretations of formulas, which can lead to omissions, overemphasis, or errors. This is why modern

mathematical studies often pair rigorous derivations with Monte Carlo simulations (Warrens, 2014). The validity of the hierarchies may be strengthened if they are also informed by empirical data. A main objective of this study is to build additional hierarchies based on simulated empirical data. We hope the two types of hierarchies may verify, complement, correct, and stimulate each other, giving us a more complete picture how the indices behave.

We started with the 22 indices in the two math-based hierarchies built by a social scientist and two mathematical statisticians (Zhao et al., 2013). Being in the existing hierarchies therefore serving as proper references for comparison, the indices also have been analyzed, reanalyzed, tested, retested, debated and re-debated (Byrt et al., 1993; Charles Feng & Zhao, 2016; Chmura Kraemer et al., 2002; Cicchetti & Feinstein, 1990; Freelon, 2010; G wet, 2008; Hoehler, 2000; Kraemer, 1979; Kraemer, 1992; Krippendorff, 2004, 2016, 2018, 2019; Zhao et al., 2022; Zhao et al., 2018; Zhao et al., 2013; Zwick, 1988). Before developing the simulation-based hierarchies, we first identified an additional index that should be incorporated into the mathematics-based hierarchies, as discussed below.

New Index from Reinterpretation of λr

Goodman and Kruskal (1959) proposed an agreement index, λr, based on a chance agreement (ac) estimation that behaves in some ways similarly to Cohen's κ (Cohen, 1960):

$$\mathbf{a_c} = \frac{\mathbf{1}}{2}\left(\frac{N_{l1}}{N} + \frac{N_{l2}}{N}\right)$$

Some, e.g., Zhao et al. (2013), interpreted N_{l1} and N_{l2} as, respectively, individual frequency reported by each coder, hence (*nl1*+*nl2*)/2, where *nl1*= N_{l1}/N and *nl2*= N_{l2}/N, represent the *average frequency* of the two coders. Suppose on a binary scale Coder 1 reports 85 cases in Category 1 and 15 cases in Category 2, while Coder 2 reports 45 cases in Category 1 and 55 cases in Category 2, N_{l1}=85, N_{l2}=55, and *ac*=(.85+.55)/2=0.7. Goodman and Kruskal's λr shares with major indices the Guttman-Bennet chance-removal formula (Bennett et al., 1954; G oodman & Kruskal, 1959). Inserting Eq. 18 into the classic formula, we have Goodman and Kruskal's λr.

Fleiss (1975), however, interpreted (nl1+nl2)/2 as the average frequency reported by two coders, which in the above example would instead produce an ac= (.85+.45)/2=.65. As Goodman and Kruskal (1959) did not provide a numerical example, it is not clear which interpretation represents the authors' intention. We therefore treat the two interpretations as distinct indices: the individual interpretation, labelled λ_i (Zhao et al., 2012), and the average interpretation, labelled λ_a (Fleiss, 1975). This brings the total number of indices considered in this study to 23 (ten Hove et al., 2018).

Expanded Math-Based Hierarchies of 23 Indices

In the above example, average interpretation produces a smaller *ac* than individual interpretation. Comparing Table 1 with Table 2, we see it is not a fluke: an *ac* estimation by average interpretation is always smaller than or equal to the counterpart *ac* estimation by individual interpretation, and a smaller *ac* leads to a larger index, λa ≥ λi. The two interpretations differ only when the two coders' estimated distributions are skewed at the opposite directions, e.g., one reports 90/10% while the other reports 5/95%, which is represented by the upper left quarter and the lower right quarter of Table 1 and Table 2. In research practice, these situations are less frequent than the

situations represented by the other two quarters, which indicate that the two coders do not disagree with each other in terms of skew directions.

Table 1. Goodman and Kruskal's Chance Agreement (ac) (Individual Interpretation) as a Function of Two Distributions *

		Distribution *1:* Positive Findings by Coder *1* (N_{p2}/N) in %**										
		0	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
Distribution *2:* Positive Findings by Coder *2* (N_{p2}/N) in % **	100	100.0	95.0	90.0	85.0	80.0	75.0	80.0	85.0	90.0	95.0	100.0
	90	95.0	90.0	85.0	80.0	75.0	70.0	75.0	80.0	85.0	90.0	95.0
	80	90.0	85.0	80.0	75.0	70.0	65.0	70.0	75.0	80.0	85.0	90.0
	70	85.0	80.0	75.0	70.0	65.0	60.0	65.0	70.0	75.0	80.0	85.0
	60	80.0	75.0	70.0	65.0	60.0	55.0	60.0	65.0	70.0	75.0	80.0
	50	75.0	70.0	65.0	60.0	55.0	50.0	55.0	60.0	65.0	70.0	75.0
	40	80.0	75.0	70.0	65.0	60.0	55.0	60.0	65.0	70.0	75.0	80.0
	30	85.0	80.0	75.0	70.0	65.0	60.0	65.0	70.0	75.0	80.0	85.0
	20	90.0	85.0	80.0	75.0	70.0	65.0	70.0	75.0	80.0	85.0	90.0
	10	95.0	90.0	85.0	80.0	75.0	70.0	75.0	80.0	85.0	90.0	95.0
	0	100.0	95.0	90.0	85.0	80.0	75.0	80.0	85.0	90.0	95.0	100.0

*: The table was adapted from Zhao et al. (2013). Main cell entries are Goodman and Kruskal's Chance Agreement (ac) in %.

**: N_{p1} is the number of positive answers by Coder *1*, N_{p2} is the number of positive answers by Coder *2*, and *N* is the total number of cases analyzed

Table 2. Goodman and Kruskal's Chance Agreement (ac) (Average Interpretation) as a Function of Two Distributions*

		Distribution *1:* Positive Findings by Coder *1* (N_{p2}/N) in %**										
		0	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
Distribution *2:* Positive Findings by Coder *2* (N_{p2}/N) in % **	100	50.0	55.0	60.0	65.0	70.0	75.0	80.0	85.0	90.0	95.0	100.0
	90	55.0	50.0	55.0	60.0	65.0	70.0	75.0	80.0	85.0	90.0	95.0
	80	60.0	55.0	50.0	55.0	60.0	65.0	70.0	75.0	80.0	85.0	90.0
	70	65.0	60.0	55.0	50.0	55.0	60.0	65.0	70.0	75.0	80.0	85.0
	60	70.0	65.0	60.0	55.0	50.0	55.0	60.0	65.0	70.0	75.0	80.0
	50	75.0	70.0	65.0	60.0	55.0	50.0	55.0	60.0	65.0	70.0	75.0

	Distribution *1:* Positive Findings by Coder *1* (N_{p2}/N) in %**										
	0	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
40	80.0	75.0	70.0	65.0	60.0	55.0	50.0	55.0	60.0	65.0	70.0
30	85.0	80.0	75.0	70.0	65.0	60.0	55.0	50.0	55.0	60.0	65.0
20	90.0	85.0	80.0	75.0	70.0	65.0	60.0	55.0	50.0	55.0	60.0
10	95.0	90.0	85.0	80.0	75.0	70.0	65.0	60.0	55.0	50.0	55.0
0	100.0	95.0	90.0	85.0	80.0	75.0	70.0	65.0	60.0	55.0	50.0

*: Main cell entries are Goodman and Kruskal's Chance Agreement (a_c) in %.

**: N_{p1} is the number of positive answers by Coder *1*, N_{p2} is the number of positive answers by Coder *2*, and N is the total number of cases analyzed.

A cell-by-cell comparison of Table 1 with Table 2 shows that λi's a_c is always larger than or equal to λa's a_c, making λi more conservative. Thus, λa is placed above λi in both hierarchies of Table 3. A further comparison between Table 2 and the corresponding table for Scott's π (cf. Table 19.3 in Zhao et al., 2013) shows that the chance-agreement estimate for λ_a is always greater than or equal to that for Scott's π. This indicates that λ_a is more conservative than π. Accordingly, λa is placed below π in both hierarchies of Table 3.

These analyses position the two indices derived from Goodman and Kruskal (1959) at the conservative ends of the expanded 23-index hierarchies, with λi at the very bottom in Table 3. Readers may compare Table 1 or Table 2 with the corresponding tables reported by Zhao et al. (2013) to verify that λa and λi are more conservative than the other indices.

This exercise expands the two 22-index hierarchies to make two 23-index hierarchies, which are shown in Table 3. The two hierarchies are *linked*, that is, the relative positions between the two hierarchies can be compared. For example, by placing *Ir* higher in Hierarchy 1 than *β* in Hierarchy 2, the table indicates that *Ir* is more liberal than *β,* even though the two indices never appear within the same hierarchy. Accordingly, Hierarchies 1&2 may be seen as two parts of one hierarchy. Table 3 assumes two coders and binary scale and large enough sample. When the number of categories increases to three or more, S, Ir, their equivalents, and AC1 may become more liberal. By contrast, when the sample size falls to 20 or below, Krippendorff's α may become highly liberal (Zhao et al., 2013, 2018, 2022).

Table 3. Two Liberal-Conservative Hierarchies Based on Mathematical Analyses of Reliability Indices* (modified from Zhao et al. (2013))

	Hierarchy 1	**Hierarchy 2**
More *liberal* estimates of reliability.	Percent Agreement (a_o) (pre 1901),	Percent Agreement (a_o) (pre 1901),
	Osgood's (1959) index, Holsti's *CR* (1969),	Osgood's (1959) index, Holsti's *CR* (1969)
More *conservative* estimates of reliability.	Rogot & Goldberg's A_1 (1966)	Rogot & Goldberg's A_1 (1966)
	Perreault & Leigh's I_r (1989)	
	Gwet's AC_1 (2008, 2010)	
	Guttman's ρ (1946),	
	Bennett et al.'s *S* (1954),	
	Guilford's *G* (1961), Maxwell's *RE* (1977),	
	Jason & Vegelius' *C* (1979),	
	Brennan & Prediger's k_n (1981),	
	Byrt et al.'s *PABAK* (1993)	
	Potter & Levine-Donnerstein's *rdf-Pi* (1999).	
		Cohen's κ (1960)
		Rogot & Goldberg's A_2 (1966)
	Krippendorff's α (1970, 1980)	Krippendorff's α (1970, 1980)
	Scott's π (1955),	Scott's π (1955),
	Siegel & Castellan's *Rev-K* (1988),	Siegel & Castellan's *Rev-K* (1988),
	Byrt et al's BAK (1993)	Byrt et al's BAK (1993)
	Goodman & Kruskal's λ_a (1954)	Goodman & Kruskal's λ_a (1954)
	Goodman & Kruskal's λ_i (1954)	Goodman & Kruskal's λ_i (1954)

* The two hierarchies assume binary scale, two coders, and sufficiently large sample. Comparisons across the dotted lines are between the general patterns in situations that are more frequent and more important for typical research, e.g., when indices are zero or above, and when the distribution estimates of two coders are not extremely skewed in opposite directions. Comparisons involving Guttman's ρ, its eight equivalents, and Perreault & Leigh's I_r assume binary scale. Comparisons involving Krippendorff's α assume sufficiently large sample.

Table 3 assumes two coders and binary scale and large enough sample. When the number of categories increases to three and beyond, *S, Ir,* their equivalents, and *AC1* can become more liberal; when a sample reduces to 20 or below, Krippendorff's α can become very liberal (Zhao et al., 2018; Zhao et al., 2013).

Six Simulation-Based Hierarchies

The two expanded hierarchies above are based on mathematical analysis. Two cells are separated by a solid line only if the index(es) in an upper cell is usually larger and never smaller than the index(es) in the lower cell. When one index is sometimes larger and sometimes smaller than the other, a judgment call must be made as to whether the two should be placed in the same cell, in two cells separated by dotted lines, or into different hierarchies.

A data-based hierarchy may help reduce uncertainty by showing whether one index consistently yields higher scores than another, thus identifying it as more liberal. Additionally, mathematical analysis is more efficient with simpler situations and is limited to binary scales. As complexity increases, for example when the number of categories increases from two to three, the mathematical analysis becomes exponentially more complex and prone to errors. A data-based approach is more efficient for handling these more complicated situations.

Accordingly, a Monte Carlo simulation was performed to build a liberal-conservative hierarchy based on data. We manipulated two between-subjects factors, i.e., the number of categories (three levels, i.e., 2, 5 and 9 categories) and sample sizes (12 levels, i.e., 10, 13, 14, 15, 16, 17, 20, 25, 30, 100, 500, and 2,000). Each condition has 5,000 contingency tables, so the total sample size is 180,000. Eight intercoder reliability indices were derived from each contingency table.

Multiple comparisons with Tukey's HSD were conducted to examine the mean differences among the indices. This procedure allows us to detect which index yields higher values and which gives lower values. In Table 4, an index is placed in a higher cell than another when it produces a higher mean value and the difference is statistically significant at $p < .05$. Indices are placed in the same cell when the difference between their mean values is not statistically significant.

Table 4. Six Liberal-Conservative Hierarchies Based on Monte Carlo Simulation

	Hierarchy 3	**Hierarchy 4**	**Hierarchy 5**	**Hierarchy 6**	**Hierarchy 7**	**Hierarchy 8**
	C=2; *N* = 2,000 Distribution restricted to 45%~55%	*C*=2; *N* = 2,000; un-restricted; distribution	*C*=2; *N* = 12 levels; un-restricted; distribution	*C*=5; *N* = 12 levels; un-restricted; distribution	*C*=9; *N* = 12 levels; un-restricted; distribution	*C* = 2, 5 & 9; *N* = 12 levels; un-restricted; distribution
Most Liberal	a_o	a_o	a_o	a_o	I_r	a_o
Most Conservative	I_r	I_r κ	I_r κ	I_r	a_o	I_r κ

Hierarchy 3	Hierarchy 4	Hierarchy 5	Hierarchy 6	Hierarchy 7	Hierarchy 8
	AC_1	AC_1	κ, AC_1	κ, AC_1, S	AC_1
	S	S	S		S
		α			α
π, α, κ, AC_1, S	π, α	π	π, α	π, α	π
λ_a	λ_a	λ_a	λ_a	λ_a	λ_a
λ_i	λ_i	λ_i	λ_i	λ_i	λ_i

When the analysis is based on the entire data, which includes all three levels of categories (*K*=2, 5&9*)*, 12 sample sizes (*N*=12 levels), and unrestricted distribution, percent agreement (*a*o) is the most liberal, followed by *Ir*, κ, *AC1* and *S*. Goodman and Kruskal's λi is consistently the most conservative. Krippendorff's α and Scott's π are in between.

With nine or five categories, a similar pattern emerged. With nine categories, the differences among κ, *AC1*, and *S*, and between π and α are no longer statistically significant. In addition, Ir appears even more liberal than percent agreement, a point we discuss further in a later section. With five categories, the differences between κ and*d A*C1, and between π and α are not statistically significant.

With two categories and 2,000 cases, the only difference that is not statistically significant is between π andα. Since some indices likeκare very sensitive to the skewness of marginal distributions, we ran another test after restricting distribution to within .45~.55. Under this condition, the differences between κ, A*C1,* π, α and Sare not statistically significant. This shows that these five indices are very similar to each other with two categories, moderately even distributions, and sufficiently large samples.

At the "starting line"—two categories, an infinitely large sample, and a 50-50% distribution—these indices are equal. As the number of categories increases while other factors remain unchanged, S and AC1 increase rapidly, while the other indices lag. When the distribution becomes more skewed from the starting line, π, κ, and α decrease, AC1 increases, and S remains stable. When the sample size decreases from the starting line, α increases while other indices remain unchanged.

In general, percent agreement is the most liberal, while λi is the most conservative. This is because percent agreement assumes no chance agreement, yet the chance agreement of λi always chooses the largest marginal distribution. The values of α are higher than those of π, lower than those of *S*. The values of *AC1* are usually between κ and *S*. *Ir* is the second most liberal index next to percent agreement. *Ir*, *S* and *AC1* are influenced by the number of categories, so their reliability values will vary with the change of number of categories. The difference between π and α is negligible when sample sizesget very large. Since κ, πandαare dependent on distribution skews, they become indistinguishable from Sand A*C1*whenmarginal distributions get even.

Non-adjusted, Category-based, and Distribution-based Indices

Comparing the hierarchies in Table 3 and Table 4, a pattern emerges: non-adjusted indices like percent agreement (ao) are the most liberal, distribution-based indices are the most conservative, and category-based indices are in between. Gwet's AC1, a double-based index, is closer to category-based indices. This pattern aligns with the underlying assumptions of the indices. Non-adjusted indices assume no chance agreement, making them the most liberal. Distribution-based indices assume that skewed distributions increase chance agreement, leading to more conservative values. Category-based indices assume that more categories reduce chance agreement, even if empty, making them more liberal than distribution-based indices.

As expected, there are differences between the mathematics-based and the simulation-based hierarchies. Most notably, Cohen's κ occupies more liberal positions in the simulation-based Hierarchies 3~8 than in the mathematics-based Hierarchies 1 & 2. This is due to κ's unique individual quota assumption, which expects low chance agreement with "contrasting skews" (e.g., one coder reports 80% positive while the other reports 80% negative). In extreme cases, such as one coder reporting 100% positive and the other 0% positive, κ expects no chance agreement removal, leading to higher κ values. In our Monte Carlo simulation, data are generated randomly, resulting in an equal number of "contrasting skews" and "congruent skews" (where both coders' distributions are skewed in the same direction). In their mathematical analysis, Zhao et al. (2012) placed less weight on contrasting skews, which are less common in typical research, leading to a more conservative placement of κ in the mathematics-based hierarchies.

Based on mathematical analysis, α is listed as more liberal than π in both hierarchies in Table 3. Based on simulated data, α is also listed as more liberal than π in two of the five hierarchies in Table 4. The difference, however, is not statistically significant in the other four hierarchies. The phenomenon is not surprising. In observed or simulated data, as sample size (compare Hierarchies 5 with 3 or 4) or categories (compare Hierarchies 8 with 6 or 7) increase, the differences between the two indices can become so small that they are not statistically significant. But in mathematical analysis, a tiny difference is still a difference, so α is still listed as more liberal than π.

A New Paradox for Perreault & Leigh's *Ir*

Perreault and Leigh (1989) produced an interesting standout in Hierarchy 4, where *Ir* appears even more liberal than percent agreement *ao*. *Ir* was designed to adjust for, which means to remove, chance agreement. Removing chance agreement is not supposed to make a reliability index larger, hence a new paradox to be added to the 22nd paradox listed by Zhao et al. (2012):

Paradox 23: Reliability index appears larger after removing random chance agreement.

The paradox is due to the combined effects of three assumptions behind Ir, maximum randomness, categories reduce chance agreements, and index needs to be elevated (Zhao et al., 2012).

In the traditional approach, the maximum randomness assumption acts as a double-edged sword. While it suppresses the index by subtracting the maximum chance agreement (*ac*) from the numerator in Equation 3, it also inflates the index by subtracting the same ac from the denominator, shortening the reference scale. On its own, this assumption would not result in a chance-adjusted index larger than *ao*.

However, *Ir* is also category-based, and its chance estimation reduces quickly as the number of categories increases, even when additional categories are unused, which further inflates the index. Yet, this alone does not make *Ir* surpass *ao*. The key factor is the assumption that the index should be elevated. In Ir, this elevation is achieved by taking the square root of S, which ultimately allows *Ir* to exceed *ao*.

Although in Table 4 the paradox appears only in Hierarchy 7, where there are nine categories, the paradox can happen in many other situations, and the underlying problem is more pervasive. Our simulation skipped *K*=6 through *K*=8 and stopped at *K*=9. The paradox did not show up in Hierarchies 6 (*K*=5) and 8 (*K*=2, 5 & 9) not because *Ir* never passed *ao*, but because *Ir* did not pass *ao* far enough or often enough to make the average *Ir* larger than average *ao*. But even at *K*=4 or *K*=3, *Ir* can be larger than *ao*. To verify, set *K*=3, and *ao* any number larger than 0.5 but smaller than 1, and calculate $Ir = ((1 - 1/K) / (ao - 1/K))^2$. This analysis confirms that the line between *ao* and *Ir* in Table 12 in Zhao et al. (2013) should indeed be dotted. In other words, while ao is more often larger *th*an *Ir*, *Ir* is also very oft*en* l*ar*ger than *ao*.

This is not just about one index slightly larger or smaller than another. Zhao et al. (2012) recommended not to use *Ir* or S when there are three or more categories (Ejbye-Ernst, 2023). The newly discovered paradox and behavior of *Ir* seem to suggest not to use *Ir* even with two categories, where *Ir* never exceeds *ao*, *Ir* may be overly inflated. *Ir* has no advantage over S under any circumstance, and in almost all circumstances S is a more reasonable alternative to *Ir*, the drawbacks of S notwithstanding. The only exceptions are when S = 0 or S = 1, in which case S = *Ir*.

Conclusion

This study extends Zhao et al.'s (2013) work by using simulated data to construct six additional hierarchies. These hierarchies show substantial similarities with the earlier mathematics-based hierarchies, while also revealing important differences. Some preliminary findings of this study were reported in a 2012 conference presentation (Zhao, 2012). We hope these hierarchies together will prove useful to workbench researchers who wish to better evaluate the inter-coder reliability indices of their instruments.

Findings from the mathematical analysis and the simulation are consistent with each other on some main points. Between groups of the indices, the non-adjusted indices tend to be the most liberal, the distribution-based indices tend to be the most conservative, and the category-based indices tend to be somewhere in between. Between the individual indices, percent agreement tends to be the most liberal, while Goodman and Kruskal's λr tends to be the most conservative. For the distribution-based indices, the hierarchy generally places κ as the most liberal, followed by α, π, and λ_r as the most conservative.

Three discrepancies emerged. First, κ appears more liberal in simulation than in mathematical analysis. Second, α appears more liberal than π in mathematical analysis while the two appeared tied statistically in the simulation analysis. Third, *Ir* appears more liberal in simulation than in mathematical analysis. Our analysis shows that mathematical analysis is more precise in the first two discrepancies, while in the third discrepancy simulation filled a gap in the mathematical analysis.

Researchers want their reliabilities to look high, and they have many indices to choose from. Two newer indices, Perreault & Leigh's *Ir* and Gwet's *AC1*, are gaining in popularity in part because they tend to produce higher numbers than other indices. Knowing that they are among the most liberal, we hope, would encourage the

researchers, reviewers and editors to be more cautious when using or interpreting the two indices. On the other hand, we should not equate low estimate with rigor, or complex formulas with sophistication. More specifically, reviewers should not require λr, π or α just for its low estimates or complicated equations. We also should not require or encourage universal application of α just because it has been repeatedly advocated.

For criminology, the contribution of this study is therefore limited but consequential. It does not propose a uniquely criminological reliability coefficient, and it does not replace the need for transparent codebooks, coder training or clear reporting of coding procedures. Its contribution is narrower: to show that commonly used reliability estimators stand in a patterned liberal–conservative relation to one another, and that this ordering gives criminological researchers one principled basis for selecting and interpreting coefficients in interview-based, observational, visual and open-source research. In a field where reliability remains unevenly reported yet is assessed through multiple non-equivalent coefficients, that clarification is methodologically useful in its own right (Copes et al., 2020; Liu, 2021).

A central need, therefore, is to develop criteria that can help researchers evaluate which index is more appropriate or accurate for a given research context. Ultimately, we need a new index(es) based on more realistic assumptions about coder behaviour. These assumptions should include 1) coders sometimes code randomly, which leads to random agreements that need to be removed; 2) the random coding is not deliberate or purposeful, therefore chance agreement is not a function of category or distribution; 3) the random coding is involuntary depending on difficulty of the task; a more difficult task produces more chance agreement, a less difficult task produces less chance agreement, a non-difficult task produces no chance agreement.

Disclosure Statement

The authors declare no competing interests.

Data Availability Statement

The datasets generated and analyzed during the current study are fully simulated and do not contain any real-world or proprietary data. They are available from the authors upon reasonable request for non-commercial research purposes.

References

[1] Adams, E. A., Morash, M., Smith, S. W., & Cobbina, J. E. (2017). Women's Experience of Motherhood, Violations of Supervision Requirements and Arrests. *The British Journal of Criminology*, *57*(6), 1420–1441. https://doi.org/10.1093/bjc/azw092

[2] Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communications Through Limited-Response Questioning*. *Public Opinion Quarterly*, *18*(3), 303–308. https://doi.org/10.1086/266520

[3] Brent, J. J., Kraska, P. B., & Hutchens, J. (2024). Qualitative Interviewing. In *Oxford Research Encyclopedia of Criminology and Criminal Justice* (pp. 0). Oxford University Press. https://doi.org/10.1093/acrefore/9780190264079.013.850

[4] Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*(5), 423–429. https://doi.org/10.1016/0895-4356(93)90018-V

[5] Charles Feng, G., & Zhao, X. (2016). Do Not Force Agreement. *Methodology*, *12*(4), 145–148. https://doi.org/10.1027/1614-2241/a000120

[6]

Chermak, S. M., Freilich, J. D., Greene-Colozzi, E., & Klein, B. R. (2025). Open-Source Research in Criminology and Criminal Justice. *Annual Review of Criminology*, *8*(Volume 8, 2025), 141–170. https://doi.org/10.1146/annurev-criminol-022422-013842

[7] Chmura Kraemer, H., Periyakoil, V. S., & Noda, A. (2002). Kappa coefficients in medical research. *Statistics in medicine*, *21*(14), 2109–2129. https://doi.org/10.1002/sim.1180

[8] Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 551–558. https://doi.org/10.1016/0895-4356(90)90159-M

[9] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

[10] Copes, H., Beaton, B., Ayeni, D., Dabney, D., & Tewksbury, R. (2020). A Content Analysis of Qualitative Research Published in Top Criminology and Criminal Justice Journals from 2010 to 2019. *American Journal of Criminal Justice*, *45*(6), 1060–1079. https://doi.org/10.1007/s12103-020-09540-6

[11] de Vet, H. C. W., Terwee, C. B., Knol, D. L., & Bouter, L. M. (2006). When to use agreement versus reliability measures. *Journal of Clinical Epidemiology*, *59*(10), 1033–1039. https://doi.org/10.1016/j.jclinepi.2005.10.015

[12] Delgado, R., & Tibau, X.-A. (2019). Why Cohen's Kappa should be avoided as performance measure in classification. *PloS one*, *14*(9), e0222916. https://doi.org/10.1371/journal.pone.0222916

[13] Dettori, J. R., & Norvell, D. C. (2020). Kappa and Beyond: Is There Agreement? *Global Spine Journal*, *10*(4), 499–501. https://doi.org/10.1177/2192568220911648

[14] Ejbye-Ernst, P. (2023). Does Third-Party Intervention Matter? A Video-Based Analysis of the Effect of Third-Party Intervention on the Continuation of Interpersonal Conflict Behaviour. *The British Journal of Criminology*, *63*(1), 78–96. https://doi.org/10.1093/bjc/azab121

[15] Feng, G. C. (2014). Estimating intercoder reliability: a structural equation modeling approach. *Quality & Quantity*, *48*(4), 2355–2369. https://doi.org/10.1007/s11135-014-0034-7

[16] Fleiss, J. L. (1975). Measuring Agreement between Two Judges on the Presence or Absence of a Trait. *Biometrics*, *31*(3), 651–659. https://doi.org/10.2307/2529549

[17] Freelon, D. G. (2010). ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science*, *5*(1), 20–33.

[18] Goodman, L. A., & Kruskal, W. H. (1959). Measures of Association for Cross Classifications. II: Further Discussion and References. *Journal of the American statistical association*, *54*(285), 123–163. https://doi.org/10.1080/01621459.1959.10501503

[19] Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British journal of mathematical and statistical psychology*, *61*(1), 29–48. https://doi.org/10.1348/000711006X126600

[20] Han, T., Zhang, L., Zhao, X., & Deng, K. (2023). Total-effect Test May Erroneously Reject So-called "Full" or "Complete" Mediation. *arXiv preprint arXiv:2309.08910*.

[21] Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664

[22] Hoehler, F. K. (2000). Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology*, *53*(5), 499–503. https://doi.org/10.1016/S0895-4356(99)00174-2

[23] Jakubauskaite, V. (2021). *Development of an observational coding scheme for parental behaviour in a play-based instructional context with 5-to 7-year-olds*. University of Kent (United Kingdom).

[24] Jiang, Y., Zhao, X., Zhu, L., Liu, J. S., & Deng, K. (2021). TOTAL-EFFECT TEST IS SUPERFLUOUS FOR ESTABLISHING COMPLEMENTARY MEDIATION. *Statistica Sinica*, *31*(4), 1961–1983.

[25] Ju, Q. R., Zhang, L., Zhao, X., & Li, D. M. (2026). Is marrying up better for mental health? Educational assortative mating, marital well-being, subjective socioeconomic status, and depressive symptoms among Chinese adults: Evidence from cross-lagged panel networks. *Journal of Affective Disorders*, *405*, 121663. https://doi.org/10.1016/j.jad.2026.121663

[26] Kokkalera, S. S., Gonzalez, C. M. F., & Williams, J. M. (2023). Introduction to the Special Issue on Qualitative Criminology and Victimology. *Crime & Delinquency*, *69*(2), 259–266. https://doi.org/10.1177/00111287221134912

[27] Kort-Butler, L. A. (2016). Content Analysis in the Study of Crime, Media, and Popular Culture. In *Oxford Research Encyclopedia of Criminology and Criminal Justice* (pp. 0). Oxford University Press. https://doi.org/10.1093/acrefore/9780190264079.013.23

[28] Kraemer, H. C. (1979). Ramifications of a Population Model for κ as a Coefficient of Reliability. *Psychometrika*, *44*(4), 461–472. https://doi.org/10.1007/BF02296208

[29] Kraemer, H. C. (1992). Measurement of reliability for categorical data in medical research. *Statistical Methods in Medical Research*, *1*(2), 183–199. https://doi.org/10.1177/096228029200100204

[30] Krippendorff, K. (1970). Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement*, *30*(1), 61–70. https://doi.org/10.1177/001316447003000105

[31] Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology.* . Sage publications.

[32] Krippendorff, K. (2004). Reliability in Content Analysis. *Human Communication Research*, *30*(3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

[33] Krippendorff, K. (2016). Misunderstanding Reliability. *Methodology*, *12*(4), 139–144. https://doi.org/10.1027/1614-2241/a000119

[34] Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

[35] Krippendorff, K. (2019). The changing landscape of content analysis: Reflections on social construction of reality and beyond. *Communication & Society*, *47*(604), 1–27.

[36] Li, D., Yi, Q., & Andrews, B. (2018). An Evaluation of Rater Agreement Indices Using Generalizability Theory. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar, *Quantitative Psychology* Cham.

[37] Li, D. M., Zhang, H. L., & Ju, Q. R. (2025). Statistical Significance, Narrative, and the Scholastic Fallacy: How Ritualized Statistics Exaggerate Social Science Theories. *Transformative Society*, *1*(2), 39–62. https://doi.org/10.63336/TransSoc.28

[38] Liu, J. (2021). Asian Criminology and Non-Western Criminology: Challenges, Strategies, and Directions. *International Annals of Criminology*, *59*(2), 103–118. https://doi.org/10.1017/cri.2021.16

[39] Liu, J., & Li, D. M. (2024). Is Machine Learning Really Unsafe and Irresponsible in Social Sciences? Paradoxes and Reconsideration from Recidivism Prediction Tasks. *Asian Journal of Criminology*, *19*(2), 143–159. https://doi.org/10.1007/s11417-024-09429-x

[40] Liu, J., Li, D. M., Ju, Q. R., & Zhang, X. S. (2025). Structuring Macau's Criminal Court Judgments with Large Language Models: Methodological Innovations for Data Accuracy and Sample Selection Bias. *Asian Journal of Criminology*, *21*(1), 14. https://doi.org/10.1007/s11417-025-09475-z

[41] Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research*, *28*(4), 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

[42] Nili, A., Tate, M., Barros, A., & Johnstone, D. (2020). An approach for selecting and using a method of inter-coder reliability in information management research. *International Journal of Information Management*, *54*, 102154. https://doi.org/10.1016/j.ijinfomgt.2020.102154

[43] Perreault, W. D., & Leigh, L. E. (1989). Reliability of Nominal Data Based on Qualitative Judgments. *Journal of Marketing Research*, *26*(2), 135–148. https://doi.org/10.1177/002224378902600201

[44] Popping, R. (1988). On Agreement Indices for Nominal Data. In W. E. Saris & I. N. Gallhofer (Eds.), *Sociometric Research: Volume 1 Data Collection and Scaling* (pp. 90–105). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-19051-5_6

[45] Riffe, D., Lacy, S., Watson, B. R., & Lovejoy, J. (2023). *Analyzing media messages: Using quantitative content analysis in research*. Routledge.

[46] Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, *7*(3), 301–317. https://doi.org/10.1177/096228029800700306

[47] Silveira, P. S. P., & Siqueira, J. O. (2023). Better to be in agreement than in bad company. *Behavior research methods*, *55*(7), 3326–3347. https://doi.org/10.3758/s13428-022-01950-0

[48] Stone, R. (2016). Desistance and Identity Repair: Redemption Narratives as Resistance to Stigma. *The British Journal of Criminology*, *56*(5), 956–975. https://doi.org/10.1093/bjc/azv081

[49] Stütz, S., Berding, F., Reincke, S., & Scheper, L. (2022). Characteristics of learning tasks in accounting textbooks: an AI assisted analysis. *Empirical Research in Vocational Education and Training*, *14*(1), 10. https://doi.org/10.1186/s40461-022-00138-2

[50] ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2018). On the Usefulness of Interrater Reliability Coefficients. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar, *Quantitative Psychology* Cham.

[51] Tong, F., Tang, S., Irby, B. J., Lara-Alecio, R., & Guerrero, C. (2020). The determination of appropriate coefficient indices for inter-rater reliability: Using classroom observation instruments as fidelity measures in large-scale randomized research. *International Journal of Educational Research*, *99*, 101514. https://doi.org/10.1016/j.ijer.2019.101514

[52] Vach, W. (2005). The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology*, *58*(7), 655–661. https://doi.org/10.1016/j.jclinepi.2004.02.021

[53] van der Knaap, L. M., Leenarts, L. E. W., Born, M. P., & Oosterveld, P. (2012). Reevaluating Interrater Reliability in Offender Risk Assessment. *Crime & Delinquency*, *58*(1), 147–163. https://doi.org/10.1177/0011128710382347

[54] von Eye, A., & von Eye, M. (2008). On the Marginal Dependency of Cohen's κ. *European Psychologist*, *13*(4), 305–315. https://doi.org/10.1027/1016-9040.13.4.305

[55] Warrens, M. J. (2014). New Interpretations of Cohen's Kappa. *Journal of Mathematics*, *2014*(1), 203907. https://doi.org/10.1155/2014/203907

[56] Weenink, D., Dhattiwala, R., & van der Duin, D. (2022). Circles of Peace. A Video Analysis of Situational Group Formation and Collective Third-Party Intervention in Violent Incidents. *The British Journal of Criminology*, *62*(1), 18–36. https://doi.org/10.1093/bjc/azab042

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

[58] Zec, S., Soriani, N., Comoretto, R., & Baldi, I. (2017). High agreement and high prevalence: the paradox of Cohen's kappa. *The open nursing journal*, *11*, 211.

[59] Zhao, X. (2011). When to use Scott's π or Krippendorff's α, if ever. Annual Conference of Association for Education in Journalism and Mass Communication, St. Louis, MO, USA,

[60] Zhao, X. (2012). *A Reliability Index (ai) that assumes honest coders and variable randomness* Annual conference of Association for Education in Journalism and Mass Communication, Chicago.

[61] Zhao, X., Feng, G. C., Ao, S. H., & Liu, P. L. (2022). Interrater reliability estimators tested against true interrater reliabilities. *BMC medical research methodology*, *22*(1), 232. https://doi.org/10.1186/s12874-022-01707-5

[62] Zhao, X., Li, D. M., Lai, Z. Z., Liu, P. L., Ao, S. H., & You, F. (2024). Percentage Coefficient (bp)--Effect Size Analysis (Theory Paper 1). *arXiv preprint arXiv:2404.19495*. https://doi.org/10.48550/arXiv.2404.19495

[63] Zhao, X., Ju, Q. R., Liu, P. L., Li, D. M., Zhang, L., Ye, J. F., Ao, S. H., & Li, M. M. (2025). Intellectual Up-streams of Percentage Scale (ps) and Percentage Coefficient (bp)--Effect Size Analysis (Theory Paper 2). *arXiv preprint arXiv: 2507.13695*. https://doi.org/10.48550/arXiv.2507.13695

[64] Zhao, X., Feng, G. C., Liu, J. S., & Deng, K. (2018). We Agreed to Measure Agreement-Redefining Reliability De-justifies Krippendorff's Alpha. *China Media Research*(2).

[65] Zhao, X., Liu, J. S., & Deng, K. (2012). 19 Assumptions behind Intercoder Reliability Indices. *Communication Yearbook 36*, *36*, 419.

[66] Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind intercoder reliability indices. *Annals of the International Communication Association*, *36*(1), 419–480.

[67] Zwick, R. (1988). Another look at interrater agreement. *Psychological bulletin*, *103*(3), 374.