Journal of Statistical Modelling: Theory and Applications
Vol. 6, No. 1, 2025, pp. 185-192
Yazd University Press 2025

Research Paper

Estimation of measurement error in binary data in surveys

Roshanak Alimohammadi[*]
Department of Statistics, Faculty of Mathematical Sciences, Alzahra University, Tehran, Iran

Received: August 18, 2025/ Revised: January 24, 2026/ Accepted: January 27, 2026

Abstract: Measurement error is an inherent and unavoidable component of nonsampling error in surveys, and its estimation is essential for assessing the quality of survey results. This paper investigates two agreement criteria, Cohen's Kappa and Gwet's AC1, for quantifying measurement error in binary survey data. A theoretical comparison is first presented to highlight the conceptual differences between the two coefficients. This is followed by a simulation study in which true binary values are generated under varying prevalence levels, and observed responses are obtained through controlled misclassification mechanisms characterized by specified levels of observed agreement. The performance and empirical variances of both agreement measures are evaluated and the results demonstrate that Gwet's AC1 provides more stable agreement estimates and variance behavior than Cohen's Kappa, particularly under conditions of extreme prevalence.

Keywords: Agreement; Cohen's Kappa; Gwet's AC1; Measurement error; Prevalence.
Mathematics Subject Classification (2010): 62D05, 62P99.

1 Introduction

Measurement error constitutes a major component of nonsampling error in survey studies and can substantially affect the precision and quality of survey estimates. Modeling of measurement error provides a framework for quantifying this error and assessing the precision of survey results. Early work by Kish (1962), Biemer and Stokes (1991), and Biemer and Trewin (1997) employed analysis of variance (ANOVA) techniques to model

[*] Corresponding author: r_alimohammadi@alzahra.ac.ir

measurement error in survey data. More recently, Alimohammadi and Navvabpour (2008) proposed a measurement error model for continuous data arising from face to face surveys.

Modeling measurement error for categorical data presents challenges that differ fundamentally from those encountered with continuous variables. In particular, measurement error in binary survey data manifests as misclassification, where the observed response does not correspond to the underlying true status. Alimohammadi (2011) introduced a modeling framework specifically designed to address measurement error in binary data. In the context of binary surveys, quantifying measurement error requires appropriate criteria that capture the agreement between observed responses and their true values. Agreement coefficients provide a natural approach for this purpose. Among these, Cohen's Kappa and Gwet's AC1 are widely used measures of agreement. While Cohen's Kappa has a long history of application, it is well documented that its value is sensitive to marginal distributions and prevalence imbalance, which are common features of survey data.

In this paper, we focus on the use of Cohen's Kappa and Gwet's AC1 as criteria for quantifying measurement error in binary survey data. We provide a theoretical comparison of these agreement measures and conduct a simulation study to examine their behavior and standard errors under varying prevalence and observed agreement scenarios. The results clarify the relative suitability of these coefficients for assessing measurement error in binary surveys.

The remainder of this paper is organized as follows. Section 2 introduces the theoretical framework, including the 2×2 contingency table structure and formal definitions of observed agreement, misclassification probabilities, Cohen's Kappa, and Gwet's AC1. Section 3 describes the simulation design, including data generation mechanisms and simulation results and compares the empirical behavior of the agreement measures. Section 4 discusses the main findings and practical implications for survey measurement error assessment.

2 Criteria for quantifying measurement error

Let $Y \in \{0, 1\}$ denote the true value of a binary survey variable and $\tilde{Y} \in \{0, 1\}$ its observed counterpart. The joint distribution of $(Y, \tilde{Y})$ can be represented by a 2×2 contingency table with cell probabilities $\pi_{ij} = \Pr(Y = i, \tilde{Y} = j)$, for $i, j \in \{0, 1\}$. The marginal probabilities are denoted by $\pi_{i.} = \sum_j \pi_{ij}$ and $\pi_{.j} = \sum_i \pi_{ij}$ (Table 1).

Table 1: Probability distribution of true and observed values for binary data.

		$\tilde{Y}$=Response Value		
		0	1	total
Y = True Value	0	π_{00}	π_{01}	$\pi_{0.}$
	1	π_{10}	π_{11}	$\pi_{1.}$
total		$\pi_{.0}$	$\pi_{.1}$	

In binary data, measurement error appears as disagreement between the observed response and the true value. A natural measure of this agreement is accuracy, defined as the probability that the observed and true values coincide. In this paper, accuracy corresponds to the observed agreement and is denoted at the population level by $\Pr(\tilde{Y} = Y) = \theta_0 = \pi_{00} + \pi_{11}$, with the corresponding sample estimator given by $P_0 = (n_{00} + n_{11})/n$, where $n_{ij},\ i, j \in \{0, 1\}$, denotes the number of units for which the true value is i and the observed response is j, and n is the sample size.

Prevalence plays a central role in determining the behavior of agreement measures. Prevalence is defined as $\pi = \Pr(Y = 1)$, which characterizes the marginal distribution of the true responses. Extreme values of π (close to 0 or 1) correspond to situations where one category is rare, a common feature of survey data. Throughout this paper, the term prevalence refers solely to the distribution of the true binary values, and does not include the observed responses affected by measurement error.

In this paper, two agreement criteria, Kappa and Gwet's AC1, are applied to evaluate the degree of concordance between observed and true values.

Cohen (1960) supposed Kappa coefficient, and Fliess and Cohen (1973) introduced weighted Kappa coefficient to assess agreement between two raters. The weighted kappa coefficient is a generalization of the simple Kappa coefficient, using weights to consider the relative difference between categories. The weights w_{ij} are constructed such that $w_{ii} = 1$ for all i and j, and $w_{ij} = w_{ji}$. The weighted Kappa coefficient is defined as

$$K_w = \frac{\theta_{0w} - \theta_{ew}}{(1 - \theta_{ew})},$$

where $\theta_{0w} = \Sigma_i \Sigma_j w_{ij} \pi_{ij}$ and $\theta_{ew} = \Sigma_i \Sigma_j w_{ij} \pi_{i.} \pi_{.j}$.

Considering Table 1 to define the Kappa coefficient for a binary variable, Cohen's Kappa is described as

$$K = \frac{\theta_0 - \theta_e}{(1 - \theta_e)},$$

where $\theta_0 = \pi_{00} + \pi_{11}$ is known as Observed agreement, and $\theta_e = \pi_{.0}\pi_{0.} + \pi_{.1}\pi_{1.}$.

The second agreement criterion, introduced by Gwet (2001), is Gwet's AC1, and defined as

$$AC1 = \frac{\theta_0 - \theta_e^*}{(1 - \theta_e^*)},$$

where $\theta_0 = \pi_{00} + \pi_{11}$ and $\theta_e^* = 2\pi^*(1 - \pi^*)$, $\pi^* = \frac{1}{2}((\pi_{10} + \pi_{11}) + (\pi_{01} + \pi_{11}))$.

The agreement criteria have been applied to quantify measurement error. Let us consider real values by examining the absolute difference, $|i - j|$, $(i, j = 0, 1$ in Table 1) as measurement error. It can be demonstrated that theses criteria take 1 if and only if $\theta_0 = 1$. Therefore, there is complete agreement between observed and true values. In other words, no measurement error occurs when the value of the criterion is 1. Conversely, these criteria take 0 if and only if $\theta_0 = \theta_e$. This condition is satisfied when agreement occurs purely by chance. The criteria may yield a negative value when the probability of agreement is less than the probability of agreement by chance. Such a scenario indicates a significant amount of measurement error.

Table 2 presents a set of theoretical examples designed to illustrate the conceptual differences between Cohen's Kappa and Gwet's AC1 under varying levels of prevalence and observed agreement (θ_0). Specifically, the table shows that when prevalence is balanced, both Kappa and AC1 yield similar values for a given level of observed agreement. However, as prevalence becomes increasingly imbalanced, Cohen's Kappa decreases substantially even when observed agreement remains high, whereas Gwet's AC1 remains comparatively stable. These examples foreshadow the simulation results presented in Table 3 and highlight why prevalence imbalance plays a central role in the interpretation of agreement coefficients.

Table 2: Theoretical values of Cohen's Kappa and Gwet's AC1

Joint probabilities				Agreement measures		
π_{00}	π_{11}	π_{01}	π_{10}	θ_0	Kappa	AC1
0.10	0.90	0.00	0.00	1.00	1.00	1.00
0.00	0.90	0.00	0.10	0.90	0.00	0.00
0.00	0.80	0.10	0.10	0.80	−0.11	−1.00
0.20	0.80	0.00	0.00	1.00	1.00	1.00
0.10	0.70	0.10	0.10	0.80	0.37	0.00
0.00	0.40	0.30	0.30	0.40	−0.43	−1.00
0.50	0.50	0.00	0.00	1.00	1.00	1.00
0.80	0.00	0.10	0.10	0.80	−0.11	0.78
0.00	0.00	0.50	0.50	0.00	−1.00	−1.00
0.10	0.10	0.40	0.40	0.20	−0.60	−0.60
0.20	0.20	0.30	0.30	0.40	−0.20	−0.20
0.30	0.30	0.20	0.20	0.60	0.20	0.20
0.40	0.40	0.10	0.10	0.80	0.60	0.60
0.00	0.00	0.10	0.90	0.00	−0.22	−1.00
0.10	0.60	0.30	0.00	0.70	−0.29	−0.20
0.20	0.30	0.20	0.30	0.50	0.00	−0.11
0.30	0.20	0.10	0.40	0.50	0.00	−0.11
0.40	0.10	0.00	0.50	0.50	0.14	0.29
0.30	0.20	0.40	0.10	0.50	0.14	0.00
0.40	0.20	0.40	0.00	0.60	0.42	0.20
0.50	0.50	0.00	0.00	1.00	1.00	1.00
0.00	0.00	0.90	0.10	0.00	−0.22	−1.00

To further clarify the behavior of the agreement criteria, we provide a theoretical example based on the contingency tables reported in Table 2. Consider the case with joint probabilities $(\pi_{00}, \pi_{11}, \pi_{01}, \pi_{10}) = (0.8, 0, 0.1, 0.1)$, for which the observed agreement is $\theta_0 = \pi_{00} + \pi_{11} = 0.8$. Despite this relatively high level of observed agreement, Cohen's Kappa yields a negative value ($K = -0.11$), suggesting agreement worse than chance. This result is caused by the strong imbalance in the marginal distributions, which inflates the expected agreement by chance θ_e in the definition of Kappa. As a result, Kappa becomes highly sensitive to prevalence imbalance and may underestimate agreement in such situations. On the other hand, Gwet's AC1 for the same contingency table equals 0.78, reflecting its alternative formulation of expected agreement.

By contrast, when the marginal distributions are balanced, such as in the case $(\pi_{00}, \pi_{11}, \pi_{01}, \pi_{10}) = (0.4, 0.4, 0.1, 0.1)$, Cohen's Kappa ($K = 0.6$), and Gwet's AC1 ($\text{AC1} = 0.6$) yield similar positive values, indicating substantial agreement. This highlights that discrepancies between the two criteria are most pronounced under marginal imbalance and extreme prevalence.

These theoretical examples highlight that discrepancies between the two criteria are intrinsic to their definitions and not due to sampling variability, thereby motivating the simulation study presented in the next section.

2.1 Variance formulas of Cohen's Kappa and Gwet's AC1

Fliess et al. (1969) presented the variance of Cohen's Kappa for a 2×2 contingency table with observed cell proportions $\pi_{ij}, (i, j = 0, 1)$. The variance of Cohen's Kappa

(K) can be approximated as follow

$$Var(K) \approx \frac{\theta_0(1-\theta_0)}{n(1-\theta_e)^2}.$$

For Gwet's AC1, Gwet (2008) show that the variance can be approximated as

$$Var(AC1) \approx \frac{1}{n}\left[\frac{\theta_0(1-\theta_0)}{(1-\theta_e^*)^2}\right],$$

where n denotes the sample size, $\theta_0 = \pi_{00} + \pi_{11}$ and $\theta_e^* = 2\pi^*(1-\pi^*)$, $\pi^* = \frac{1}{2}((\pi_{10} + \pi_{11}) + (\pi_{01} + \pi_{11}))$.

The variance of a reliability coefficient influences its confidence interval and interpretability. While Cohen's Kappa remains an important metric, it is not always reliable in scenarios with unbalanced data. Gwet's AC1 provides a theoretically justified, more stable alternative with superior variance properties. It is recommended for applications involving low prevalence, small sample sizes, or unbalanced marginal totals.

3 Simulation Methodology

This section describes the simulation design used to compare Cohen's Kappa and Gwet's AC1 when observed survey responses are subject to misclassification relative to an underlying true binary status. The simulation framework is explicitly defined to ensure full reproducibility.

3.1 Data-generating process

We consider a binary latent variable $Y \in \{0,1\}$ representing the true status of a survey unit (e.g., presence or absence of a characteristic). For each simulation replication, a sample of size n is generated independently according to a Bernoulli distribution

$$Y_i \sim \text{Bernoulli}(\pi), \quad i = 1, \ldots, n,$$

where $\pi \in (0,1)$ denotes the prevalence, defined as the proportion of units with true value equals 1 in the population. Observed survey responses $\tilde{Y}$ are generated by introducing misclassification relative to the true status. Specifically, conditional on Y, the observed response is defined as

$$\tilde{Y_i} = \begin{cases} Y_i & \text{with probability } \alpha \\ 1 - Y_i & \text{with probability } 1-\alpha, \end{cases}$$

where $\alpha \in (0,1)$ denotes the accuracy of the survey response. Accuracy is defined as the overall probability that the observed response equals the true latent status $\alpha = P(\tilde{Y} = Y)$. In the simulation design, accuracy is defined as $\Pr(\tilde{Y} = Y)$, which coincides with the observed agreement $\theta_0 = \pi_{00} + \pi_{11}$.

3.2 Contingency table construction

For each simulated dataset, a 2×2 contingency table is constructed by cross-tabulating $\tilde{Y}$ and Y. The cell counts correspond to $n_{ij},\ i,j \in \{0,1\}$, the number of units for which $Y = i$ and $\tilde{Y} = j$. Agreement coefficients are computed from this contingency table.

Sample observed agreement: In practice, observed agreement, θ_0, is estimated by the sample observed agreement as $P_o = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\tilde{Y}_i = Y_i)$, where $\mathbb{I}(\cdot)$ is the indicator function. In this setting, P_o is numerically equal to the realized sample accuracy.

3.3 Simulation parameters

A range of prevalence values $\pi \in \{0.05, 0.10, 0.30, 0.50, 0.70, 0.90, 0.95\}$, and observed agreement levels $\theta_0 \in \{0.50, 0.70, 0.80, 0.95\}$ are considered in the simulation study.

Rationale for sample size and replications: For each combination of prevalence and observed agreement, the sample size was set to $n = 1000$, which is consistent with common practice in survey methodology and agreement studies (Fliess et al., 2003; Kish, 1965). Each simulation scenario was replicated $R = 1000$ times to ensure stable estimation of the mean and variance of the agreement coefficients, in line with established recommendations for the design of simulation studies (Burton et al., 2006; Morris et al., 2019). The simulation was repeated $R = 1000$ times to approximate the sampling distributions of the agreement coefficients. For each replication, Cohen's Kappa and Gwet's AC1 were computed, and empirical means and variances were estimated across replications.

Several limitations of this study should be acknowledged. First, the simulation assumes symmetric misclassification, with equal probabilities π_{01} and π_{10} in Table 1. While this assumption simplifies interpretation and consistent with common methodological studies, real-world survey data may exhibit asymmetric error structures. Second, the analysis focuses on binary outcomes; extensions to ordinal or nominal outcomes may reveal additional complexities. Finally, although the simulation parameters were chosen to reflect realistic survey settings, specific applications may involve different sample sizes or error mechanisms.

3.4 Simulation study to compare Cohen's Kappa and Gwet's AC1

In this section, a simulation study is conducted to evaluate the performance of Cohen's Kappa and Gwet's AC1. Table 3 summarizes the results of the simulation study. Each row corresponds to a specific combination of prevalence π and observed agreement θ_0. For each scenario, the table reports the mean and standard deviation of the sample observed agreement P_o, Cohen's Kappa, and Gwet's AC1 across $R = 1000$ replications.

In Table 3, Kappa is the average of Kappa values across $R = 1000$ simulation runs, $\bar{K} = \frac{1}{R}\sum_{i=1}^{R} K_i$, Standard Deviation of Kappa is $K_{SD} = \sqrt{\frac{1}{R-1}\sum_{i=1}^{R}(K_i - \bar{K})^2}$, AC1 is indeed the mean of AC1 as $\overline{AC1} = \frac{1}{R}\sum_{i=1}^{R} AC1_i$ and Standard Deviation of AC1 is $AC1_{SD} = \sqrt{\frac{1}{R-1}\sum_{i=1}^{R}(AC1_i - \overline{AC1})^2}$.

Table 3: Simulation results to compare agreement criteria.

Prevalence	θ_0	P_o	Kappa	K_{SD}	AC1	$AC1_{SD}$
0.01	0.95	0.951	0.310	0.017	0.890	0.012
0.05	0.95	0.951	0.346	0.018	0.902	0.013
0.10	0.80	0.884	0.423	0.029	0.765	0.018
0.30	0.70	0.794	0.512	0.031	0.612	0.027
0.50	0.80	0.800	0.600	0.025	0.600	0.025
0.70	0.70	0.798	0.540	0.032	0.635	0.028
0.90	0.70	0.798	0.490	0.034	0.611	0.029
0.95	0.95	0.951	0.346	0.018	0.902	0.013
0.99	0.95	0.951	0.310	0.017	0.890	0.012

The prevalence parameter π controls the proportion of true positive cases in the population and distinguishes balanced scenarios (e.g., $\pi \approx 0.5$) from imbalanced scenarios (e.g., $\pi \leq 0.10$ or $\pi \geq 0.90$). The reported standard deviations provide empirical estimates of the sampling variability of each agreement coefficient.

The results in Table 3 show that, for a fixed level of observed agreement (θ_0), Cohen's Kappa decreases sharply as prevalence becomes more extreme, while Gwet's AC1 remains relatively stable. This pattern indicates that AC1 more accurately reflects the underlying measurement error process, whereas Kappa is strongly influenced by prevalence imbalance. Additionally, the empirical variance of AC1 is consistently lower than that of Kappa in highly imbalanced settings, suggesting greater statistical stability.

4 Discussion and conclusions

This study examined the behavior of Cohen's Kappa and Gwet's AC1 in a survey measurement error framework, where observed binary responses are imperfect measurements of an underlying true status. By explicitly modeling misclassification through controlled prevalence and observed agreement parameters, the simulation results provide insight into how agreement coefficients behave under the considered assumptions.

The results demonstrate that Cohen's Kappa is highly sensitive to prevalence imbalance. In scenarios where the true prevalence is extreme, Kappa values remain relatively low despite high observed agreement. From a measurement error perspective, this implies that Kappa may substantially understate the quality of survey measurements in applications involving rare events, even when misclassification rates are low.

In contrast, Gwet's AC1 exhibits substantially greater stability across prevalence levels. The simulation results show that AC1 remains closely aligned with observed agreement and displays lower empirical variance, particularly under severe prevalence imbalance. This stability arises from the alternative formulation of expected agreement in AC1, which reduces sensitivity to marginal distributions. As a result, AC1 provides a more consistent and interpretable summary of measurement quality when survey responses are subject to classification error.

References

Alimohammadi, R. and Navvabpour, H. (2008). Response error modeling in face to face surveys. *Research Journal of Science, Isfahan University*, **33**(4):1-14, https://sid.ir/

paper/56025/en.

Alimohammadi, R. (2011). Comparison of variance components estimation methods of response error model in surveys. *Applied Mathematical Sciences*, **5**(48):2405-2410.

Biemer, P.P. and Stokes, S.L. (1991). Approaches of modeling of measurement error. In: Lyberg, L.E., Kasprzyk, D., Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A. and Sudman, S. (Eds) *Measurement Error in Surveys*, New York: Wiley.

Biemer, P.P. and Trewin, D. (1997). A review of measurement error effects on the analysis of survey data. *Survey measurement and Process Quality*, 601-632.

Burton, A., Altman, D.G., Royston, P. and Holder, R.L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, **25**(24):4279-4292.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1):37-46.

Fliess, J.L., Cohen, J. and Everitt, B.S. (1969). Large sample standard errors of Kappa and weighted Kappa. *Psychological Bulletin*, **72**(5):323-327.

Fliess, J.L. and Cohen, J. (1973). The equivalence of weighted Kappa and the itraclass correlation coefficient as measure of reliability. *Educational and Psychological measurement*, **33**(3):613-619.

Fliess, J.L., Levin, B. and Paik, M.C. (2003). *Statistical Methods for Rates and Proportions.* 3rd ed., New York: Wiley.

Gwet, K.L. (2001). *Statistical Tables for Inter-Rater Reliability Studies.* Gaithersburg, MD: STATAXIS.

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.

He, H. and Garcia, E.A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, **21**(9):1269-1284.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American statistical association*, **57**(297):92-115.

Kish, L. (1965). *Survey Sampling.* New York: Wiley.

Morris, T.P., White, I.R. and Crowther, M.J. (2019). Using simulation studies to evaluate statistical methods, *Statistics in Medicine*, **38**(11):2074-2102.