CHIS Working Paper Series

Intercoder Reliability Assessment for Questions with "Other-Specify" Responses in CHIS 2024

Xinyu Zhang, UCLA Center for Health Policy Research
Jiangzhou Fu, UCLA Center for Health Policy Research
Andrew Juhnke, UCLA Center for Health Policy Research
Todd Hughes, UCLA Center for Health Policy Research
Ninez A. Ponce, UCLA Center for Health Policy Research

June 2025

UCLA Center for Health Policy Research

Background

In the current practice of the California Health Interview Survey (CHIS), "other specify" responses that exactly match wording from previous survey waves are automatically assigned the corresponding existing codes. However, all other unique responses are manually coded by human coders. CHIS 2024 has 75 questions with 16,182 "other specify" responses that have to be coded by human coders in the adult, adolescent, and child questionnaires. The subjectivity among human coders raises concerns about the reliability of the qualitative data. Intercoder reliability refers to the degree of agreement between different coders when they independently categorize the same set of data using a standardized coding scheme. The assessment of intercoder reliability yields several benefits, including enhancing the consistency, transparency, and overall systematic nature of the coding process. We aim to assess and improve the reliability of coding for questions with "other specify" responses.

We primarily focus on the questions with more than 100 open-ended responses. Figure 1 shows a flowchart describing the selection process of questions. These questions with a larger number of open-ended responses tend to have a greater impact on survey estimates. One question (AJ115, number of days missed work due to illness, injury, or disability) was excluded from the analysis, because the responses are objective, and the coding is expected to be consistent across different coders. The CHIS Data Access Center (DAC) team also selected three questions with fewer than 100 open-ended responses (CA10A, child physical, behavioral, or mental conditions; CB23, main reason did not visit dentist past year; AJ50, language the doctor spoke to the respondent); CA10A has 24 categories with 78 open-ended responses, CB23 has 12 categories with 69 open-ended responses, and AJ50 has 25 categories with 60 open-ended responses. In total, the CHIS DAC team examined 46 questions with "other specify" responses.

Data & Methods

Our coding team consisted of three coders. Each selected question was coded by two coders independently. The coding frames, which were predefined lists of codes, were based on previous years of CHIS data collection. All responses for each question were ordered alphabetically, which might help reduce the burden of coding. This was because responses that began with the same first few letters were likely to be coded into the same category. For example, for AJ200 (Change in job status due to care recipient), there were multiple unique responses related to similar ideas, such as "Retirement" and "Retiring earlier than planned."

For questions with fewer than 200 open-ended responses (18 questions), we double-coded 50% using systematic sampling, selecting every other response starting with the first response. For questions with more than 200 open-ended responses (25 questions), we double-coded 25%, selecting every fourth response starting with the first response. For questions with fewer than 100 open-ended responses (three questions), we double-coded all responses. The detailed counts of double-coded responses are provided in Appendix Table S1.

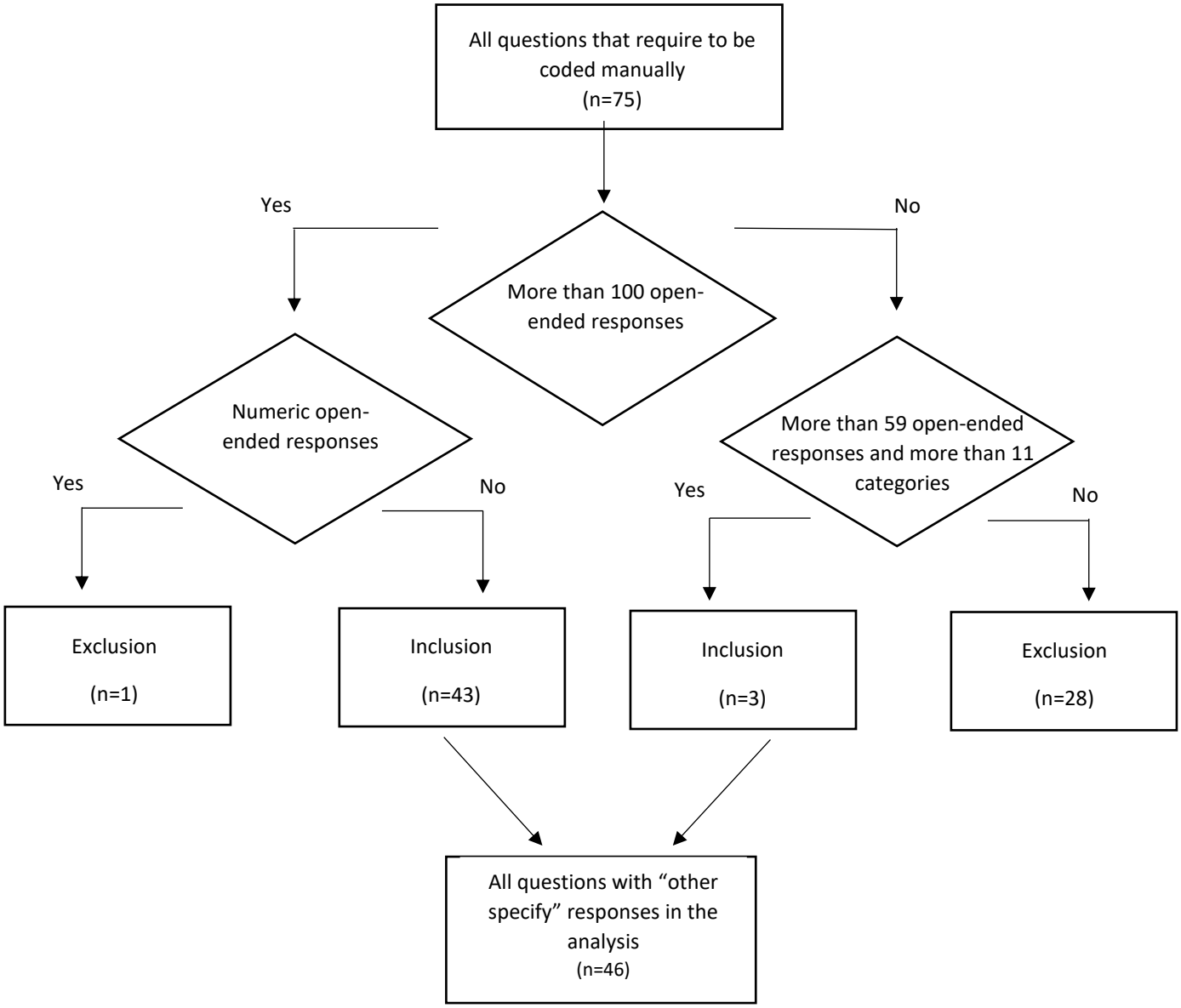

Figure 1. Flowchart of Selecting Questions with "Other Specify" Responses.

For each question, at least 50 responses were double-coded, following the general heuristic that a minimum sample size should be at least 30 (McHugh, 2012). Additionally, Bujang and Baharum (2017) outlined the minimum sample size required for Cohen's Kappa calculation in different scenarios.

In a series of questions where respondents were first asked to select all that apply and then to provide a main reason, only the first open-ended response was coded for both questions to

maintain consistency. Therefore, analysis in this report was limited to the first open-ended response for these types of questions.

In assessing intercoder reliability, we used two key measures: percent agreement and Cohen's Kappa. This allowed for a more comprehensive assessment of intercoder reliability. If coders were likely to make random or uncertain judgements rather than confidently assign codes based on clear criteria, Cohen's Kappa provided a more accurate measure of agreement. However, when coding categories were well-defined and coders were properly trained, percent agreement would still serve as a reliable metric.

The percent agreement measures the proportion of times all raters agree on a set of data, calculated by

$$\text{Percent agreement} = \frac{\text{Number of Agreements}}{\text{Total Number of Observations}} \times 100\%.$$

The percent agreement ranges from 0 to 1, where 0 indicates no agreement between coders, and 1 represents perfect agreement. This measure has the advantage of straightforward interpretation allowing researchers to easily identify variables that may be problematic. However, it does not correct for the possibility that coders guessed on scores. Thus, the percent agreement may overestimate the true agreement among coders.

To address the shortcomings of percent agreement, Cohen's Kappa introduced an adjustment for chance agreement. This measure, similar to correlation coefficients, ranges from -1 to +1, where 0 represents agreement expected by random chance, and 1 indicates perfect agreement between raters. Although negative values are possible, they are rare and would indicate agreement worse than random guessing, meaning systematic disagreement between coders.

Let $p_o = \frac{\text{Number of Agreements}}{\text{Total Number of Observations}}$ be the relative observed agreement, and $p_e = \sum_{i=1}^{k}(p_{i1} \times p_{i2})$ be the hypothetical probability of chance agreement, where k is the number of total categories, p_{i1} is the proportion of responses assigned to category i by Code 1, and p_{i2} is the proportion of responses assigned to category i by Code 2.

Then, Cohen's Kappa κ is calculated as

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

Compute the Standard Error (SE) of Kappa using the formula

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

where *N* is the total number of observations.

The 95% Confidence Interval (CI) of κ is given by

$$CI = \kappa \pm 1.96 \times \mathrm{SE}_{\kappa}.$$

As a general guideline, Cohen suggested the following interpretation of kappa values:

Table 1. Cohen's Kappa Interpretation.

< 0.00	Poor agreement (worse than chance)
0.00 – 0.20	None to slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

However, Cohen's Kappa also has several limitations. One limitation is its sensitivity to the prevalence of agreement in the data. When some categories being coded are significantly more frequent than others, Cohen's Kappa tend to be biased and may not accurately reflect the agreement between two coders.

Results

Overall, the coding by the independent coders is highly consistent across the 46 questions in CHIS 2024. Each question corresponds to a variable in the analysis. Figures 2 - 4 show that the percent agreement for these variables. We found that the percent agreement for all variables exceeded 0.8, with values ranging from 0.81 to 1.0 for each variable, across all three groups—adult, child, and multiple age groups—indicating strong agreement between the two coders.

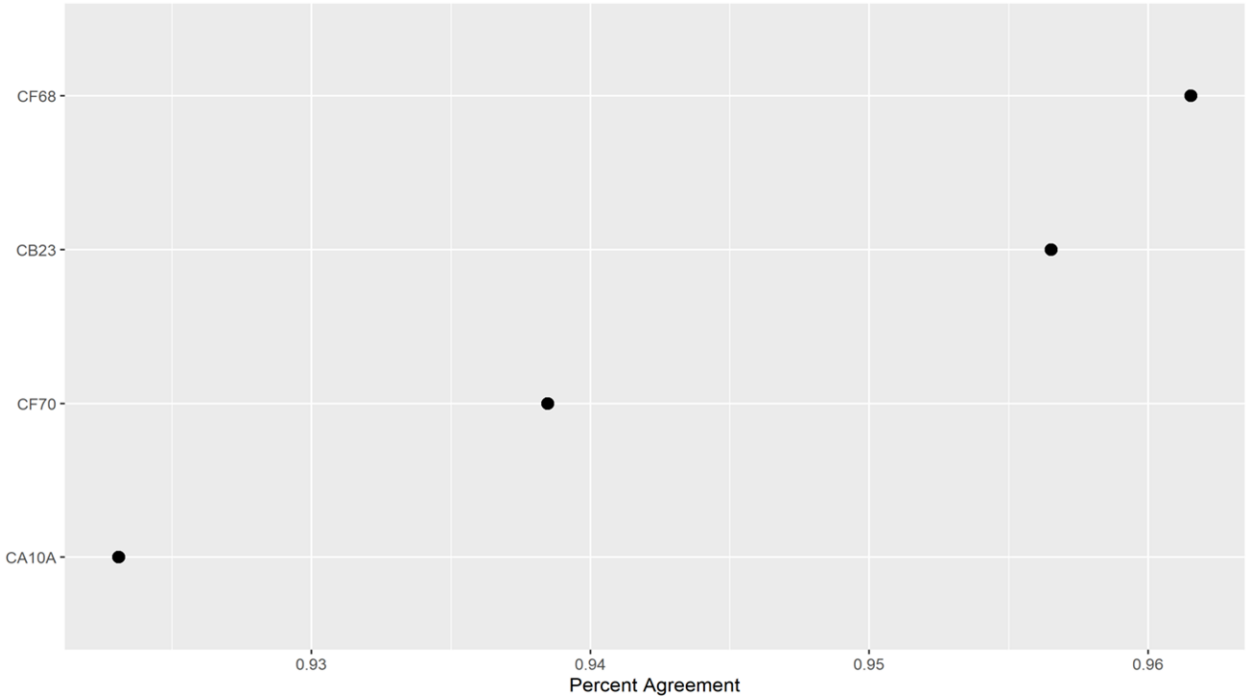

Figure 2. Percent Agreement: Child Variables.

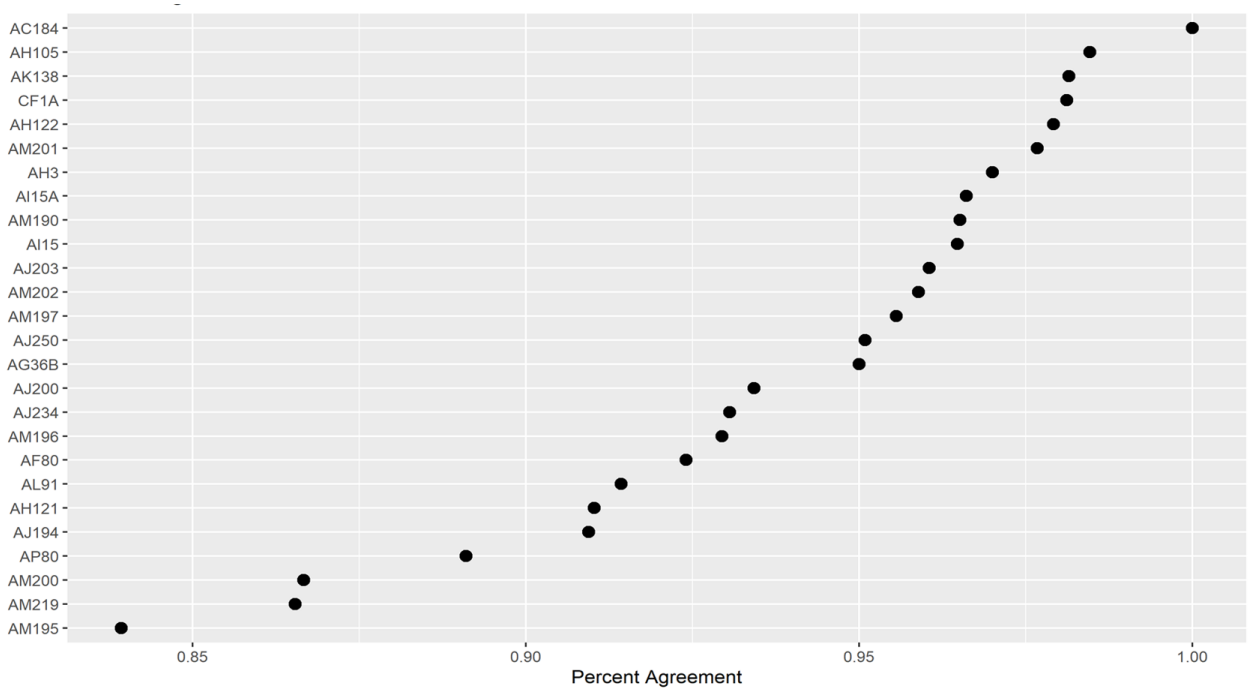

Figure 3. Percent Agreement: Adult Variables.

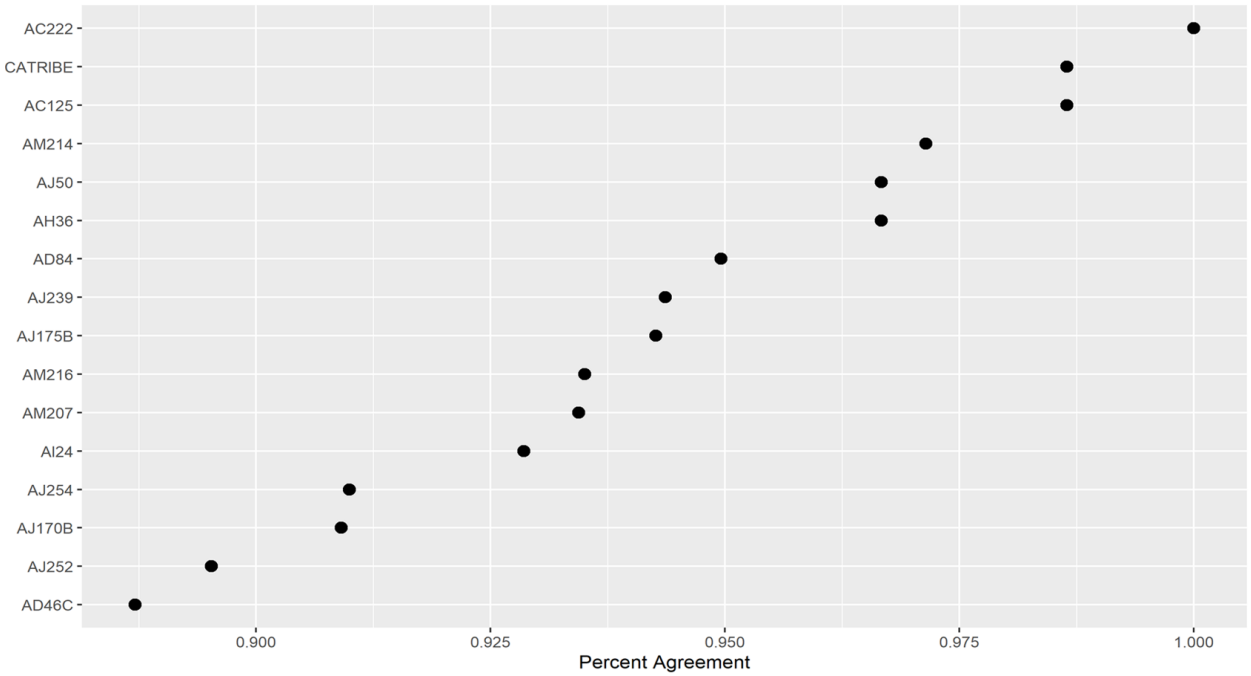

Figure 4. Percent Agreement: Variables Across Multiple Age Groups.

Figures 5 to 7 show that 44 out of 46 variables had Cohen's Kappa of 0.81 or more, reflecting almost perfect agreement. Two variables (AC184, adult used e-cigarette flavor; AC222, reason for prescription painkiller) reached the 100 percent agreement with Cohen's Kappa of 1. However, we observed six variables where their lower bounds of the 95% confidence intervals falling below 0.8. Despite this, the results still indicated strong agreement.

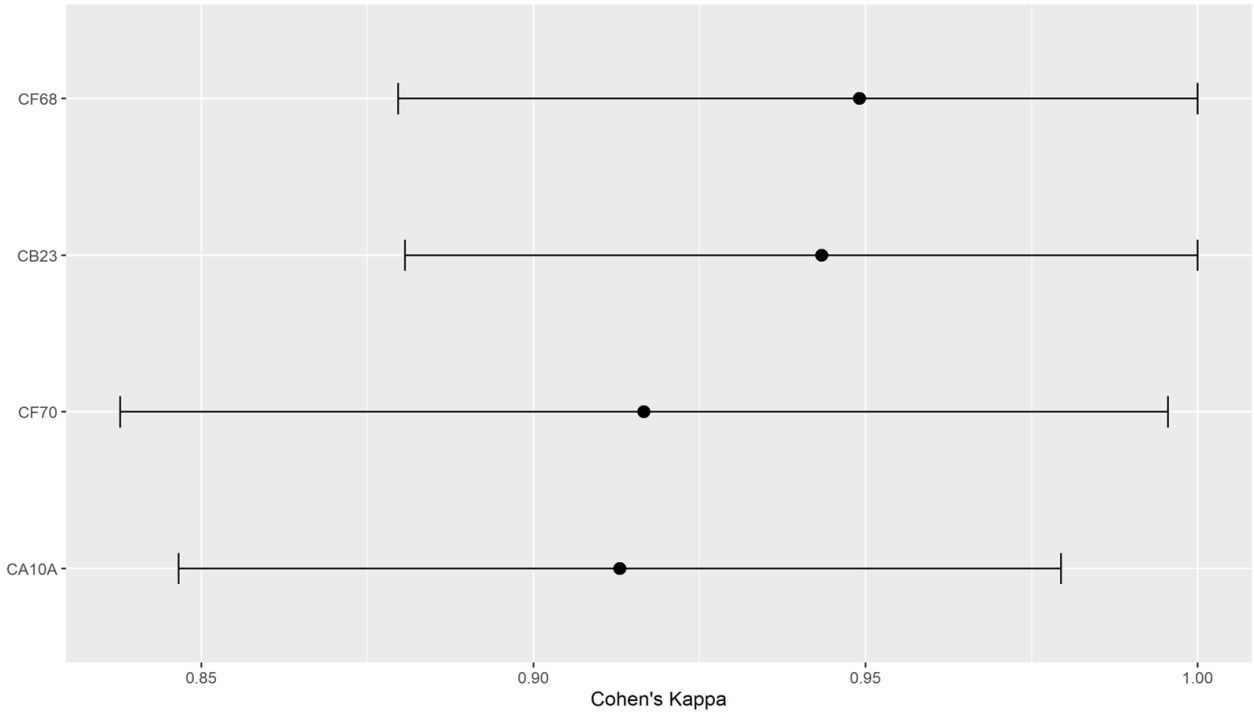

Figure 5. Cohen's Kappa Estimates with 95% Confidence Intervals: Child Variables

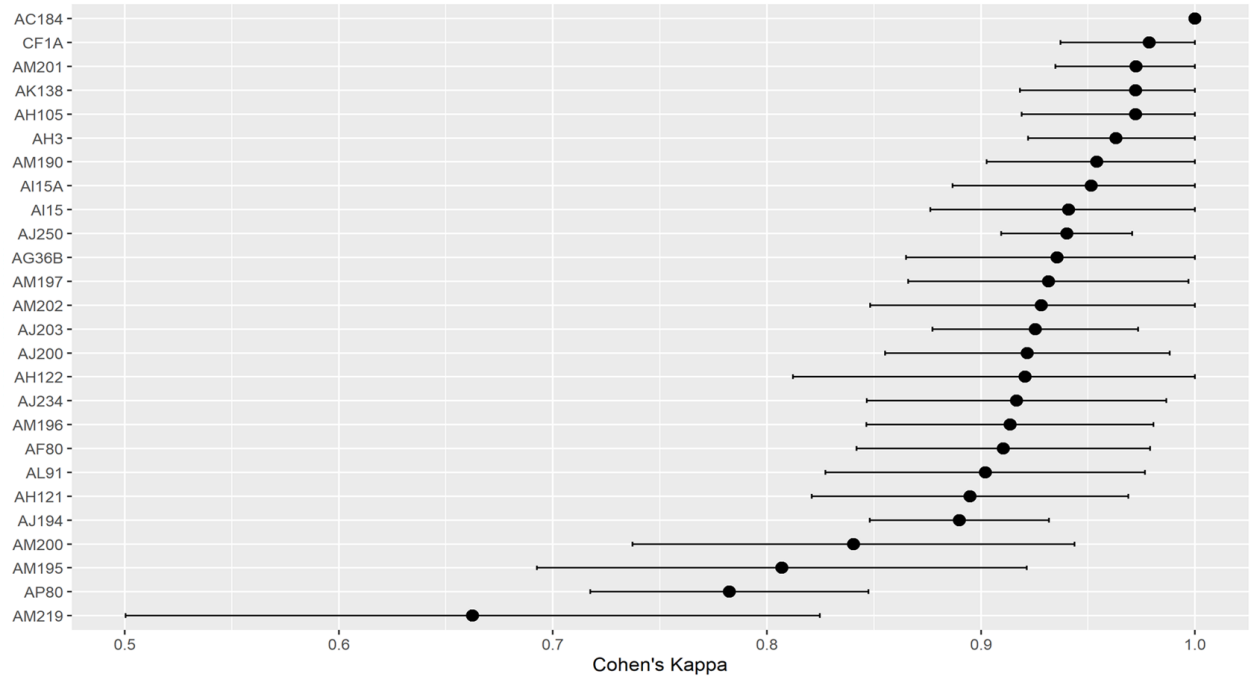

Figure 6. Cohen's Kappa Estimates with 95% Confidence Intervals: Adult Variables.

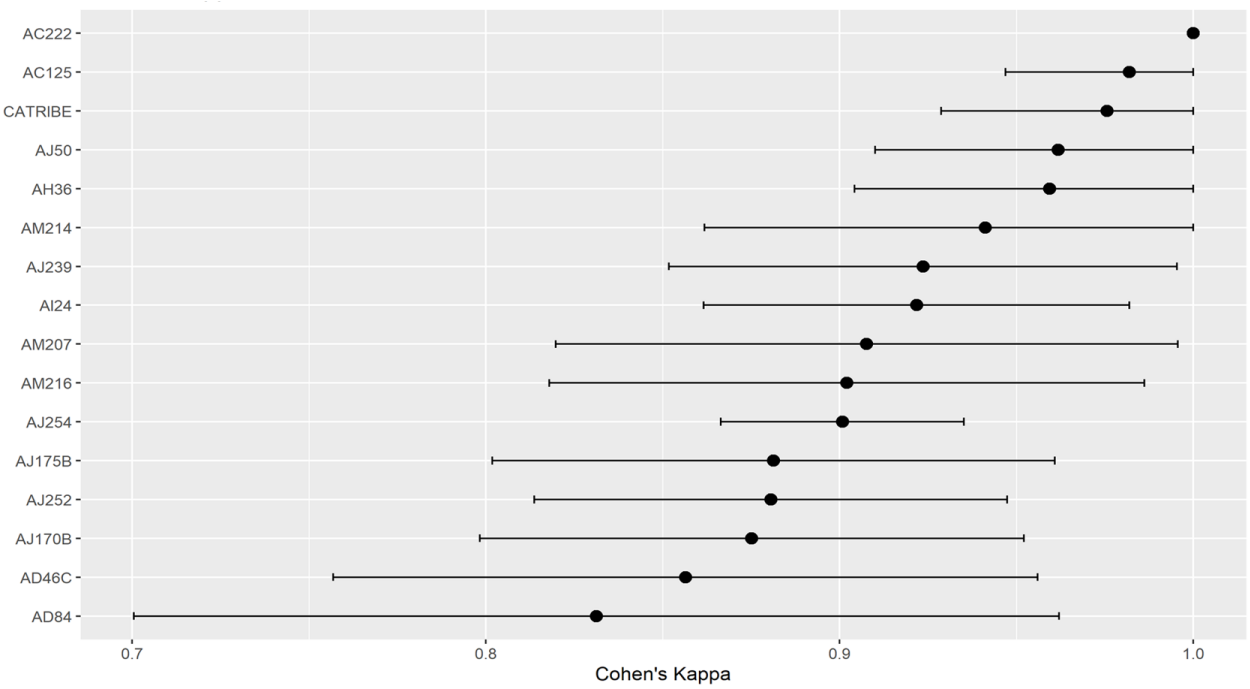

Figure 7. Cohen's Kappa Estimates with 95% Confidence Intervals: Variables Across Multiple Age Groups.

Table 2 presents Cohen's Kappa for the six variables from the CHIS 2024 dataset that exhibited lower inter-coder reliability, with the lower bounds of their 95% confidence intervals falling below 0.8. These variables apply to either adults or multiple age groups.

Table 2. Variables with relatively low Cohen's Kappa.

Age group	Var ID	Content	Cohen's Kappa	95% CI
Adult	AM200	Type of hate incident you witnessed	0.84	[0.73, 0.94]
Adult	AM195	Type of hate incident you experienced	0.81	[0.69, 0.92]
Adult	AM219	Offender of the hate incident	0.66	[0.50, 0.82]
Adult	AP80	Main reason why you did not vote	0.78	[0.71, 0.84]
Multiple	AD46C	Sexual Orientation	0.85	[0.76, 0.96]
Multiple	AD84	HIV testing reason	0.83	[0.70, 0.96]

The variable "Offender of the hate incident" (AM219) had the lowest reliability, with a Cohen's Kappa of 0.66 (95% CI: 0.50 - 0.82), indicating only moderate agreement. The variable "Main reason why you did not vote" (AP80) also showed a relatively low agreement, with a Cohen's Kappa of 0.78 (95% CI: 0.71 - 0.84). Other variables, including "Type of hate incident you experienced" (AM195) and "Type of hate incident you witnessed" (AM200), had Kappa values above 0.8 but their 95% CIs covered 0.8.

For variables in the multiple-age group category, "Sexual orientation" (AD46C) and "HIV testing reason" (AD84) both had relatively strong agreement, with Kappa values of 0.85 and 0.83, respectively. However, their lower bounds of the confidence intervals slightly fell below 0.8.

There are two reasons that account for the disagreements between two coders. First, the linguistic similarity between the codes can lead to complexities in accurate assignment. For example, in "Main reason why you did not vote" (AP80), there are two codes that are close in meanings, one is "Voting has little to do with the way real decisions are made" and the other is "My one vote is not going to affect how things turn out". We observed that two coders used these two codes interchangeably for several responses.

Second, vague and abstract write-in responses can lead to difference interpretations by different coders. Therefore, these responses can be coded differently. For example, in "Offender of the hate incident" (AM219), one coder used "I don't know or I didn't see" while the other coder used "Other" throughout the up-coding for several similar write-in responses.

Overall, while these six variables exhibited slightly lower agreement compared to others, most still fell within the range of substantial agreement.

References

Bujang, M. A., & Baharum, N. (2017). Guidelines of the Minimum Sample Size Requirements for Kappa Agreement Test. *Epidemiology, Biostatistics, and Public Health, 14*(2).

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

Appendix Table S1

VARNAME	AGE	LABEL	CATEGORIES	# Double-coded
AJ254	ADULT	REASON DELAYED/DIDN'T GET NEEDED CARE	21	326
AP80	ADULT	MAIN REASON DID NOT VOTE	14	312
AJ250	ADULT	MAIN REASON NOT VISITED DENTIST IN PAST 12 MOS	7	285
AJ194	ADULT	DISABILITY OR ILLNESS OF CARE RECIPIENT	18	267
AJ203	ADULT	HEALTH PROBLEM FOR TELE-MEDICAL CARE	6	232
AJ175B	ADULT	MAIN REASON YOU ARE NOT CURRENTLY USING BIRTH CONTROL	15	122
AD84	ADULT	OFFERED OR ASKED FOR HIV TEST	5	119
AM219	ADULT	WHO WAS THE OFFENDER OF THE MOST SEVERE HATE INCIDENT	11	114
AJ252	ADULT	REASON DELAYED/DIDN'T GET PRESCRIBED MEDICINE	13	108
AH3	ADULT	KIND OF PLACE FOR USUAL SOURCE OF HEALTH CARE	8	100
AJ170B	ADULT	MAIN REASON YOU ARE NOT CURRENTLY USING BIRTH CONTROL	15	99
AH122	ADULT	IS HEALTH PLAN A PPO OR EPO	5	96
AM197	ADULT	REASON TARGETED FOR HATE INCIDENT	11	94
AM201	ADULT	LOCATION OF WITNESSED HATE INCIDENT	11	91
AM196	ADULT	LOCATION OF HATE INCIDENT	11	89
AM190	ADULT	REASON TARGETED FOR HOUSING-RELATED DISCRIMINATION/HARASSMENT	13	88
AI15	ADULT	MAIN REASON NOT IN EMPLOYER'S HEALTH PLAN	8	85
AI24	ADULT	MAIN REASON NO HEALTH INS AT ALL	17	84
AL91	ADULT	AREAS COUNTY OFFICE CAN IMPROVE	13	83
AF80	ADULT	MAIN REASON QUIT MENTAL HEALTH TREATMENT	12	79
AH121H	ADULT	MOST IMPORTANT REASON CHOSE PLAN	9	78
CA10A	CHILD	CHILD'S CONDITIONS	24	78
AJ200	ADULT	CHANGE IN JOB STATUS DUE TO CARE RECIPIENT	10	64
AM216	ADULT	WHAT HELP OR SUPPORT DID YOU FEEL YOU NEEDED BUT DID NOT RECEIVE	11	77
AM202	ADULT	REASON PERSON TARGETED FOR HATE INCIDENT WITNESSED	11	76
AJ234	ADULT	PHONE/VIDEO APPT PROBLEM EXPERIENCED	10	74
AC125	ADULT	METHOD OF MARIJUANA USE IN PAST 30 DAYS	8	74
CATRIBE	ADULT	CALIFORNIA TRIBE	3	74
AM214	ADULT	AFTER YOU EXPERIENCED THE MOST SEVERE HATE INCIDENT WITHIN THE PAST 12 MONTHS	11	74

AD46C	ADULT	SEXUAL ORIENTATION	6	72
CF70	CHILD	WEHRE GET BOOKS FOR CHILD	9	71
AJ239	ADULT	WHERE REC'D MAIN BIRTH CONTROL METHOD IN PAST 12 MOS	13	71
CB23	CHILD	MAIN REASON DID NOT VISIT DENTIST PAST YEAR	12	69
AH105	ADULT	GET INS THRU EMPLOYER, UNION, SHOP PROGRAM	4	65
AM200	ADULT	TYPE OF HATE INCIDENT WITNESSED	10	64
AM207	ADULT	DURING THE PAST 12 MONTHS HAVE ANY OF THE FOLLOWING HAPPENED TO YOU BECAUSE	7	63
AH36	ADULT	LANGUAGES AT HOME	23	61
AG36B	ADULT	CURRENTLY HERE ON WHAT TYPE OF IMMIGRATION DOCUMENT	8	60
AJ50	ADULT	LANGUAGE THE DOCTOR SPOKE TO THE RESPONDENT	25	60
AI15A	ADULT	MAIN REASON INELIGIBLE FOR EMPLOYER'S HEALTH PLAN	8	59
AC184	ADULT	USED E-CIG FLAVOR	8	57
AM195	ADULT	TYPE OF HATE INCIDENT EXPERIENCED	10	57
CF68	CHILD	CHALLENGES PREVENTING READING TO YOUNG CHILD	7	55
AK138	ADULT	THE REASON LEAVE FROM WORK	4	54
CF1A	ADULT	MAIN REASON CHILD NOT ENROLLED IN THE MEDI-CAL PROGRAM	16	53
AC222	ADULT	REASON FOR PRESCRIPTION PAINKILLER	9	53