Using Pooled Kappa to Summarize Interrater Agreement across Many Items

HAN DE VRIES
RAND Europe
MARC N. ELLIOTT
DAVID E. KANOUSE
STEPHANIE S. TELEKI
RAND Health

We propose the pooled estimator of kappa, an efficient estimator when summarizing the interrater agreement for qualitative data with many items but few subjects. We evaluate this estimator through a simulation of proposed and alternative (average kappa) estimators and subsequently apply our method to calculate pooled and average kappas over 2,176 rated items from six semistructured interviews with sponsors of the CAHPS. The proposed pooled kappa estimator efficiently summarizes interrater agreement by domain. It is more widely applicable and makes better use of scarce subjects than simply averaging item-level kappas.

Keywords: *qualitative analysis; interviews; reliability; simulations*

Cohen's kappa statistic (Cohen 1960) is a widely used measure to evaluate interrater agreement compared to the rate of agreement expected from chance alone on the basis of the overall coding rates of each rater. This chance-corrected statistic is an important measure of the reliability of qualitative data, and although it is still sensitive to the base rates of coding, it more fully considers the effect of base rates than simple measures of agreement. It is applicable to both dichotomous items and items with multiple response categories (Bakeman and Gottman 1986; Simon 2006) and has been extended in a variety of ways. For example, it has been extended to include discrepancies caused by differences in the discretization of units (Bakeman and Gottman 1986; Kravitz et al. 2002), where one coder rates

We thank Kate Sommers-Dawes for assistance in the preparation of this manuscript. The research reported here was supported by the Agency for Healthcare Research and Quality (AHRQ 5U18HS09204-10). Marc Elliott is supported in part by the Centers for Disease Control and Prevention (CDC U48/DP000056). The statements expressed in this article are those of the authors and do not necessarily reflect the views of AHRQ or the CDC.

Field Methods, Vol. 20, No. 3, August 2008 272–282
DOI: 10.1177/1525822X08317166

an event that another coder does not note as an event. It has also been extended to eliminate discrepancies caused by differing base rates between raters (Brennan and Prediger 1981) and to accommodate both covariates and alternative chance models (von Eye 2006).

In this article, we consider a situation in which Cohen's (1960) original kappa is applicable, but the choice of estimator for that kappa may be important. In particular, we wish to summarize kappa values for large sets of related items (which have either dichotomous or multiple response options) when relatively few unique cases have been multiply coded. The sparseness of multiple-coded cases may limit the precision of estimates of kappa unless an estimator makes efficient use of multiple items within a given domain. We therefore compare the precision of two ways of estimating Cohen's kappa in this situation.

More specifically, we consider the situation in which we have two observers, a small number of subjects, and a large number of measurements per subject, a situation not uncommon in anthropology and other fields of social science (Bernard 2001). The measurements arise from the coding by two independent coders of transcripts from a small number of semistructured interviews, with a large number of items per interview.[1]

Calculating a separate kappa value for each item would yield a large number of relatively unstable measures of interrater agreement while failing to provide a meaningful assessment of the overall agreement between coders. It would often be most helpful in such a situation to have a few well-estimated summary kappa statistics for each of several domains. We demonstrate that the choice of a pooled versus averaged estimator of kappa can substantially affect the stability and hence the utility of this approach. We then make specific recommendations for analysts in this common situation.

BACKGROUND

Before discussing the pooled estimator of the kappa statistic, we will briefly review the calculation of the kappa statistic for a single item. Consider two coders who independently judge whether a predefined code matches a respondent's answer to an interview question. When this has been done for a number of interviews, we can count the number of times they agree and disagree and set up the usual agreement table (see Table 1).

Here, $P_{11} \ldots P_{22}$ sum to 1 and represent the observed relative frequencies of a particular outcome (i.e., the number of times each agreement combination was observed, divided by the total number of interviews coded).

TABLE 1
Definitions of Quantities for the Calculation of Kappa

	Coder 2	
Coder 1	*Assigned Code*	*Did Not Assign Code*
Assigned code	P_{11}	P_{12}
Did not assign code	P_{21}	P_{22}

We can calculate the observed agreement as

$$P_O = P_{11} + P_{22}$$

and the agreement we would expect to see by chance alone as

$$P_E = (P_{11} + P_{21}) * (P_{11} + P_{12}) + (P_{12} + P_{22}) * (P_{21} + P_{22}).$$

Kappa is then defined as

$$\kappa = \frac{P_O - P_E}{1 - P_E}.$$

Although there are references to a "pooled" kappa estimator in the applied social science literature, rarely is there a specification of how such an estimator is calculated (Adamson, Bakeman, and Deckner 2004). Two articles in the field of ophthalmology focused on the calculation of kappa when two graders rate pairs of eyes (Oden 1991; Schouten 1993). We build on the idea presented in these articles—that a summary kappa estimator can be defined in two different ways, which we will refer to as pooled kappa (κ_{pooled}) and averaged kappa (κ_{ave}).

Typically, coders do not assign just one code to code an entire interview, but rather tens, hundreds, or even thousands of codes, resulting in a large number of kappa statistics. To summarize these, it might be tempting to take a simple arithmetic mean of the separate *J* kappas, which we will call "averaged kappa":

$$\kappa_{\text{ave}} = \frac{1}{J}\sum_{j=1}^{J} \kappa_j.$$

Alternatively, we may average over the separate P_O and P_E instead of kappa and substitute these averages into the kappa estimator:

$$\kappa_{\text{pooled}} = \frac{\bar{P}_O - \bar{P}_E}{1 - \bar{P}_E},$$

with

$$\bar{P}_O = \frac{1}{J}\sum_{j=1}^{J} P_{O_j}$$

and

$$\bar{P}_E = \frac{1}{J}\sum_{j=1}^{J} P_{E_j}$$

Whereas variance in the numerator of individual kappas has an additive effect on total variance, variance in the denominator of individual kappas has a multiplicative effect. Averaging individual kappas may not efficiently reduce the variance-increasing effects of small denominators for individual kappas. Instead, use of a pooled kappa estimator that stabilizes the denominator variance by pooling estimates of numerator and denominator terms before division may be a more efficient way to reduce the effect of variability in small denominators. To determine empirically which approach provides a better estimate of the true pooled kappa, we performed a simulation.

METHOD

Our simulation consisted of the following steps:

1. The design points at which we evaluated our simulation consisted of all 20,089 possible combinations of $\{P_{11}, P_{12}, P_{21}, P_{22}\}$, for which $\kappa_{true} > 0$, with a resolution of 0.01, after eliminating redundant evaluations points that were identical except for exchanging rater 1 and rater 2. We refer to a given design point as a set $S^* = \{P_{11}, P_{12}, P_{21}, P_{22}\}$. This resulted in observed agreement rates of 19%–98% and true kappa values of 0–0.96.
2. For each set S*, we randomly and independently drew a series of 100 sets $S_1 \ldots S_{100}$ of agreement matrices from a multinomial distribution with probabilities S*. This step generates observed relative frequencies from specified probabilities. $S_1 \ldots S_{100}$ represent 100 separate items over which we wish to calculate a pooled kappa statistic. Because they are randomly drawn from the same distribution, the ideal pooled κ statistic calculated over these series would be equal to κ_{true}. We calculated both κ_{ave} and κ_{pooled}, in addition to the squared error for both, defined as the squared difference between the pooled estimator and κ_{true}. Sampling was multinomial (only total sample sizes fixed), rather than product multinomial (row totals fixed). This simulation assumes no additional dependence in agreement in raters beyond that determined by $S^* = \{P_{11}, P_{12}, P_{21}, P_{22}\}$.

3. For each set S*, we repeated step 2 as a Monte Carlo simulation with 1,000 replications and calculated the mean of the 1,000 squared errors for each pooled kappa. The average squared error over these simulation runs for a given method is the mean squared error (MSE) for that method at that design point or the sum of variance and squared bias (if any). We present errors in terms of the square root of MSE, also known as root MSE (RMSE) because this measure is in the same units as kappa and is comparable to the standard error of kappa.

In some cases, κ_{ave} is not defined because at least one of the item-level kappas is not defined (which occurs when there is 100% expected agreement). The only situation in which κ_{pooled} is not defined is when there is 100% expected agreement for all item-level kappas. When either κ_{ave} or κ_{pooled} is not defined, we exclude both κ_{ave} and κ_{pooled} from the analysis.

The simulation was performed using R version 2.4.1 (R Development Core Team 2006). Pseudo-random numbers were drawn using the default Mersenne-Twister algorithm in R (Matsumoto and Nishimura 1998).

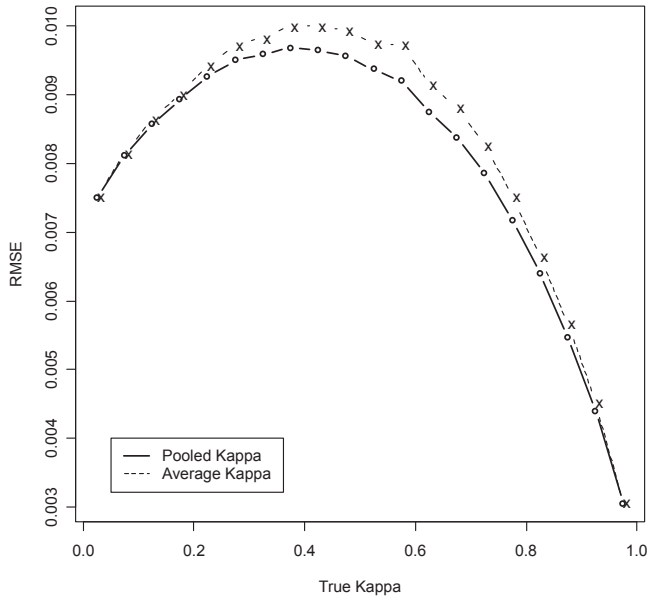
RESULTS

Simulation

From the simulation described above, we obtained 20,089 pairs of MSEs (each based on 1,000 replications at each unique design point) associated with κ_{ave} and κ_{pooled}. Figure 1 shows for each κ_{true} the RMSE of both estimators. To facilitate comparison, we divided the horizontal axis into bins of 0.05 and averaged the RMSEs of each estimator within each bin. RMSEs for both estimators are largest for true kappas of 0.4–0.7 and smallest near true kappas of 0 and 1. Within each bin, the RMSE associated with κ_{pooled} is smaller on average than the MSE associated with κ_{ave}. The advantages of the former over the latter are greatest at middle values of true kappa, where the error and hence the need for precision is greatest.

Figure 2 compares the RMSE of each estimator to the probability of chance agreement (P_E), which ranged from 0.188 to 0.887 with these simulated values. For both methods, RMSE increases with the probability of chance agreement. This is not unexpected, because the denominator approaches 0 when the probability of chance agreement approaches 1, and ratio estimates are known to be more sensitive to variance in the denominator than in the numerator when quantities are close to 0 (Basilevsky 1980). The methods differ little at lower values of P_E, but for $P_E > 0.6$, the RMSE for κ_{pooled} increases more slowly than for $\kappa_{ave.}$ The relative efficiency of κ_{ave}

FIGURE 1
Simulation Results Showing RMSE of Average and Pooled Kappas by True Kappa



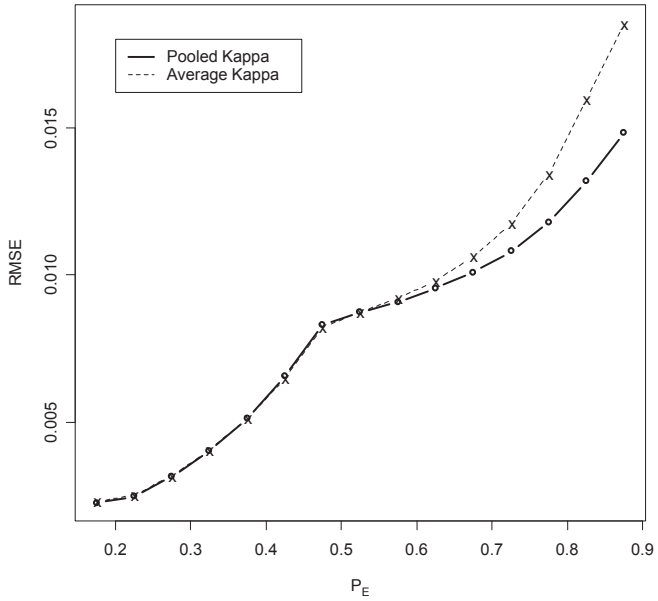
NOTE: RMSE = root mean squared error.

to κ_{pooled} (the inverse ratio of their variances) is less than 91% beyond P_E = 0.75 and less than 72% beyond P_E = 0.85, meaning that κ_{ave} would require as much as 39% more sample (100% / 72% – 100%) to obtain the same precision as κ_{pooled} for $P_E \geq 0.85$. Values of $P_E \geq 0.75$ will typically occur when both raters score a dichotomous event in more than 85% or fewer than 15% of cases.

Application

We applied the pooled kappa estimator in a practical setting in which two interviewers coded twenty-five interview notes from 1-hour interviews with sponsors of the CAHPS Health Plan Survey (Teleki et al. 2007). Six randomly selected interviews were coded by both coders. To assess the reliability of the coding, we calculated κ_{pooled} over subsets of coding items for these six interviews.

FIGURE 2
Simulation Results Showing RMSE of Average and Pooled Kappas by Expected Rates of Agreement



NOTE: RMSE = root mean squared error; P_E = probability of chance agreement.

The interview consisted of 2,176 items in twelve sections,[2] with between 6 and 492 items per section, as detailed in Table 2. For 50% of items, however, only one code was assigned. We did not include these items in the calculation of the pooled kappa statistic (as kappa would be undefined at the item level), so restricted to the 1,076 items (6–274 per section) that showed variation in coding by the entity evaluated. Table 2 displays the pooled kappa statistic and the averaged kappa statistic for each section.

The absolute difference between pooled kappa and average kappa varies between 0 (section K) and 0.08 (sections D and E2). For three sections, average kappa is larger than pooled kappa; for one section, both are equal, and for eight sections, pooled kappa is larger than average kappa. For section G, the choice of estimator would also determine whether the agreement would be classified as "moderate" (with kappa between 0.41 and 0.60) or "substantial" (with kappa between 0.61 and 0.80) according to the Landis

TABLE 2
Application of Pooled Kappa to Semistructured Interviews with Sponsors of the CAHPS Health Plan Survey

Section of Interview	*Topic*	*Total Items in Section*	*Proportion of Items with No Variation*	*Items per Section on which Kappa Was Based*	*Pooled Kappa (standard error)*	*Averaged Kappa (standard error)*
A	General	114	0.47	60	0.82 (0.031)	0.76 (0.044)
B	Reasons for reporting	102	0.53	48	0.69 (0.042)	0.67 (0.044)
C	Audience	189	0.50	95	0.88 (0.021)	0.84 (0.034)
D	Report content	408	0.58	173	0.81 (0.020)	0.73 (0.025)
E1	Report format	263	0.58	110	0.84 (0.021)	0.83 (0.034)
E2	Report format	492	0.44	274	0.71 (0.018)	0.63 (0.018)
F	CAHPS SUN	16	0.00	16	0.70 (0.080)	0.76 (0.086)
G	Dissemination	402	0.51	196	0.63 (0.024)	0.57 (0.022)
H	Actual experience/use	20	0.30	14	0.68 (0.085)	0.64 (0.088)
I	Evaluation	66	0.00	66	0.75 (0.035)	0.76 (0.040)
K	Questions for limited reporters only	6	0.00	6	0.94 (0.055)	0.94 (0.124)
L	Questions for business groups on health only	93	0.81	18	0.92 (0.038)	0.93 (0.083)

SOURCE: Teleki et al. (2007).
NOTE: CAHPS SUN = Consumer Assessment of Healthcare Providers and Systems Survey Users Network.

and Koch (1977) classification. These differences can largely be attributed to greater unsystematic error in average kappa relative to pooled kappa, as the standard error of pooled kappa is smaller than the standard error of average kappa for all but two sections. In one section (E2, report format) these standard errors were the same for the two methods, and for another (G, dissemination), the standard error was slightly larger for the pooled kappa than for averaged kappa. Examination of the latter section revealed some interitem correlation in rates of agreement.

As demonstrated in our simulations, pooled kappa is an equal or more efficient estimator than average kappa when interitem dependence in agreement is 0 or negligible. Small advantages to average kappa are possible when interitem dependence is substantial, which is likely the case only when heterogeneity within a set of items makes a summary measure of agreement questionable or suggests regrouping items to increase homogeneity within the groupings. We believe that, as was the case in this application, dependence is more likely to be negligible, as there is little reason to expect cross-item correlation in the classification of a varied set of interview items. In actual applications, one can calculate and compare the standard errors of pooled and averaged kappa and could consider regrouping items to improve the standard error of pooled kappa in the occasional circumstances where its standard exceeded that of average kappa.

CONCLUSION

Qualitative data, such as those generated by semistructured interviews, often provide extensive, richly detailed observations on a small number of subjects. The former aspect makes summarizing the characteristics of such data especially important, and the latter aspect places a premium on making efficient use of information that is often time consuming and expensive to collect. We propose a means of summarizing the reliability of coding of qualitative data by using a pooled estimator of kappa. We further demonstrate that this estimator is more often defined and more precise for a given sample size than simple averaging of kappa values at the item level. We propose that pooled estimator kappa be routinely used in situations such as those described here.

NOTES

1. Scholars working in the field of content analysis generally regard the textual passage, not the person or persons from whom the passages were derived, as the unit of analysis. Viewed from that perspective, the sample size could be defined as the number of passages to

be coded, not the number of subjects. Because we are trying to make inferences about the subject (the person interviewed), not just the passage, we adhere to a tradition commonly found in disciplines such as psychology and educational research, where the sample consist of "subjects" (persons) and the content collected on the subjects is categorized into "items."

2. Because of the size of section E, we have split this section into two natural subsections and calculated pooled kappa for each.

REFERENCES

Adamson, L. B., R. Bakeman, and D. F. Deckner. 2004. The development of symbol-infused joint engagement. *Child Development* 75 (4): 1171–87.

Bakeman, R., and J. M. Gottman. 1986. *Observing interaction. An introduction to sequential analysis*. London: Cambridge University Press.

Basilevsky, A. 1980. The ratio estimator and maximum-likelihood weighted least squares regression. *Quality and Quantity* 14 (3): 377–95.

Bernard, H. R. 2001. *Research methods in anthropology: Qualitative and quantitative approaches*. Walnut Creek, CA: AltaMira.

Brennan, R. L., and D. J. Prediger. 1981. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement* 41 (3): 687–99.

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

Kravitz, R. L., R. A. Bell, C. E. Franz, M. N. Elliott, C. Willis, and L. Silverio. 2002. Characterizing patient requests and physician responses in office practice. *Health Services Research* 37 (1): 218–38.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33 (1): 159–74.

Matsumoto, M., and T. Nishimura. 1998. Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation* 8 (1): 3–30.

Oden, N. L. 1991. Estimating kappa from binocular data. *Statistics in Medicine* 10 (8): 1303–11.

R Development Core Team. 2006. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org (accessed February 29, 2008).

Schouten, H. J. A. 1993. Estimating kappa from binocular data and comparing marginal probabilities. *Statistics in Medicine* 12 (23): 2207–17.

Simon, P. 2006. Including omission mistakes in Cohen's kappa and an analysis of the coefficient's paradox features. *Educational and Psychological Measurement* 66 (5): 765–77.

Teleki, S. S., D. E. Kanouse, M. N. Elliott, L. Hiatt, H. de Vries, and D. D. Quigley. 2007. Understanding the reporting practices of CAHPS sponsors. *Health Care Financing Review* 28 (3): 1–14.

von Eye, A. 2006. Alternatives to Cohen's k. *European Psychologist* 11 (1): 12–24.

HAN DE VRIES is a senior analyst at RAND Europe, Cambridge, United Kingdom. He does research on health economics, quality of care, econometric modeling, and causal inference. Some recent publications are, with S. Teleki et al., "Understanding the

*Reporting Practices of CAHPS Sponsors" (*Health Care Finance Review*, 2007); with M. Elliott et al., "Equivalence of Mail and Telephone Modes of HCAHPS Data Collection" (*Health Services Research*, 2005); and, with M. M. Samama et al., "An Electronic Tool for Venous Thromboembolism Prevention in Medical and Surgical Patients" (*Haematologica*, 2006).*

*MARC N. ELLIOTT is a senior statistician at RAND Health, Santa Monica, California. He is interested in statistical research on sampling (including weights and nonresponse adjustment), categorical data analysis, case-mix adjustment and survey measurement (including mode effects), propensity score techniques, and experimental design. He also does substantive research on Medicare, racial/ethnic disparities, consumer evaluation of health care, adolescent health, and vulnerable populations. Some recent publications are, with A. Haviland, "Use of a Web-Based Convenience Sample to Supplement and Improve the Accuracy of a Probability Sample" (*Survey Methodology*, 2007); with M. Beckett et al., "Problem-Oriented Reporting of CAHPS® Consumer Evaluations of Healthcare" (*Medical Care Research and Review*, 2007); and, with D. McCaffrey and J. R. Lockwood, "How Important Is Exact Balance in Treatment and Control Sample Sizes to Evaluations?" (*Journal of Substance Abuse Treatment*, 2007).*

*DAVID E. KANOUSE is a senior behavioral scientist, RAND Health, Santa Monica, California. His research interests include report cards on health care quality, health information interventions, sexual behavior, and HIV/STD risk, HIV treatment, and prevention services Some recent publications are, with L. M. Bogart et al., "Patterns and Correlates of Deliberate Abstinence among Men and Women with HIV/AIDS" (*American Journal of Public Health*, 2006); with F. H. Galvan et al., "Religiosity, Denominational Affiliation, and Sexual Behaviors among People with HIV in the United States" (*Journal of Sex Research*, 2007); and, with S. Teleki et al., "Understanding the Reporting Practices of CAHPS Sponsors" (*Health Care Finance Review*, 2007).*

*STEPHANIE S. TELEKI is a policy analyst, RAND Health, Santa Monica, California. She does research on quality improvement, quality of health care, consumer and physician information/behavior (including report cards for these audiences), pay for performance (P4P), and patient safety. Some recent publications include, with D. E. Kanouse et al., "Understanding the Reporting Practices of CAHPS Sponsors" (*Health Care Financing Review*, 2007); with C. L. Damberg et al., "Will Financial Incentives Stimulate Physician Behavior Change to Improve Quality? Reactions from the Frontline" (*American Journal of Medical Quality*, 2006); and, with M. M. Spence et al., "Direct-to-Consumer Advertising of COX-2 Inhibitors: Effect on Appropriateness of Prescribing" (*Medical Care Research and Review*, 2005).*