ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Safety Science 42 (2004) 771–778

SAFETY SCIENCE

www.elsevier.com/locate/ssci

# Technical note: measurement issues in taxonomic reliability

A.J. Ross *, B. Wallace, J.B. Davies

*CASP University of Strathclyde, 40 George Street, Glasgow G1 1 QE, UK*

Received 5 February 2003; accepted 30 October 2003

**Abstract**

Work in safety management often involves *classification* of events using coding schemes or 'taxonomies'. Such schemes contain separate categories, and users have to *reliably* choose which codes apply to the events in question. The usefulness of any system is limited by the reliability with which it can be employed, that is the consensus that can be reached on application of codes. This technical note is concerned with practical and theoretical issues in defining and measuring such reliability. Three problem areas are covered: the use of correlational measures, the reporting and calculating of indices of concordance and the use of correction coefficients.

*Keywords:* Taxonomies; Classification; Reliability; Inter-rater consensus

## 1. Introduction

Work in safety management often involves *classification* of events using coding schemes or 'taxonomies'. Often, such schemes contain separate categories, and users have to choose which codes apply to the events in question. Taxonomies (especially those which are used 'after-the-fact') typically involve classification of features of a

* Corresponding author. Address: University of Strathclyde, 141 St. James Road, Glasgow, UK. Tel.: +44-141-548-4503; fax: +44-141-548-4508.

*E-mail address:* alastair.j.ross@strath.ac.uk (A.J. Ross).

0925-7535/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ssci.2003.10.004

system (for example, human behaviours, organisational/environmental factors, or cognitive factors) [1] which can be examined to help avoid unwanted events in the future.

Classifying events using taxonomies designed for the purpose is a common technique in the human sciences (e.g. psychology, sociology, psychiatry) and has been shown to be highly useful, dependent on certain important criteria. The most critical of these has usually been called 'inter-judge' (Cohen, 1960), 'inter-observer' (Caro et al., 1979) or 'inter-rater' (Posner et al., 1990) *reliability*. Tests of this criterion have usually been simply called *reliability studies* (Grove et al., 1981). (The related concept of 'intra-rater' reliability refers to a comparison between the judgments made by the same judge about the same data on different occasions.)

Importantly, reliability in event classification is the extent to which independent users of a coding scheme, taxonomy or similar diagnostic technique can agree on discrete events to be coded (e.g. Caro et al., 1979). A basic principle of coding reliability is that agreement refers to the ability to discriminate *for individual subjects, events or cases* (e.g. Cohen, 1960; Fleiss, 1971; James et al., 1993). This type of agreement (consensus on individual codes) is a pre-requisite for the validity and pragmatic usefulness of a coding device (e.g. Groeneweg, 1996, p. 134; Stanton and Stevenage, 1998, p. 1746).

## 2. Using patterns or frequencies as a test of consensus

In safety management there is a tendency to call a *consistent* pattern of 'codes' from a coding scheme *reliable*, and to assume that people can *agree* on which codes to pick for individual events. For example, Stanton and Stevenage (1998, p. 1740) define 'consistency' thus: 'this criterion is the same as inter-rater reliability, i.e. the degree to which two analysts make the same error predictions'. But equating consistency with inter-rater reliability is misleading. Coding schemes can actually produce highly *consistent* data in the absence of independent *agreement* on discrete events. Indeed, in the most extreme case, *consistent* or reliable coding can be demonstrated in the absence of any *agreement* at all. (This distinction is discussed in detail by James et al., 1993, p. 306.)

For example, if two coders assign codes to a set of events, and use each individual code the same number of times overall, coding is highly consistent, i.e. there will be a rank order correlation of 1 in *overall code utilization*. If this happens over repeated trials the coding output is also 'reliable'. However, this is not evidence for *inter-rater reliability* (as defined above), as the coders may have disagreed on which codes to assign to individual events. This is why we prefer *inter-rater consensus* (IRC) rather

[1] These have been described respectively as the 'external mode of malfunction' (Rasmussen et al., 1981) or 'external error mode' (Isaac et al., 2000); the 'general failure type' (Reason, 1990) or 'external performance shaping factor' (Rasmussen et al., 1981); and the 'psychological failure mechanism' (Hollnagel, 1998) or 'internal performance shaping factor' (Rasmussen et al., 1981).

than *reliability* as a term to denote the capacity of users to independently agree on classification (Davies et al., 2003).

Wallace et al. (2002, p. 2) provide evidence to show that this is the case (see also Davies et al., 2003). These authors found a highly consistent pattern of codes from a root-cause coding system in the nuclear industry despite low reliability (consensus) between experienced users of the technique (index of concordance = 42%; for definitions as to what constitutes acceptable consensus see Borg and Gall, 1989). In addition, an analysis of the distribution of codes assigned *during the reliability trial* showed there was a strong correlation between these and the usual pattern ($\rho = 0.837$). These data are of general interest because they negate the assumption that correlations between patterns are evidence for *consensus* in code or category assignment. Lack of consensus means predictive and discriminatory utility of the database is lost. For examples from the literature where correlations and overall frequencies have been used to test 'reliability', leading to inter-rater *consensus* being overlooked, see Kirwin (1988, p. 99), Stanton and Stevenage (1998, p. 1737), and Groeneweg (1996, p. 229).

The appropriate method for calculating inter-rater consensus is to calculate the 'index of concordance' (e.g. Martin and Bateson, 1993, p. 120). This is done by applying the formula $A/A + D$, where $A =$ the total number of agreements and $D =$ the total number of disagreements. Inter-rater consensus can then be reported as a figure between 0 and 1 or as a percentage by multiplying by 100, and is often 'corrected' using coefficients such as Cohen's $\kappa$ (Cohen, 1960) (see discussion below). However, even where these appropriate measures are used, problems may still arise in terms of calculation and reporting.

## 3. Calculating and reporting on indices of consensus

It is usually desirable that an overall 'inter-rater consensus' score can be calculated when there are more than two coders involved in a trial. However, a measure of consensus should be computed for each pair of coders separately, from which an average may be computed. Fleiss (1971) outlines clearly how agreement between multiple raters is to be calculated. Raw agreement arises from 'the proportion of agreeing pairs out of all the ... possible pairs of assignments' (p. 379).

Problems can arise if researchers try to calculate average agreement for more than two coders at a time. For example, Stratton et al. (1988) provide reliability data for a coding system [2] by stating that '... a total of 315 ... statements was extracted and of these, 220 were identified by independent agreement of *at least two of the three raters*' (p. 89) (*emphasis added*). But if '2 out of 3' people assign a code the agreement for that case is 33% (two pairs disagree and one agree; see Davies et al., 2003). So in this case, even if, for example, half the 220 statements in the 'two or more' group were

[2] This system has been used in an organisational context by Silvester et al. (1999) and Munton et al. (1999).

agreed by *all three* raters, then the total agreement is 56% (the average of 110 trials at 100% and 205 at 33%), which still leaves doubt as to the usefulness of the system. The safest method with multiple raters is to score the mean of all possible paired comparisons as outlined by Fleiss (1971).

## 4. Pre-selecting events

Wallace et al. (2002) report a trial of SECAS (Strathclyde Event Coding and Analysis System) which was developed for use in the nuclear industry. There were two distinct parts to the trial. In the first part, events to be coded were identified and coded *independently* by raters so that agreement could be tested on both selection and coding of events. (This process involved coding events directly from reports in the 'natural language' of reporters, not rewritten or summarized reports which can sometimes be used to make coding easier.)

In the second trial, events for coding were first selected from a sample of reports by one coder, who then coded those events. These events were then passed on to a second coder who coded them independently. In this way, agreement (or disagreement) due to the coding scheme could be tested independently from agreement (or disagreement) as to what constituted a 'codeable' event. Agreement was around 20% higher when the second coder did not have to read whole reports and *decide what events to code*. As it is often the case that events to be classified in a reliability trial are pre-identified (i.e. highlighted in reports before the coding process), we would recommend factoring in a drop in agreement of at least 10% when users have to identify events themselves, although it can be conceded that, in practice, events to be coded are often decided upon prior to any independent use of a system. [3]

## 5. Ambiguity in reporting

Baber and Stanton (1996, p. 126) report high reliability for use of SHERPA (Systematic Human Error Reduction and Prediction Approach, Embrey, 1986). However, the data presented in the paper are somewhat ambiguous. The system involves identifying errors and classifying them using a coding system for error types. First, it is stated that 'Analyst two found 47 errors, 44 of which were found by analyst one'. This appears at first glance to be evidence for consensual coding (i.e. agreement on individual cases). However, it is of course possible that *an uneven number of coding attempts were made*. Suppose analyst one found 276 errors, of which 44 matched the 44 (out of 47) coded by analyst two, reliability would then be

[3] In addition, trials were carried out on consecutive days, so that 'practice' could be evaluated. Agreement did indeed increase from day 1 (56% for selection and coding; 72% for coding only) to day 2 (66% and 89% respectively). These data show how consensus needs to be continuously evaluated and that 'snapshot' trials can be misleading.

around 16%. This example shows why clear presentation of results (including the number of codes assigned by each coder) is essential.

It must be stressed that any implied criticism of those reporting on reliability (consensus) trials is intended as entirely constructive, as it must be noted that such data is apparently not available *at all* for many common techniques. (Wagenaar and van der Schrier, 1997, evaluated a number of techniques (including MORT (Johnson, 1980); STEP (Hendrick and Benner, 1987); and FTA (Ferry, 1988)) and describe these techniques as presenting no 'inter-rater reliability' data. Further they state that 'there is no real excuse for the lack of reliability testing ... since it is not difficult to measure between-raters reliability ...' (p. 31).)

## 6. Issues with statistical measures of agreement

Finally, we have come to be concerned with the common use of the $\kappa$ coefficient (Cohen, 1960, 1968; Fleiss, 1971). This is typically used in the human sciences to correct the Index of Concordance for agreement 'expected by chance'. [4] A discussion seems wise as to the applicability of this technique for calculating the reliability of safety management schemes.

In keeping with the distinction outlined above between correlation and consensus on individual cases, $\kappa$ is used precisely because it can be interpreted 'as a measure of the amount of *agreement* (*as opposed to correlation or association*) between two raters ...' (Spitznagel and Helzer, 1985 *emphasis added*). However, $\kappa$ has been extensively critiqued, and its use in this context is questioned here.

$\kappa$ is a simple formula for correcting the number of categorical agreements between independent judges for the number of agreements that would be expected purely by chance. The probability of chance agreement on a single code is calculated from the probability of each rater using a code relative to the total number of codes assigned (Cohen, 1960, p. 38). $\kappa$ is then computed as $[(A/A+D) - \text{chance agreement}]/(1 - \text{chance agreement})$ *where* $A =$ *number of agreements and* $D =$ *number of disagreements*. It is useful to note that Cohen (1960, p. 38) is quite clear as to the conditions under which $\kappa$ can be used:

(a) the events/scenarios/errors to be coded are independent;
(b) the categories used are independent, mutually exclusive and exhaustive;
(c) the coders operate independently.

Perhaps the most fundamental criticism of $\kappa$ is that the concept of correcting for 'chance agreement' is inherently flawed (Grove et al., 1981; Carey and Gottesman, 1978; Janes, 1979). This is related to violation of the third condition above, independence of raters. As Spitznagel and Helzer (1985), put it 'the assumption about the

[4] Examples of emergent use of $\kappa$ in safety management work include Isaac et al. (2000) and Lehane and Stubbs (2001).

independence of errors (i.e. coders' decisions) is probably never correct' (p. 727). Maxwell (1977) argues that it is absurd to argue that coders start from a position of complete ignorance, and thus rejects the idea that $\kappa$ can measure agreement relative to the proportion of cases arising from chance alone.

Another assumption underlying the use of the $\kappa$ statistic is that codes tested will be mutually exclusive and exhaustive codes. There appears to be a paradox here. One essentially tests how much overlap and redundancy there is in a coding taxonomy by calculating consensus (IRC). If trained coders cannot agree on classification then we would argue that the codes are not functioning as exclusive categories. Yet mutually exclusive categories are a pre-requisite for $\kappa$. There is no easy solution here—developers must simply endeavor to design coding frames where definitions are clear and choices are absolute. As a guideline here, we would recommend computing *raw agreement* (i.e. the Index of Concordance (Martin and Bateson, 1993)) first. $\kappa$ coefficients calculated on the basis of *low raw agreement* can be assumed to violate the principle of mutual exclusivity and should be avoided.

A final problem with $\kappa$ is that it is sensitive to prevalence (also called base rate; Spitznagel and Helzer, 1985). This issue has been extensively discussed in the literature on psychiatric diagnosis and epidemiology. In simple terms, the problem is that $\kappa$ varies not just with consensus between raters but with distribution of cases to be coded (see Davies et al., 2003). Grove et al. (1981) point out that $\kappa$ is actually a series of reliabilities, one for each base rate. As a result, $\kappa$s are seldom comparable across studies, procedures, or populations (Thompson and Walter, 1988; Feinstein and Cicchetti, 1990; Davies et al., 2003).

Alternatives to $\kappa$ have been proposed. A solution to the prevalence problem was proposed by Spitznagel and Helzer (1985), who also provide a detailed discussion of these issues (see also Cicchetti and Feinstein, 1990). Spitznagel and Helzer recommended a statistic called the coefficient of colligation ($Y$) (Yule, 1912). $Y$ remains more stable than $\kappa$ (it is effectively independent of prevalence) for all but high prevalence rates. However, Lee and Del Fabbro (2002) argue that a fundamental problem with both $\kappa$ and $Y$ is that they are 'frequentist' (in simple terms '10 agreements and 10 disagreements' is treated the same as "50 agreements and 50 disagreements" in the calculation). If a Bayesian approach is adopted (Carlin and Louis, 2000; Leonard and Hsu, 1999; Sivia, 1996), the observed data are used to revise prior beliefs. So, as a picture of agreement builds up, changes in *frequencies* rather than *ratios* alters the calculation (i.e. the more *actual* agreements there are the higher the coefficient becomes). [5] Davies et al. (2003) recommend consideration of the BK Coefficient of Agreement for Binary Decisions (Lee and Del Fabbro, 2002) for correcting raw consensus.

[5] A coefficient which is not sensitive to prevalence (RE) was proposed by Maxwell (1977), however RE is still depends on ratios, and is not sensitive to absolute numbers of observations. Interestingly, Maxwell argues that it is false to argue that raters start from a position whereby 'chance' agreement is a possibility, and so rejected $\kappa$ where agreement is measured against chance.

## 7. Conclusion

The concept of ‘inter-rater reliability’ is essential for a workable taxonomy (e.g. Groeneweg, 1996, p. 134). We have proposed the term ‘inter-rater consensus’ (IRC) (Davies et al., 2003) in order to avoid current confusion around the term ‘reliability’. We have shown how consensus between users of taxonomies may:

(a) be overlooked (see reviews by Kirwin, 1992; Wagenaar and van der Schrier, 1997);
(b) be examined inappropriately by looking at correlations between overall frequencies of codes assigned (Stanton and Stevenage, 1998; Groeneweg, 1996). (Wallace et al., 2002, show data which demonstrate why this does not provide evidence for consensus, and outline a method for a consensus trial.)
(c) be reported ambiguously or incompletely.

Recommendations for analysis and reporting have been made, including an encouragement to consider alternative Bayesian measurements (e.g. BK; Lee and Del Fabbro, 2002) because of problems with, for example, the $\kappa$ correction coefficient (Cohen, 1960).

## References

Baber, C., Stanton, N.A., 1996. Human error identification techniques applied to public technology: predictions compared with observed use. Applied Ergonomics 27 (2), 119–131.

Borg, W., Gall, M., 1989. Educational Research. Longman, London.

Carey, G., Gottesman, I.I., 1978. Reliability and validity in binary ratings: areas of common misunderstanding in diagnosis and symptom ratings. Archives of General Psychiatry 35, 1454–1459.

Carlin, B.P., Louis, T.A., 2000. Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall, New York.

Caro, T.M., Roper, R., Young, M., Dank, G.R., 1979. Inter-observer reliability. Behaviour 69 (3–4), 303–315.

Cicchetti, D.V., Feinstein, A.R., 1990. High agreement but low kappa II: resolving the paradoxes. Journal of Clinical Epidemiology 43 (6), 551–558.

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

Cohen, J., 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70, 213–220.

Davies, J.B., Ross, A.J., Wallace, B., Wright, L., 2003. Safety management: A Qualitative Systems Approach. Taylor and Francis, London.

Embrey, D.E., 1986. SHERPA: a systematic human error reduction and prediction approach. In: Proceedings of the International Topical Meeting on Advances in Human Factors in Nuclear Power Systems. American Nuclear Society, LaGrange Park.

Feinstein, A.R., Cicchetti, D.V., 1990. High agreement but low kappa: the problems of two paradoxes. Journal of Clinical Epidemiology 43 (6), 543–549.

Ferry, T.S., 1988. Modern Accident Investigation and Analysis. Wiley & Sons, New York.

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin 76, 378–381.

Groeneweg, J., 1996. Controlling the Controllable: The Management of Safety, third revised ed. DSWO Press, Leiden.

Grove, W.M., Andreasen, N.C., McDonald-Scott, P., Keller, M.B., Shapiro, R.W., 1981. Reliability studies of psychiatric diagnosis. Archives of General Psychiatry 38, 408–413.

Hendrick, K., Benner, L., 1987. Investigating Accidents with STEP. Dekker, New York.

Hollnagel, E., 1998. Cognitive Reliability and Error Analysis Method. Elsevier Science, Oxford.

Isaac, A., Shorrock, S., Kirwin, B., Kennedy, R., Anderson, H., Bove, T., 2000. Learning from the past to protect the future—the HERA approach. In: 24th European Association for Aviation Psychology Conference, Crieff.

James, L.R., Demaree, R.G., Wolf, G., 1993. $r_{wg}$: an assessment of Within-Group Interrater Agreement. Journal of Applied Psychology 78 (2), 306–309.

Janes, C.L., 1979. Agreement measurement and the judgement process. Journal of Nervous and Mental diseases 167, 343–347.

Johnson, W.G., 1980. MORT Safety Assurance Systems. Marcel Dekker, New York.

Kirwin, B.A., 1988. A comparative study of five human reliability assessment techniques. In: Sawyer, B.A. (Ed.), Human Factors and Decision Making: Their Influence on Safety and Reliability. Elsevier Applied Science, London, pp. 87–109.

Kirwin, B., 1992. Human error identification in human reliability assessment. Part 2: Detailed comparison of techniques. Applied Ergonomics 23, 371–381.

Lee, M.D., Del Fabbro, P.H., 2002. A Bayesian coefficient of agreement for binary decisions. Available from <http://www.psychology.adelaide.edu.au/members/staff/michaellee/homepage/bayeskappa.pdf>.

Lehane, P., Stubbs, D., 2001. The perceptions of managers and accident subjects in the service industries towards slip and trip accidents. Applied Ergonomics 32, 119–126.

Leonard, T., Hsu, J.S.J., 1999. Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers. Cambridge University Press, New York.

Martin, P., Bateson, P., 1993. Measuring Behaviour: An Introductory Guide. CUP, Cambridge.

Maxwell, A.E., 1977. Coefficients of agreement between observers and their interpretation. British Journal of Psychiatry 130, 79–83.

Munton, A.G., Silvester, J., Stratton, P., Hanks, H., 1999. Attributions in Action: A Practical Approach to Coding Qualitative Data. Wiley, Chichester.

Posner, K.L., Sampson, P.D., Capln, R.A., Ward, R.J., Cheney, F.W., 1990. Measuring interrater reliability among multiple raters: an example of methods for nominal data. Statistics in Medicine 9, 1103–1115.

Rasmussen, J., Pedersen, O.M., Mancini, G., Carnino, A., Griffon, M., Gagnolet, P., 1981. Classification System for Reporting Events Involving Human Malfunctions. Risø National Laboratory, Roskilde.

Reason, J., 1990. Human Error. CUP, Cambridge.

Silvester, J., Anderson, N.R., Patterson, F., 1999. Organizational culture change: an inter-group attributional analysis. Journal of Occupational and Organisational Psychology 72, 1–23.

Sivia, D.S., 1996. Data Analysis: A Bayesian Tutorial. Clarendon Press, Oxford.

Spitznagel, E.L., Helzer, J.E., 1985. A proposed solution to the base rate problem in the kappa statistic. Archives of General Psychiatry 42, 725–728.

Stanton, N.A., Stevenage, S.V., 1998. Learning to predict human error: issues of acceptability, reliability and validity. Ergonomics 41 (11), 1737–1756.

Stratton, P., Munton, A.G., Hanks, H., Heard, D.H., Davidson, C., 1988. Leeds Attributional Coding System (LACS) Manual. LFTRC, Leeds.

Thompson, W.D., Walter, S.D., 1988. A reappraisal of the $k$ coefficient: $k$ and the concept of independent errors. Journal of Clinical Epidemiology 41, 949–958, 969–970.

Wagenaar, A., van der Schrier, J., 1997. Accident analysis: the goal, and how to get there. Safety Science 26 (1), 25–33.

Wallace, B., Ross, A., Davies, J.B., Wright, L., 2002. The creation of a new minor event coding system. Cognition Technology and Work 4, 1–8.

Yule, G.U., 1912. On the methods of measuring association between two attributes. Journal of the Royal Statistical Society 75, 581–642.