Journal of Modern Applied Statistical Methods

Volume 3 | Issue 1 Article 25

5-1-2004

On The Reporting Of Reliability In Content Analysis

Patric R. Spence
Wayne State University, spencepa@wayne.edu

Recommended Citation
Spence, Patric R. (2004) "On The Reporting Of Reliability In Content Analysis," *Journal of Modern Applied Statistical Methods*: Vol. 3 : Iss. 1 , Article 25.
DOI: 10.22237/jmasm/1083371100

Journal of Modern Applied Statistical Methods
May, 2004, Vol. 3, No. 1, 234-238

1538 – 9472/04/$95.00

On The Reporting Of Reliability In Content Analysis

Patric R. Spence
Department of Communication
Wayne State University

This article explores one type of misreporting of reliability that has been seen in recent conference papers and articles using the method of content analysis. The reporting of reliability is central to the validity of claims made using this method. A brief overview of content analysis is offered, followed by the exploration of one type of misreporting of reliability. Suggestions are offered to address the problem.

Key words: Content analysis, intercoder reliability

Introduction

Though many definitions of content analysis have been offered over the years (Berelson, 1952; Weber, 1990; Berger, 1991), a complete and concise contemporary definition is offered by Neuendorf (2002), who defines it as “summarizing, quantitative analysis of messages that relies on the scientific method (including attention to objectivity-intersubjectivity, a prior design, reliability, validity, generalizability, replicability, and hypothesis testing) and is not limited to the types of variables that may be measured or the context in which the messages are created or presented” (p. 10).

Content analysis is used for numerous purposes in several fields of study. Examples include; settling disputed authorships (Berelson, 1952), during World War II, the technique was employed to gather information from enemy literature (George, 1959), rule making among jury members (Seibold, 1998), interactions in adolescent peer groups (England & Petro, 1998), advertising in children's television (Stern & Harmon, 1984), the role of face in organizational relationships (Redding & Ng, 1982), minority

advertising in children's television (Stern & Harmon, 1984), the role of face in organizational relationships (Redding & Ng, 1982), minority

representation on television (Tamborini et al., 2000) and several accounts of media topics.

More specifically, in education the method has been used to research issues such as, the press as a resource for teaching science (Dimopoulous, Koulaidis & Sklaveniti, 2003), the treatment of gender in teacher education textbooks (Zittleman & Sadker, 2002), and materials in specific textbooks (Harmon, Hedrick & Fox, 2000; Plucker & Beghetto, 2000).

Content analysis is a popular method used in the behavior sciences because of its ability to be utilized for both written and oral communication as well as its ability to compare data across time and context. The method allows the researcher to identify particular words, phrases or concepts within the text(s) being examined. The text(s) that are used can be transcripts of communication, classroom interactions, historical documents, newspaper articles, magazine articles, books, interviews, essays, speeches, and almost any behavioral event that is recorded in some manner.

The importance of intercoder reliability is of central concern when content analysis is used. Intercoder reliability is “the extent to which independent coders evaluate a characteristic of a message or artifact and reach the same conclusion” (Lombard et al., 2002, p. 589). This provides a validation to the coding

Patric R. Spence (M. A., Michigan State University) is a graduate student Department of Communication at Wayne State University. Email: spencepa@wayne.edu.

scheme. Thus, intercoder reliability demonstrates that more than one person can use the coding scheme and obtain similar results. The validity of the data and any subsequent interpretations are suspect if intercoder reliability is not established or reported. Further, not only does the establishment of intercoder reliability help ensure validity, but it also allows the work of coding to be distributed among multiple coders (Neuendorf, 2002).

Much of the concern within the method is whether separate coders achieve agreement on the values assigned to an examined data point. The simplest method of assessing reliability between coders is a percent agreement. This statistic represents the number of between coder agreements divided by the total measures observed. Percent agreement is the most common measure of intercoder reliability; however, while it is intuitively appealing and simple to calculate, it is a misleading measure that overestimates the true score. The statistic has a range from .00 (no agreement) to 1.00 (perfect agreement).

$$PA_o = A / n \qquad (1)$$

PAo concerns the proportion agreement, observed, where *A* is the number of agreements between the two coders and *n* represents the total number of units the coders have coded (Neuendorf, 2002).

Cohen's *kappa* (1968) is the most popular reliability assessment used (Zwick, 1988), particularly because of its accessibility in SPSS. The kappa accounts for the role of chance in agreements in coding which the percent agreement does not. However, it is only used for nominal level variables. The kappa's range is from .00 (agreement at chance level) to 1.00 (perfect agreement), a value that is less than .00 illustrates an agreement that is less than chance.

$$\frac{PA_o - PA_E}{1 - PA_E} \qquad (2)$$

PAo concerns the proportion agreement, observed, and PA_E refers to the proportion agreement that is expected by chance (Neuendorf, 2002).

Some other measures of reliability include Kripendorff's *alpha* (Krippendorff, 1980), Scott's *pi* (1955) and Lin's concordance correlation coefficient (Lin, 1989), each of which have their own advantages and disadvantages.

Although there "is no simple right way to do content analysis" (Weber, 1990, p. 13) most have the following elements in common. After the research question is asked a decision needs to be made on what will be analyzed or what social artifacts will be studied. Then a decision needs to be made on the unit of analysis. Following this a categorical system needs to be developed in which the responses can be filled. Next, it needs to be determined how the data will be coded. It is a good idea to take a sample or even do a pilot study to determine if the coding structure needs to be modified.

Methodology

Recently, some researchers have used a more uncommon coding scheme that entails multiple steps in coding. In the scheme coders first code a variable in a context for its presence (variable A). If in the experimental condition variable A exists the coders then look for or categorize a next variable (variable B). The process can either stop at this point or continue. Therefore, the process of coding the second variable is contingent upon the existence of the first variable.

Consider a hypothetical study examining aggressive behaviors of children in a classroom. Variable A is a particular instance, and can be observed through video taping, in class observation or vignette. In this situation there are two coders examining the interactions (coder 1 and 2). The coders either code the behavior as (1) not aggressive or (2) aggressive. After the experiment the results of the coders are compared for intercoder reliability. This is demonstrated in table 1 (C1va and C2va). Using Cohen's kappa, the intercoder reliability is .83; no problems exist in the reporting thus far.

Consider that the next behavior coded is dependent (contingent) upon whether or not the first behavior was identified as aggressive. Thus, if the behavior in the condition was (2) aggressive, was it (1) physical aggression or (2) verbal aggression? The coding process continues but the analysis is dependent upon the first code. It is at this point in reporting the results that a reliability reporting bias can occur. In table 1 (C1va and C1v2) the reporting of the behavior can be seen. There are 47 agreements between the coders and 4 disagreements, producing a kappa of .83. This represents excellent agreement beyond the role of chance (Banerjee et al. 1999).

Three instances exist however, where one of the coders moved on to coding the type of physical aggression (variable B) while the second coder did not. When reporting the reliability of variable B the researcher must include the non-agreements from variable A in order to give the reader an accurate assessment of the intercoder reliability. This does not always happen. Increasingly authors report the reliability without the addition of the non-agreements from the first variable under examination, which inflates reliability.

Consider in table 1 (C1vb, C2vb, C1vb2 and C1vb2) the reporting of variable B (type of physical aggression). In this situation there are 21 instances of aggressive behavior coded from the first condition. Coder 1 and 2 agreed on the type of the aggression in 18 of the 21 instances. If the researcher fails to include the non-agreements from the examination of variable A the reliability in the condition is .70: still a good measure of reliability beyond chance.

Compare those results to table 1 (C1vb3 and C2vb3) where the researcher includes the first wave of reliability assessments. The agreement is 18 out of 25 cases, producing a kappa of .49. This is considerably lower, and it is considered poor agreement beyond chance (Banerjee et al. 1999). Moreover, consider what would be the case if this continued in a study and the author failed to include the non-agreements for variables C, D, and E. In reporting the reliabilities for variable E, the reported score would be far removed from the true value.

Conclusion

A few suggestions follow concerning this problem. The first and simplest is for the researcher to report the reliability with the inclusion of all coded responses as was done in table 1 (C1vb3 and C2vb3). When this is done the reader has an accurate assessment of the true score concerning the reliability and can have more confidence in the conclusions the data support.

If a researcher believes that due to some aspect of the research design the inclusion of the non-agreements from the first condition is unwarranted, then he or she should outline the reason behind the exclusion of the non-agreements in the results section of the article or paper. Accompanying this should be the scores from each coder and an explanation indicating that the previous condition produced X number of agreements that is not calculated in the reliability kappa. Another alternative is for the researcher may include both reliability scores within the results.

The above example used only a few instances of disagreement between the coders. In a study that has more disagreement the reporting bias can be larger. Although there are no rules explaining exactly how a researcher should report reliability, care needs to be taken in reporting and the author needs to justify the use of any reporting scheme.

Table 1. Comparison of responses between Coder 1 and 2

C1va	C2va	C1vb	C2vb	C1vb1	C2vb2	C1vb3	C2vb3
1	1			1	1	1	0
1	1			1	1	1	1
1	1			1	1	1	1
1	1			1	1	1	1
2	1	1		2	2	2	0
2	2	1	1	2	2	1	1
2	2	1	1	2	1	2	2
2	2	1	1	1	1	2	2
2	1	2		1	1	2	1
1	1			1	2	1	1
1	1			1	1	0	2
1	1			2	2	0	2
1	1			2	2	1	1
2	2	1	1	2	2	1	2
2	2	2	2	1	2	1	1
2	2	2	2	2	2	2	2
2	2	2	1	2	2	2	2
2	2	1	1	2	2	2	2
1	2		2	2	2	1	2
1	1			2	2	2	2
1	1			2	2	2	2
1	1					2	2
1	2		2			2	2
1	1					2	2
1	1					2	2
1	1						
1	1						
2	2	1	1				
2	2	1	2				
2	2	1	1				
2	2	2	2				
2	2	2	2				
2	2	2	2				
2	2	1	2				
1	1						
1	1						
1	1						
1	1						
1	1						
1	1						
1	1						
1	1						
1	1						
1	1						
2	2	2	2				
2	2	2	2				
2	2	2	2				
2	2	2	2				
2	2	2	2				

C1va = Coder 1 variable A (presence of aggression).
C2va = Coder 2 variable A (presence of aggression).
C1vb and C2vb show the progression in coding from variable A to B.
C1vb1 and C2vb2 is the progression in coding from variable A to B collapsed.
C1vb3 and C2vb3 is the progression in coding from variable A to B with all inclusion of all instances of reliability assessments.

References

Banerjee, M., Capozzoli, M., McSweeney, L., Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, *27*, 2-23.

Berelson, B. (1952). *Content analysis in communication research*. Glencoe, IL. Free Press.

Berger, A. A. (1991). *Media research techniques*. Newbury Park, CA: Sage.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement of partial credit. *Psychological Bulletin*, *70*, 213-220.

Dimopoulous, K., Koulaidis, V., & Sklaveniti, S. (2003). Towards an analysis of visual images in school science textbooks and press articles about science and technology. *Research in Science Education*, *33*, 189 – 216.

Harmon, J. M., Hedrick, W. B., & Fox, E. A. (2000). A content analysis of vocabulary instruction in social studies textbooks for grades K-8. *The Elementary School Journal*, *100*, 253-271.

Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, *45*, 255-268.

Lombard, M., Snyder-Duch, J, & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting intercoder reliability. *Human Communication Research*, *28*, 587-604.

England, E. M., & Petro, K. D. (1988). Middle school students' perceptions of peer groups: Relative judgments about group characteristics. *The Journal of Early Adolescence*, *18*, 349-373.

George, A. L. (1959). *Propaganda analysis*. Evanston: Row and Peterson.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. CA: Sage.

Neuendorf, K.A. (2001). *The content Analysis Guidebook*. London: Sage Publications.

Plucker, J. A., & Beghetto, R. A. (2000). Needles in haystacks or field of plenty? A content analysis of popular creativity texts. *Gifted Child Quarterly*, *44*, 135-138.

Redding, S. G., & Ng, M. (1982). The role of 'face' in the organizational perceptions of Chinese managers. *Organizational Studies*, *3*, 201-219.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*, 321-325.

Seibold, D. R. (1998). Jurors intuitive rules for deliberation: A structurational approach to communication in jury decision making. *Communication Monographs*, *65*, 282-307.

Stern, B. L., & Harmon, R. R. (1998). The incidence and characteristics of disclaimers in children's television advertising. *Journal of Advertising*, *13*, 12-17.

Tamboini, R., Mastro, D. E., Chory-Assad, R. M., Huang, R. H. (2000). The color and crime of the court: A content analysis of minority representation. *Journalism and Mass Communication Quarterly*, *77*, 639-653.

Weber, R. P. (1990). *Basic content analysis*. CA: Sage.

Zittleman, K., & Sadker, D. (2002). Gender bias in teacher education texts. *Journal of Teacher Education*, *53* (2), 168-180.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, *103*, 374-378.