Quantitative Methods for Estimating the Reliability of Qualitative Data

Jason W. Davey
Fenwal, Inc.

P. Cristian Gugiu
Western Michigan University

Chris L. S. Coryn
Western Michigan University

Background: Measurement is an indispensable aspect of conducting both quantitative and qualitative research and evaluation. With respect to qualitative research, measurement typically occurs during the coding process.

Purpose: This paper presents quantitative methods for determining the reliability of conclusions from qualitative data sources. Although some qualitative researchers disagree with such applications, a link between the qualitative and quantitative fields is successfully established through data collection and coding procedures.

Setting: Not applicable.

Intervention: Not applicable.

Research Design: Case study.

Data Collection and Analysis: Narrative data were collected from a random sample of 528 undergraduate students and 28 professors.

Findings: The calculation of the kappa statistic, weighted kappa statistic, ANOVA Binary Intraclass Correlation, and Kuder-Richardson 20 is illustrated through a fictitious example. Formulae are presented so that the researcher can calculate these estimators without the use of sophisticated statistical software.

Keywords: *qualitative coding; qualitative methodology; reliability coefficients*

The rejection of using quantitative methods for assessing the reliability of qualitative findings by some qualitative researchers is both frustrating and perplexing from the vantage point of quantitative and mixed method researchers. While the need to distinguish one methodological approach from another is understandable, the reasoning sometimes used to justify the wholesale rejection of all concepts associated with quantitative analysis are replete with mischaracterizations, overreaching arguments, and inadequate substitutions.

One of the first lines of attack against the use of quantitative analysis centers on philosophical arguments. Healy and Perry (2000), for example, characterize qualitative methods as flexible, inductive, and multifaceted, whereas quantitative methods are often characterized as inflexible and fixed. Moreover, most qualitative researchers view quantitative methods as characteristic of a positivist paradigm (e.g., Stenbacka, 2001; Davis, 2008; Paley, 2008)—a term that has come to take on a derogatory connotation. Paley (2008) states that "doing quantitative research entails commitment to a particular ontology and, specifically, to a belief in a single, objective reality that can be described by universal laws" (p. 649).

However, quantitative analysis should not be synonymous with the positivist paradigm because statistical inference is concerned with probabilistic, as opposed to deterministic, conclusions. Nor do statisticians believe in a universal law measured free of error. Rather, statisticians believe that multiple truths may exist and that despite the best efforts of the researcher these truths are measured with some degree of error. If that were not the case, statisticians would ignore interaction effects, assume that measurement errors do not exist, and fail to consider whether differences may exist between groups. Yet, most statisticians consider all these factors before formulating their conclusions. While statisticians may be faulted for paying too much attention to measures of central tendency (e.g., mean, median) at the expense of interesting outliers, this is not the same as believing in one all-inclusive truth.

The distinction between objective research and subjective research also appears to emerge from this paradigm debate. Statisticians are portrayed as detached and neutral investigators while qualitative researchers are portrayed as embracing personal viewpoints and even biases to describe and interpret the subjective experience of the phenomena they study (Miller, 2008). While parts of these characterizations do, indeed, differentiate between the two groups of researchers, they fail to explain why a majority of qualitative researchers dismiss the use of statistical methods. After all, the formulas used to conduct such analyses do not know or care whether the data were gathered using an objective rather than a subjective method. Moreover, certain statistical methods lend themselves to, and were even specifically developed for, the analysis of qualitative data (e.g., reliability analysis). Other qualitative researchers have come to equate positivism, and by extension quantitative analysis, with causal explanations (Healy & Perry, 2000). To date, the gold standard for substantiating causal claims is through the use of a well-conducted experimental design. However, the implementation of an experimental design does not necessitate the use of quantitative analysis. Furthermore, quantitative analysis may be conducted for any type of research design, including qualitative research, as is the central premise of this paper.

For some qualitative researchers (e.g., Miller, 2008; Stenbacka, 2001), the wholesale rejection of all concepts perceived to be quantitative has extended to general research concepts like reliability and validity. According to Stenbacka (2001), "reliability has no relevance in qualitative research, where it is impossible to differentiate between researcher and method" (p. 552). From the perspective of quantitative research, this statement is inaccurate because several quantitative methods have been developed for differentiating between the researcher, data collection method, and informant (e.g., generalizability theory), provided, of course, data are available for two or more researchers and/or methods.

Stenbacka (2001) also objected to traditional forms of validity because "the purpose in qualitative research never is to measure anything. A qualitative method seeks for a certain quality that is typical for a phenomenon or that makes the phenomenon different from others" (p. 551). It would seem to the present authors, however, that this notion is inconsistent with traditional qualitative research. Measurement is a indispensable aspect of conducting research, regardless if it is quantitative or qualitative.

With respect to qualitative research, measurement occurs during the coding process. Illustrating the integral nature of coding in qualitative research, Benaquisto (2008) noted:

> The coding process refers to the steps the researcher takes to identify, arrange, and systematize the ideas, concepts, and categories uncovered in the data. Coding consists of identifying potentially interesting events, features, phrases, behaviors, or stages of a process and distinguishing them with labels. These are then further differentiated or integrated so that they may be reworked into a smaller number of categories, relationships, and patterns so as to tell a story or communicate conclusions drawn from the data. (p. 85)

Clearly, in absence of utilizing a coding process, researchers would be forced to provide readers with all of the data, which, in turn, would place the burden of interpretation on the reader. However, while the importance of coding to qualitative research is self-evident to all those who have conducted such research, the role of measurement may not be as obvious. In part, this may be attributed to a misunderstanding on the part of many researchers as to what is measurement.

Measurement is the process of assigning numbers, symbols, or codes to phenomena (e.g., events, features, phrases, behaviors) based on a set of prescribed rules (i.e., a coding rubric). There is nothing inherently quantitative about this process or, at least, there does not need to be. Moreover, it does not limit qualitative research in any way. In fact, many times, measurement may only be performed in a qualitative context.

For example, suppose that a researcher conducts an interview with an informant who states that "the bathrooms in the school are very dirty." Now further suppose that the researcher developed a coding rubric, which, for the sake of simplicity, only contained two levels: cleanliness and academic performance. Clearly, the informant's statement addressed the first level (cleanliness) and not the second. Whether the researcher chooses to assign this statement a checkmark for the cleanliness category or a 1, and an 'X' or 0 (zero) for the academic performance category, does not make a difference. The researcher clearly used his or her judgment to transform the raw statement made by the informant into a code. However, when the researcher decided that the statement best represented cleanliness and not academic performance, he or she also performed a measurement process. Therefore, if one accepts this line of reasoning, qualitative research depends upon measurement to render judgments. Furthermore, three questions may be asked. First, does statement X fit the definition of code Y? Second, how many of the statements collected fit the definition of code Y? And third, how reliable is the definition of code Y for differentiating between statements within and across researchers (i.e., intrarater and interrater reliability, respectively)?

Fortunately, not every qualitative researcher has accepted Stenbacka's notion, in part, because qualitative researchers, like quantitative researchers, compete for funding and therefore, must persuade funders of the accuracy of their methods and results (Cheek, 2008). Consequently, the concepts of reliability and validity permeate qualitative research. However, owing to the desire to differentiate itself from quantitative research, qualitative researchers have espoused the use of "interpretivist alternatives" terms (Seale, 1999). Some of the most popular terms substituted for reliability include confirmability, credibility, dependability, and replicability (Coryn, 2007; Golafshani, 2003; Healy & Perry, 2000; Morse, Barrett, Mayan, Olson, & Spiers, 2002; Miller, 2008; Lincoln & Guba, 1985).

In the qualitative tradition, confirmability is concerned with confirming the researcher's interpretations and conclusions are grounded in actual data that can be verified (Jensen, 2008; Given & Saumure, 2008). Researchers may address this reliability indicator through the use of multiple coders, transparency, audit trails, and member checks. Credibility, on the other hand, is concerned with the research methodology and data sources used to establish a high degree of harmony between the raw data and the researcher's interpretations and conclusions. Various means can be used to enhance credibility, including accurately and richly describing data, citing negative cases, using multiple researchers to review and critique the analysis and findings, and conducting member checks (Given & Saumure, 2008; Jensen, 2008; Saumure & Given, 2008). Dependability recognizes that the most appropriate research design cannot be completely predicted *a priori*. Consequently, researchers may need to alter their research design to meet the realities of the research context in which they conduct the study, as compared to the context they predicted to exist a priori (Jensen, 2008). Dependability can be addressed by providing a rich description of the research procedures and instruments used so that other researchers may be able to collect data in similar ways. The idea being that if a different set of researchers use similar methods then they should reach similar conclusions (Given & Saumure, 2008). Finally, replicability is concerned with repeating a study on participants from a similar background as the original study. Researchers may address this reliability indicator by conducting the new study on participants with similar demographic variables, asking similar questions, and coding data in a similar fashion to the original study (Firmin, 2008).

Like qualitative researchers, quantitative researchers have developed numerous definitions of reliability, including interrater and intrarater reliability, test-retest reliability, internal consistency, and interclass correlations to name a few (Crocker & Algina, 1986; Hopkins, 1998). A review of the qualitative alternative terms revealed them to be indirectly associated with quantitative notions of reliability. However, although replicability is conceptually equivalent to test-retest reliability, the other three terms appear to describe research processes tangentially related to reliability. Moreover, they have two major liabilities. First, they place the burden of assessing reliability squarely on the reader. For example, if a reader wanted to determine the confirmability of a finding they would need to review the audit trail and make an independent assessment. Similar reviews of the data would be necessary, if a reviewer wanted to assess the credibility of a finding or dependability of a study design.

Second, they fail to consider interrater reliability, which, in our experience, accounts for a considerable amount, if not a majority, of the variability in findings in qualitative studies. Interrater reliability is concerned with the degree to which different raters or coders appraise the same information (e.g., events, features, phrases, behaviors) in the same way (van den Hoonaard, 2008). In other words, do different raters interpret qualitative data in similar ways? The process of conducting an interrater reliability analysis, which is detailed in the next section, is relatively straightforward. Essentially, the only additional step beyond development and finalization of a coding rubric is that, at least two or more raters must *independently* rate all of the qualitative data using the coding rubric. Although collaboration, in the form of consensus agreement, may be used to finalize ratings after each rater has had an opportunity to rate all data, each rater must work independently of the other to reduce bias in the first phase of analysis. Often, this task is greatly facilitated by use of a database system that, for example, (1) displays the smallest codable unit of a transcript (e.g., a single sentence), (2) presents the available coding options, and (3) records the rater's code before displaying the next codable unit.

While it is likely that qualitative researchers who prescribe to a constructionist paradigm may object to the constraint of forcing qualitative researchers to use the same coding rubric for a study, rather than developing their own, this is an indispensable process for attaining a reasonable level of interrater reliability. An example of the perils of not attending to this issue may be found in an empirical study conducted by Armstrong, Gosling, Weinman, and Marteau (1997). Armstrong and his colleagues invited six experienced qualitative researchers from Britain and the United States to analyse a transcript (approximately 13,500 words long) from a focus group comprised of adults living with cystic fibrosis that was convened to discuss the topic of genetic screening. In return for a fee, each researcher was asked to prepare an independent report in which they identified and described the main themes that emerged from the focus group discussion, up to a maximum of five. Beyond these instructions, each researcher was permitted to use any method for extracting the main themes they felt was appropriate. Once the reports were submitted, they were thematically analyzed by one of the authors, who deliberately abstained from reading the original transcript to reduce external bias.

The results uncovered by Armstrong and his colleagues paint a troubling picture. On the surface, it was clear that a reasonable level of consensus in the identification of themes was achieved. Five of the six researchers identified five themes, while one identified four themes. Consequently, only four themes are discussed in the article: visibility; ignorance; health service provision; and genetic screening. With respect to the presence of each theme, there was unanimous agreement for the visibility and genetic screening themes, while the agreement rates were slightly lower for the ignorance and health service provision themes (83% and 67%, respectively). Overall, these are good rates of agreement. However, a deeper examination of the findings revealed two troubling issues. First, a significant amount of disagreement existed with respect to how the themes were organized. Some researchers classified a theme as a basic structure whereas others organized it under a larger basic structure (i.e., gave it less importance than the overarching theme they assigned it to). Second, a significant amount of disagreement existed with respect to the manner in which themes were interpreted. For example, some of the researchers felt that the ignorance theme suggested a need for further education, other researchers raised concern about the eugenic threat, and the remainder thought it provided parents with choice. Similar inconsistencies with regard to interpretability occurred for the genetic screening theme where three researchers indicated that genetic screening provided parents with choice while one linked it with the eugenic threat.

These results serve as an example of how “reality” is relative to the researcher doing the interpretation. However, they also demonstrate how the quality of a research finding requires knowledge of the degree to which consensus is reached by knowledgeable researchers. Clearly, by this statement, we are assuming that reliability of findings across different researchers is a desirable quality. There certainly may be instances in which reliability is not important because one is only interested in the findings of a specific researcher, and the perspectives of others are not desired. That being the case, one may consider examining intrarater reliability. In all other instances, however, it is reasonable to assume that it is desirable to differentiate between the perspectives of the informants and those of the researcher. In other words, are the researcher’s findings truly grounded in the data or do they reflect his or her personal ideological perspectives. For a politician, for example, knowing the answer to this question may mean the difference between passing and rejecting a policy that allows parents to genetically test embryos.

Although qualitative researchers can address interrater reliability by following the method used by Armstrong and his colleagues, the likelihood of achieving a reasonable level of reliability will be low simply due to researcher differences (e.g., the labels used to describe themes, structural organization of themes, importance accorded to themes, interpretation of data). In general, given the importance of reducing the variability in research findings attributed solely to researcher variability, it would greatly benefit qualitative researchers to utilize a common coding rubric. Furthermore, use of a common coding rubric does not greatly interfere with normal qualitative procedures, particularly if consensus is reached beforehand by all the researchers on the rubric that will be used to code all the data. Of equal importance, this procedure permits the researcher to remain to be the instrument by which data are interpreted (Brodsky, 2008).

Reporting the results of, to this point, this qualitative process should considerably improve the credibility of research findings. However, three issues still remain. First, reporting the findings of multiple researchers places the burden of synthesis on the reader. Therefore, researchers should implement a method to synthesize all the findings through a consensus-building procedure or averaging results, where appropriate and possible. Second, judging the reliability of a study requires that deidentified data are made available to anyone who requests it. While no one, to the best of our knowledge, has studied the degree to which this is practiced, our experience suggests it is not prevalent in the research community. Third, reporting the findings of multiple researchers will only permit readers to get an approximate sense of the level of interrater reliability or whether it meets an acceptable standard. Moreover, comparisons between the reliability of the study to another qualitative study are impractical for complex studies.

Fortunately, simple quantitative solutions exist that enable researchers to report the reliability of their conclusions rather than shift the burden to the reader. The present paper will expound upon four quantitative methods for calculating interrater reliability that can be specifically applied to qualitative data and thus, should not be regarded as products of a positivist position. In fact, reliability estimates, which can roughly be conceptualized as the degree to which variability of research findings are or are not due to differences in researchers, illustrate the degree to which reality is socially constructed or not. Data that are subject to a wide range of interpretations will likely produce low reliability estimates, whereas data whose interpretations are consistent will likely produce high reliability estimates. Finally, calculating interrater reliability in addition to reporting a narrative of the discrepancies and consistencies between researchers can be thought of as a form of methodological triangulation.

Method

Data Collection Process

Narrative data were collected from 528 undergraduate students and 28 professors randomly selected from a university population. Data were collected with the help of an open-ended survey that asked respondents to identify the primary challenges facing the university that should be immediately addressed by the university's administration. Data were transcribed from the surveys to an electronic database (Microsoft Access) programmed to resemble the original questionnaire. Validation checks were performed by upper-level graduate students to assess the quality of the data entry process. Corrections to the data entered into the database were made by the graduate students in the few instances in which discrepancies were found between the responses noted on the survey and those entered in the database. Due to the design of the original questionnaire, which encouraged respondents to bullet their responses, little additional work was necessary to further break responses into the smallest codable units (typically 1-3 sentences). That said, it was possible for the smallest codable units to contain multiple themes although the average number of themes was less than two per unit of analysis.

Coding Procedures

Coding qualitative data is an arduous task that requires iterative passes through the raw data in order to generate a reliable and comprehensive coding rubric. This task was conducted by two experienced qualitative researchers who independently read the original narratives and identified primary and secondary themes, categorized these themes based on their perception of the internal structure (selective coding; Benaquisto, 2008), and produced labels for each category and subcategory based on the underlying data (open coding; Benaquisto, 2008). Following this initial step, the two researchers further differentiated or integrated their individual coding rubric (axial coding; Benaquisto, 2008) into a unified coding rubric. Using the unified coding rubric, the two researchers attempted an initial coding of the raw data to determine (1) the ease with which the coding rubric could be applied, (2) problem areas that needed further clarification, (3) the trivial categories that could be eliminated or integrated with other categories, (4) the extensive categories that could be further refined to make important distinctions, and (5) the overall coverage of the coding rubric. Not surprisingly, several iterations were necessary before the coding rubric was finalized. In the following section, for ease of illustration, reliability estimates are presented only for a single category.

Statistical Procedures

Very often, coding schemes follow a binomial distribution. That is, coders indicate whether a particular theme either is or is not present in the data. When two or more individuals code data to identify such themes and patterns, the reliability of coder's efforts can be determined, typically by coefficients of agreement. This type of estimate can be used as a measure that objectively permits a researcher to substantiate that his or her coding scheme is replicable.

Most estimators for gauging the reliability of continuous agreement data predominately evolved from psychometric theory (Cohen, 1968; Lord & Novick, 1968; Gulliksen, 1950; Rozeboom, 1966). Similar methods for binomial agreement data shortly followed (Cohen, 1960; Lord & Novick, 1968). Newer forms of these estimators, called binomial intraclass correlation coefficients (ICC), were later developed to handle more explicit patterns in agreement data (Fleiss & Cuzick, 1979; Kleinman, 1973; Lipsitz, Laird, & Brennan, 1994; Mak, 1988; Nelder & Pregibon, 1987; Smith, 1983; Tamura & Young, 1987; Yamamoto & Yanagimoto, 1992).

In this paper four methods that can be utilized to assess the reliability of binomial coded agreement data are presented. These estimators are the kappa statistic (κ), the weighted kappa statistic (κ_W), the ANOVA binary ICC, and the Kuder-Richardson 20 (KR-20). The kappa statistic was one of the first statistics developed for assessing the reliability of binomial data between two or more coders (Cohen, 1960; Fleiss, 1971). A modified version of this statistic introduced the use of numerical weights. This statistic allows the user to apply different probability weights to cells in a contingency table (Fleiss, Cohen, & Everitt, 1969) in order to apply different levels of importance to various coding frequencies. The ANOVA binary ICC is based on the mean squares from an analysis of variance (ANOVA) model modified for binomial data (Elston, Hill, &

Smith, 1977). The last estimator was developed by Kuder and Richardon (1937), and is commonly known as KR-20 or KR (20), because it was the 20th numbered formula in their seminal article. This estimator is based on the ratios of agreement to the total discrete variance.

These four reliability statistics are functions of *i* x *j* contingency tables, also known as cross-tabulation tables. The current paper will illustrate the use of these estimators for a study dataset that comprises the binomial coding patterns of two investigators. Because these coding patterns are from two coders and the coded responses are binomial (i.e., theme either is or is not present in a given interview response, the contingency table has two rows (i = 2) and two columns (j = 2).

The layout of this table is provided in Table 1. The first cell, denoted (i_1 = Present, j_1 = Present), of this table consists of the total frequency of cases where Coder 1 and Coder 2 both agree that a theme is present in the participant interview responses. The second cell, denoted (i_1 = Present, j_2 = Not Present), of this table consists of the total frequency of cases where Coder 1 feels that a theme is present in the interview responses, and the second coder does not agree with this assessment. The third cell, denoted (i_2 = Not Present, j_1 = Present), of this table consists of the total frequency of cases where Coder 2 feels that a theme is present, and the first coder does not agree with this assessment. The fourth cell, denoted (i_2 = Not Present, j_1 = Not Present), of this table consists of the total frequency of cases where both Coder 1 and Coder 2 agree that a theme is not present in the interview responses (Soeken & Prescott, 1986).

Table 1
General Layout of Binomial Coder Agreement Patterns for Qualitative Data

		Coder 1	
		Theme Present (j_1)	Theme Not Present (j_2)
Coder 2	Theme Present (i_1)	Cell$_{11}$	Cell$_{21}$
	Theme Not Present (i_2)	Cell$_{12}$	Cell$_{22}$

Participants

Interview data were collected for and transcribed from 28 professors and 528 undergraduate students randomly selected from a university population. The binomial coding agreement patterns for these two groups of interview participants are provided in Table 2 and Table 3.

For the group of professor and student interview participants, the coders agreed that one professor and 500 students provided a response that pertains to overall satisfaction of university facilities.

Coder 1 felt that an additional seven professors and two students made a response pertinent to overall satisfaction, whereas Coder 2 did not feel that response from these two professors pertained to the interview response of interest. Coder 2 felt that one professor and one student made a response pertinent to overall satisfaction, whereas Coder 1 did not feel that response from this professor pertained to the interview response of interest. Coder 1 and Coder 2 agreed that responses from the final 19 professors and 25 students did not pertain to the topic of interest.

Table 2
Binomial Coder Agreement Patterns for Professor Interview Participants

		Coder 1	
		Theme Present (j_1)	Theme Not Present (j_2)
Coder 2	Theme Present (i_1)	1	1
	Theme Not Present (i_2)	7	19

Table 3
Binomial Coder Agreement Patterns for Student Interview Participants

		Coder 1	
		Theme Present (j_1)	Theme Not Present (j_2)
Coder 2	Theme Present (i_1)	500	1
	Theme Not Present (i_2)	2	25

Four Estimators for Calculating the Reliability of Qualitative Data

Kappa

According to Brennan and Hays (1992), the κ statistic “determines the extent of agreement between two or more judges exceeding that which would be expected purely by chance” (p. xx). This statistic is based on the observed and expected level of agreement between two or more raters with two or more levels. The observed level of agreement (p_o) equals the frequency of records where both coders agree that a theme is present plus the frequency of records where both coders agree that a theme is not present divided by the total number of ratings. The expected level of agreement (p_e) equals the summation of the cross product of the marginal probabilities. In other words, this is the expected rate of agreement by random chance alone. The kappa statistic (κ) then equals (p_o-p_e)/(1-p_e). The traditional formulae for p_o and p_e are

$$p_o = \sum_{i=1}^{c}\sum_{j=1}^{c} p_{ij} \quad \text{and} \quad p_e = \sum_{i=1}^{c}\sum_{j=1}^{c} p_{i.}\, p_{.j} \quad ,$$

where c denotes the total number of cells, i denotes the i^{th} row, and j denotes the j^{th} column (Fleiss, 1971; Soeken & Prescott, 1986). These formulae are illustrated in Table 4.

Table 4
2 x 2 Contingency Table for the Kappa Statistic

		Coder 1		Marginal Row Probabilities
		Theme present	Theme not present	p$_{i.}$
Coder 2	Theme present	c_{11}	c_{21}	$p_{1.} = (c_{11} + c_{21}) / N$
	Theme not present	c_{12}	c_{22}	$p_{2.} = (c_{12} + c_{22}) / N$
Marginal Column Probabilities	$p_{.j}$	$p_{.1} = (c_{11} + c_{12})$ / N	$p_{.2} = (c_{21} + c_{22}) / N$	$N = (c_{11} + c_{21} + c_{12} + c_{22})$

$$p_o = \sum_{i=1}^{c}\sum_{j=1}^{c} p_{ij} = \frac{c_{11} + c_{22}}{N}$$

and

$$p_e = \sum_{i=1}^{c}\sum_{j=1}^{c} p_{i.}p_{.j} = p_{1.}p_{.1} + p_{2.}p_{.2}$$

Estimates from professor interview participants for calculating the kappa statistic are provided in Table 5. The observed level of agreement for professors is (1+19)/556 = 0.0360. The expected level of agreement for professors is 0.0036(0.0144) + 0.0468(0.0360) = 0.0017.

Table 5
Estimates from Professor Interview Participants for Calculating the Kappa Statistic

		Coder 1		Marginal Row Probabilities
		Theme present	Theme not present	$p_{i.}$
Coder 2	Theme present	1	1	$p_{1.}$ = 2/556 = 0.0036
	Theme not present	7	19	$p_{2.}$ = 26/556 = 0.0468
Marginal Column Probabilities	$p_{.j}$	$p_{.1}$ = 8/556 = 0.0144	$p_{.2}$ = 20/556 = 0.0360	N = 28 + 528 = 556

Estimates from student interview participants for calculating the kappa statistic are provided in Table 6. The observed level of agreement for students is (500+25)/556 = 0.9442. The expected level of agreement for students is 0.9011(0.9029) + 0.0486(0.0486) = 0.8160.

Table 6
Estimates from Student Interview Participants for Calculating the Kappa Statistic

		Coder 1		Marginal Row Probabilities
		Theme present	Theme not present	$p_{i.}$
Coder 2	Theme present	500	1	$p_{1.}$ = 501/556 = 0.9011
	Theme not present	2	25	$p_{2.}$ = 27/556 = 0.0486
Marginal Column Probabilities	$p_{.j}$	$p_{.1}$ = 502/556 = 0.9029	$p_{.2}$ = 26/556 = 0.0468	N = 556

The total observed level of agreement for the professor and student interview groups is $p_o = 0.0360 + 0.9442 = 0.9802$. The total expected level of agreement for the professor and student interview groups is $p_e = 0.0017 + 0.8160 = 0.8177$. For the professor and student and professor groups, the kappa statistic equals $\kappa = (0.9802 – 0.8177)/(1 – 0.8177) = 0.891$. The level of agreement between the two coders is 0.891 beyond that which is expected purely by chance.

Weighted Kappa

The reliability coefficient, κ_W, has the same interpretation as the kappa statistic, κ, but the researcher can differentially weight each cell to reflect varying levels of importance. According to Cohen (1968), κ_W is “the proportion of weighted agreement corrected for chance, to be used when different kinds of disagreement are to be differentially weighted in the agreement index” (p. xx). As an example, the frequencies of coding patterns where both raters agree that a theme is present can be given a larger weight than patterns where both raters agree that a theme is not present. The same logic can be applied where the coders disagree on the presence of a theme in participant responses.

The weighted observed level of agreement (p_{ow}) equals the frequency of records where both coders agree that a theme is present times a weight plus the frequency of records where both coders agree that a theme is not present times another weight divided by the total number of ratings. The weighted expected level of agreement (p_{ew}) equals the summation of the cross product of the marginal probabilities, where each cell in the contingency table has its own weight. The weighted kappa statistic κ_W then equals $(p_{ow}-p_{ew})/(1-p_{ew})$. The traditional formulae for p_{ow} and p_{ew} are

$$p_{ow} = \sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}\, p_{ij} \text{ and } p_{ew} = \sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}\, p_{i.}\, p_{.j}\ ,$$

where c denotes the total number of cells, i denoted the i^{th} row, j denotes the j^{th} column, and w_{ij} denotes the i, j^{th} cell weight (Fleiss, Cohen, & Everitt, 1969; Everitt, 1968). These formulae are illustrated in Table 7.

Table 7
2 x 2 Contingency Table for the Weighted Kappa Statistic

		Coder 1		Marginal Row Probabilities
		Theme present	Theme not present	$p_{i.}$
Coder 2	Theme present	$w_{11}c_{11}$	$w_{21}c_{21}$	$p_{1.} = (w_{11}c_{11} + w_{21}c_{21})$ / N
	Theme not present	$w_{12}c_{12}$	$w_{22}c_{22}$	$p_{2.} = (w_{12}c_{12} + w_{22}c_{22})$ / N
Marginal Column Probabilities	$p_{.j}$	$p_{.1} = (w_{11}c_{11} + w_{12}c_{12}) / N$	$p_{.2} = (w_{21}c_{21} + w_{22}c_{22}) / N$	$N = (c_{11} + c_{21} + c_{12} + c_{22})$

$$p_o = \sum_{i=1}^{c}\sum_{j=1}^{c} p_{ij} = \frac{w_{11}c_{11} + w_{22}c_{22}}{N}$$ and

$$p_e = \sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}\, p_{i.}\, p_{.j}$$.

Karlin, Cameron, and Williams (1981) provided three methods for weighting probabilities as applied to the calculation of a kappa statistic. The first method equally weights each pair of observations. This weight is calculated as $\mathrm{w_i} = \frac{\mathrm{n_i}}{\mathrm{N}}$, where n_i is the sample size of each cell and N is the sum of the sample sizes from all of cells of the contingency table. The second method equally weights each group (e.g., undergraduate students and professors) irrespective of its size. These weights can be calculated as $\mathrm{w_i} = \frac{1}{\mathrm{kn_i}(\mathrm{n_i} - 1)}$, where k is the number of groups (e.g., k = 2). The last method weights each cell according to the sample size in each cell. The formula for this weighting option is $\mathrm{w_i} = \frac{1}{\mathrm{N}(\mathrm{n_i} - 1)}$.

There is no single standard for applying probability weights to each cell in a contingency table. For this study, the probability weights used are provided in Table 8. In the first row and first column, the probability weight is 0.80. This weight was chosen arbitrarily to reflect the overall level of importance in the agreement of a theme being present as identified by both coders. In the second row and first column, the probability weight is 0.10. In the first row and second column, the probability weight is 0.09. These two weights were used to reduce the impact of differing levels of experience in qualitative research between the two raters. In the second row and second column, the probability weight is 0.01. This weight was employed to reduce the effect of the lack of existence of a theme from the interview data.

Table 8
Probability Weights on Binomial Coder Agreement Patterns for Professor and Student Interview Participants

		Coder 1	
		Theme Present (j_1)	Theme Not Present (j_2)
Coder 2	Theme Present (i_1)	0.80	0.09
	Theme Not Present (i_2)	0.10	0.01

Estimates from professor interview participants for calculating the weighted kappa statistic are provided in Table 9. The observed level of agreement for professors is [0.8(1)+0.01(19)]/556 = 0.0018. The expected level of agreement for professors is 0.0016(0.0027) + 0.0016(0.0005) = 0.00001.

Table 9
Estimates from Professor Interview Participants for Calculating the Weighted Kappa Statistic

		Coder 1		Marginal Row Probabilities
		Theme present	Theme not present	$p_{i.}$
Coder 2	Theme present	0.8(1) = 0.8	0.09(1) = 0.09	$p_{1.}$ = 0.89/556 = 0.0016
	Theme not present	0.1(7) = 0.7	0.01(19) =0.19	$p_{2.}$ = 0.89/556 = 0.0016
Marginal Column Probabilities	$p_{.j}$	$p_{.1}$ = 1.5/556 = 0.0027	$p_{.2}$ = 0.28/556 = 0.0005	N = 28 + 528 = 556

Estimates from professor interview participants for calculating the weighted kappa statistic are provided in Table 10. The observed level of agreement for professors is [0.8(500)+0.01(25)]/556 = 0.7199. The expected level of agreement for professors is 0.7196(0.7198) + 0.0008(0.0006) = 0.5180.

Table 10
Estimates from Student Interview Participants for Calculating the Weighted Kappa Statistic

		Coder 1		Marginal Row Probabilities
		Theme present	Theme not present	$p_{i.}$
Coder 2	Theme present	0.8(500) = 400	0.09(1) = 0.09	$p_{1.}$ = 400.09/556 = 0.7196
	Theme not present	0.1(2) = 0.2	0.01(25) =0.25	$p_{2.}$ = 0.45/556 = 0.0008
Marginal Column Probabilities	$p_{.j}$	$p_{.1}$ = 400.2/556 = 0.7198	$p_{.2}$ = 0.34/556 = 0.0006	N = 28 + 528 = 556

The total observed level of agreement for the professor and student interview groups is $p_{ow} = 0.0018 + 0.7199 = 0.7217$. The total expected level of agreement for the professor and student interview groups is $p_{ew} = 0.00001 + 0.5180 = 0.5181$. For the professor and student and professor groups, the weighted kappa statistic equals $\kappa_W = (0.7217 - 0.5181)/(1 - 0.5181) = 0.423$. The level of agreement between the two coders is 0.423 beyond that which is expected purely by chance after applying importance weights to each cell. This reliability statistic is notably smaller than the unadjusted kappa statistic because of the number of down-weighted cases where both coders agreed that the theme is not present in the interview responses.

ANOVA Binary ICC

From the writings of Shrout and Fleiss (1979), the currently available ANOVA Binary ICC that is appropriate for the current data set is based on what they refer to as ICC(3,1). More specifically, this version of the ICC is based on within mean squares and between mean squares for two or more coding groups/categories from an analysis of variance model modified for binary response variables by Elston (1977). This reliability statistic measures the consistency of the two ratings (Shrout and Fleiss, 1979), and is appropriate when two or more raters rate the same interview participants for some item of interest. ICC(3,1) assumes that the raters are fixed; that is, the same raters are utilized to code multiple sets of data. The statistic ICC(2,1) that assumes the coders are randomly selected from a larger population of raters (Shrout and Fleiss, 1979) is recommended for use but not currently available for binomial response data.

The traditional formulae for these mean squares within and between along with an adjusted sample size estimate are provided in Table 11. In these formulae, k denotes the total number of groups or categories. Y_i denotes the frequency of agreements (both coders indicate a theme is present, or both coders indicate a theme is not present) between coders for the i^{th} group or category, n_i is the total sample size for the i^{th} group or category, and N is the total sample size across all groups or categories. Using these estimates, the reliability estimate equals $\hat{\rho}_{AOV} = \frac{MS_B - MS_W}{MS_B + (n_0 - 1)MS_W}$ (Elston, Hill, & Smith, 1977; Ridout, Demétrio, & Firth, 1999).

Estimates from professor and student interview participants for calculating the ANOVA Binary ICC are provided in Table 11. Given that *k* = 2 and *N* = 556, the adjusted sample size equals 54.2857. The within and between mean squares equal 0.0157 and 2.0854, respectively. Using these estimates, the ANOVA binary ICC equals

$$\frac{MS_B - MS_W}{MS_B + (n_0 - 1)MS_W} = \frac{2.0854 - 0.0157}{2.0854 + (54.5827 - 1)0.0157}$$

= 0.714, which denotes the consistency of coding between the two coders on the professor and student interview responses.

Table 11
Formulae and Estimates from Professor and Student Interview Participants for Calculating the ANOVA Binary ICC

Description of Statistic	Statistic	Formula
Mean Squares Within	MS_W	$\frac{1}{N-k}\left[\sum_{i=1}^{k} Y_i - \sum_{i=1}^{k} \frac{Y_i^2}{n_i}\right] = \frac{1}{556-2}\left[545 - 536.303\right] = 0.0157$
Mean Squares Between	MS_B	$\frac{1}{k-1}\left[\sum_{i=1}^{k} \frac{Y_i^2}{n_i} - \frac{1}{N}\left(\sum_{i=1}^{k} Y_i\right)^2\right] = \frac{1}{2-1}\left[536.303 - \frac{545^2}{556}\right] = 2.0854$
Adjusted Sample Size	n_0	$\frac{1}{k-1}\left[N - \frac{1}{N}\sum_{i=1}^{k} n_i^2\right] = \frac{1}{2\text{-}1}\left[556 \text{-} \frac{1}{556}(528^2 + 28^2)\right] = 54.5827$

Note: ΣY_i denotes the total number of cases where both coders indicate that a theme either is or is not present in a given response.

Kuder-Richardson 20

In their landmark article, Kuder and Richardson (1937) presented the derivation of the KR-20 statistic, a coefficient that they used to determine the reliability of test items. This estimator is a function of the sample size, summation of item variances, and total variance. Two observations in these formulae require further inquiry. These authors do not appear to discuss the distributional requirements of the data in relation to the calculation of the correlation r_{ii}, possibly due to its time of development in relation to the infancy of mathematical statistics. This vagueness has lead to some incorrect calculations of the KR-20. Crocker and Algina (1986) present examples on the calculation of the KR-20 in Table 7.2 based on data from Table 7.1 (pp. 136-140). In Table 7.1, the correlation on the two split-halves is presented as $\hat{\rho}_{AB} = 0.34$. It is not indicated that this statistic is the Pearson correlation. This is problematic because this statistic assumes that the two random variables are continuous, when in actuality they are discrete. An appropriate statistic is Kendall-τ_c and this correlation equals 0.35. As can be seen, the correlation may be notably underestimated as well as the KR-20 if the incorrect distribution is assumed. For the remainder of this paper, the Pearson correlation will be substituted with the Kendall-τ_c correlation.

Second, Kuder and Richardson (1937) present formulae for the calculation of σ_t^2 and r_{ii} that are not mutually exclusive. This lack of exclusiveness has caused some confusion in appropriate calculations of the total variance σ_t^2. Lord and Novick (1968) indicated that this statistic is equal to coefficient α (continuous) under certain circumstances, and Crocker and Algina (1986) elaborated on this statement by indicating "This formula is identical to coefficient alpha with the substitution of p_iq_i for $\hat{\sigma}_i^2$" (p. 139). This is unfortunately incomplete.

Not only must this substitution be made for the numerators variances, the denominator variances must also be adjusted in the same manner. That is, if the underlying distribution of the data is binomial, all estimators should be based on the level of measurement appropriate for the distribution. Otherwise, KR-20 formula will be based on a ratio of a discrete variance to a continuous variance. The resulting total variance will be notably to substantially inflated. For the current paper, the KR-20 will be a function of a total variance based on the discrete level of measurement. This variance will equal the summation of the main and off diagonals of a variance-covariance matrix. These calculations are further detailed in the next section.

KR-20 will be computed using the formula $\frac{N}{N-1}\left[1-\frac{1}{\sigma_T^2}\sum_{i=1}^{k}\frac{Y_i}{n_i}\left(1-\frac{Y_i}{n_i}\right)\right]$, where k denotes the total number of groups or categories, Y_i denotes the number of agreements between coders for the ith group or category, n_i is the total sample size for the ith group or category, and N is the total sample size across all groups or categories (Lord & Novick, 1968). The total variance $\left(\sigma_T^2\right)$ for coder agreement patterns equals the summation of elements in a variance-covariance matrix for binomial data (i.e., $\sigma_1^2+\sigma_2^2+2COV(X_1,X_2)$ = $\sigma_1^2+\sigma_2^2+2\rho_{12}\sigma_1\sigma_2$) (Stapleton, 1995). The variance-covariance matrix takes the general form

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{ij}\sigma_i\sigma_j \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \vdots & \cdots \\ \cdots & \cdots & \ddots & \vdots \\ \rho_{ij}\sigma_i\sigma_j & \cdots & \cdots & \sigma_n^2 \end{bmatrix}$$

(Kim & Timm, 2007), and reduces to

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 \end{bmatrix}$$

for a coding scheme comprised of two raters. In this matrix, the variances $\left(\sigma_1^2,\sigma_2^2\right)$ of agreement for the ith group or category should be based on discrete expectations (Hogg, McKean, & Craig, 2004). The form of this variance equals the second moment minus the square of the first moment; that is, $E(X^2) - [E(X)]^2$ (Ross, 1997). For continuous data, $E\left(X^2\right)=\int_{-\infty}^{+\infty}x^2\cdot f(x)\partial x$ and $E(X)=\int_{-\infty}^{+\infty}x\cdot f(x)\partial x$ where $f(x)$ denotes the probability density function (pdf) for continuous data. For the normal pdf, for example, $E(X)=\mu$ and $E\left(X^2\right)-E(X)=\sigma^2$.

For discrete data, $\mathrm{E}(\mathrm{X}^2) = \sum_i x_i^2 \,\mathrm{Pr}\,ob(X = x_i)$ and $\mathrm{E}(\mathrm{X}) = \sum_i x_i \,\mathrm{Pr}\,ob(X = x_i)$, where Prob($X$ = x_i) denotes the pdf for discrete data (Hogg & Craig, 1995). For the binomial pdf, $\mathrm{E}(\mathrm{X}) = n \cdot p$ and $\mathrm{E}(\mathrm{X}^2) - E(X) = n \cdot p(1-p)$ (Efron & Tibshirani, 1993). If a discrete distribution cannot be assumed or is unknown, it is most appropriate to use the distribution-free expectation (Hettmansperger & McKean, 1998). Basic algebra is only needed to solve for $\mathrm{E}(\mathrm{X}^2)$ and $E(X)$. For this last scenario it is important to also note that if the underlying distribution is discrete, methods assuming continuity for calculating $\mathrm{E}(\mathrm{X}^2)$ and $E(X)$ should not be utilized because the standard error can be substantially inflated, and reducing the accuracy of statistical inference (Bartoszynski & Niewiadomska-Bugaj, 1996).

As with the calculation of $\mathrm{E}(\mathrm{X}^2)$ and $E(X)$ the distribution of data must also be considered in the calculation of correlations. Otherwise, standard errors will be inflated. For data that take the form as either the presence or absence of a theme, which clearly have a discrete distribution, the correlation should be based on distributions suitable for this type of data. In this paper, the correlation ρ_{12} for agreement patterns between the coders will be Kendall-τ_c (Bonett and Wright, 2000). This correlation can be readily estimated using the PROC CORR procedure in the statistical software package SAS.

Estimates for calculating the KR-20 based on coder agreement patterns for the professor and student interview groups are provided in Table 12. Letting x_2 = 2 for non-agreed responses, the variance is 0.816 and 0.023, respectively, for the professor and undergraduate student groups. The Kendall-τ_c correlation equals 0.881. Using these estimates, the covariance between the groups equals 0.121. The total variance then equals 10.081. The final component of the KR-20 formula is the proportion of agreement times one minus this proportion (i.e., p_i(1-p_i)) for each of the groups. This estimate for the professor and undergraduate student interview groups equals 0.204 and 0.006. The sum of these values is 0.210. The KR-20 reliability estimate thus equals $\frac{556}{556-1}\left[1-\frac{1}{1.081}0.210\right]$ = 0.807, which equals the reliability between professor and student interview responses on the theme of interest.

Table 12
Estimates from Professor and Student Interview Participants for Calculating the KR 20

Estimate	Professor Group	Student Group
Individual Variances	0.816	0.023
Kendall-τ_c Correlation	0.881	
Covariance	$(0.881)(0.816)^{1/2}(0.023)^{1/2} = 0.121$	
Total Variance	0.816 + 0.023 + 2*0.121 = 1.081	
$p_i(1\text{-}p_i)$	0.714(1-0.714) = 0.204	0.994(1-0.994) = 0.006
$\Sigma p_i(1\text{-}p_i)$	0.210	

Discussion

This paper presented four quantitative methods for gauging interrater reliability of qualitative findings following a binomial distribution (theme is present, theme is absent). The κ statistic is a measure of observed agreement beyond the expected agreement between two or more coders. The κ_W statistic has the same interpretation as the kappa statistic, but permits the differential weights of cell frequencies reflecting patterns of coder agreement. The ANOVA (binary) ICC measures the degree to which two or more ratings are consistent. The KR-20 statistic is a reliability estimator based on the ratio of variances. That being said, it is important to note that the reliability of binomial coding patterns is invalid if based on continuous agreement statistics (Maclure & Willett, 1987).

Some researchers have developed tools for interpreting reliability coefficients, but do not provide guidelines for determining the sufficiency of such statistics. According to Landis and Koch (1977), coefficients of 0.41-0.60, 0.61-0.80, and 0.81-1.00 have 'Moderate,' 'Substantial,' and 'Almost Perfect' agreement, in that order. George and Mallery (2003) indicate that reliability coefficients of 0.9-1.0 are "Excellent," of 0.8-0.9 are "Good," of 0.7-0.8] are "Acceptable," of 0.6-0.7 are "Questionable," of 0.5-0.6] are "Poor," and less than 0.5 are "Unacceptable," where coefficients of at least 0.8 should be a researcher's target.

According this tool, the obtained κ of 0.891 demonstrates 'Almost Perfect' to 'Good' agreement between the coders. The κ_W statistic of 0.423 demonstrates 'Fair' to 'Unacceptable' agreement between the coders. The obtained ANOVA ICC of 0.714 demonstrates 'Substantial' to 'Acceptable' agreement between the coders. Last, the obtained KR-20 of 0.807 demonstrates 'Substantial' to 'Good' agreement between the coders.

The resulting question from these findings is "Are these reliability estimates sufficient?" The answer is dependent upon on the focus of the study, the complexity of the theme(s) under investigation, and the comfort level of the researcher. The more complicated the topic being investigated, the lower the proportion of observed agreement between the coders may be. According to Nunnally (1978), Cascio (1991), and Schmitt (1996), reliabilities of at least 0.70 are typically sufficient for use. The κ statistic, ANOVA ICC, and KR-20 meet this cutoff, demonstrating acceptable reliability coefficients.

What happens if the researcher has an acceptable level of reliability in mind, but does not meet the requirement? What methods should be employed in this situation? If a desired reliability coefficient is not achieved, it is recommended that the coders revisit their coding decisions on patterns of disagreement on the presence of themes in the binomial data (e.g., interview responses). After the coders revisit their coding decisions, the reliability coefficient would be re-estimated. This process would be recursive until a desired reliability coefficient is achieved. Although this process may seem tedious, the confidence in which the coders identified themes increases and thus improves the interpretability of the data.

Future Research

Three areas of research are recommended for furthering the use of reliability estimators for discrete coding patterns of binomial responses (e.g., qualitative interview data). In the current paper estimators that can be used to gauge agreement pattern reliability within a theme were presented. The development of quality reliability estimators applicable across themes should be further developed and investigated. This would allow researchers to determine the reliability of one's grounded theory, for example, as opposed to a component of the theory.

Sample size estimation methods also should be further developed for reliability estimators, but are presently limited to the κ statistic (Bonett, 2002; Feldt & Ankenmann, 1998). Sample size estimation would inform the researcher, in the example of the current paper, as to how many interviews should be conducted in order to achieve a desired reliability coefficient on their coded qualitative interview data with a certain likelihood prior to the initiation of data collection.

The current study simulated coder agreement data that follow a binomial probability density function. Further investigation should be conducted to determine if there are more appropriate discrete distributions to model agreement data. Possible densities may include the geometric, negative binomial, beta-binomial, and Poisson, for example. This development could lead to better estimators of reliability coefficients (e.g., for the investigation of 'rare' events).

References

Armstrong, D., Gosling, A., Weinman, J., & Marteau, T. (1997). The place of inter-rater reliability in qualitative research: An empirical study. *Sociology, 31*(3), 597-606.

Bartoszynski, R., & Niewiadomska-Bugaj, M. (1996). *Probability and statistical inference*. New York, NY: John Wiley.

Benaquisto, L. (2008). Axial coding. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (Vol. 1, pp. 51-52). Thousand Oaks, CA: SAGE.

Benaquisto, L. (2008). Coding frame. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (pp. 88-89). Thousand Oaks, CA: Sage.

Benaquisto, L. (2008). Open coding. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (Vol. 2, pp. 581-582). Thousand Oaks, CA: Sage.

Benaquisto, L. (2008). Selective coding. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods*. Thousand Oaks, CA: Sage.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*, 335-340.

Bonett, D. G. & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall, and Spearman correlations. *Psychometrika*, *65*, 23-28.

Brodsky, A. E. (2008). Researcher as instrument. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (Vol. 2, p. 766). Thousand Oaks, CA: Sage.

Burla, L., Knierim, B., Barth, J., Liewald, K., Duetz, M., & Abel, T. (2008). From text to codings: Intercoder reliability assessment in qualitative content analysis. *Nursing Research*, *57*, 113-117.

Cascio, W. F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall International.

Cheek, J. (2008). Funding. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (Vol. 1, pp. 360-363). Thousand Oaks, CA: Sage.

Cohen, J. (1960). A coefficient of agreement from 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.

Coryn, C. L. S. (2007). The holy trinity of methodological rigor: A skeptical view. *Journal of MultiDisciplinary Evaluation*, *4*(7), 26-31.

Creswell, J. W. (2007). *Qualitative inquiry & research design: Choosing among five approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Fort Worth, TX: Holt, Rinehart, & Winston.

Davis, C. S. (2008). Hypothesis. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (Vol. 1, pp. 408-409). Thousand Oaks, CA: SageAGE.

Dillon, W. R., & Mulani, N. (1984). A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioral Research*, *19*, 438-458.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall/CRC.

Elston, R. C., Hill, W. G., & Smith, C. (1977). Query: Estimating "Heritability" of a dichotomous trait. *Biometrics*, *33*, 231-236.

Everitt, B. S. (1968). Moments of the statistics kappa and weighted kappa. *The British Journal of Mathematical and Statistical Psychology*, *21*, 97-103.

Feldt, L. S. & Ankenmann, R. D. (1998). Appropriate sample size for comparison alpha reliabilities. *Applied Psychological Measurement*, *22*, 170-178.

Firmin, M. W. (2008). Replication. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (Vol. 2, pp. 754-755). Thousand Oaks, CA: Sage.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378-382.

Fleiss, J. L., Cohen, J., & Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, *72*, 323-327.

Fleiss, J. L., & Cuzick, J. (1979). The reliability of dichotomous judgments:

Unequal numbers of judges per subject. *Applied Psychological Measurement*, *3*, 537-542.

George, D. & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston, MA: Allyn & Bacon.

Given, L. M., & Saumure, K. (2008). Trustworthiness. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (Vol. 2, pp. 895-896). Thousand Oaks, CA: Sage.

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report, 8*(4), 597-607.

Greene, J. C. (2007). *Mixed methods in social inquiry*. Thousand Oaks, CA: Sage.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hettmansperger, T. P. & McKean, J. (1998). *Kendalls library of statistics 5, robust nonparametric statistical models*. London: Arnold.

Hogg, R. V. & Craig, A. T. (1995). *Introduction to mathematical statistics* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Hogg, R. V., McKean, J. W., & Craig, A. T. (2004). *Introduction to mathematical statistics* (6th ed.). Upper Saddle Rover, NJ: Prentice Hall.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Boston, MA: Allyn and Bacon.

Jensen, D. (2008). Confirmability. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (Vol. 1, p. 112). Thousand Oaks, CA: Sage.

Jensen, D. (2008). Credibility. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (Vol. 1, pp. 138-139). Thousand Oaks, CA: Sage.

Jensen, D. (2008). Dependability. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (Vol. 1, pp. 208-209). Thousand Oaks, CA: Sage.

Karlin, S., Cameron, P. E., & Williams, P. (1981). *Sibling and parent-offspring correlation with variable family age*. Proceedings of the National Academy of Science, U.S.A. 78, 2664-2668.

Kim, K. & Timm, N. (2007). *Univariate and multivariate general linear models: Theory and applications with SAS* (2nd ed.). New York, NY: Chapman & Hall/CRC.

Kleinman, J. C. (1973). Proportions with extraneous variance: Single and independent samples. *Journal of the American Statistical Association*, *68*, 46-54.

Krippendorf, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika*, *2*, 151-160.

Landis, J. R., & Koch, G. C. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.

Lipsitz, S. R., Laird, N. M., & Brennan, T. A. (1994). Simple moment estimates of the κ-coefficient and its variance. *Applied Statistics*, *43*, 309-323.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Maclure, M. & Willett, W. C. (1987). Misinterpretation and misuse of the kappa statistic. *Journal of Epidemiology*, *126*, 161-169.

Magee, B. (1985). *Popper*. London: Routledge Falmer.

Mak, T. K. (1988). Analyzing intraclass correlation for dichotomous variables. *Applied Statistics*, *37*, 344-252.

Marshall, C., & Rossman, G. B. (2006). *Designing qualitative research* (4th ed.). Thousand Oaks, CA: Sage.

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

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage.

Miller, P. (2008). Reliability. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (Vol. 2, pp. 753-754). Thousand Oaks, CA: Sage.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin*, *86*, 376-390.

Morse, J. M., Barrett, M., Mayan, M., Olson, K., & Spiers, J. (2002). Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods, 1*(2), 13-22.

Nelder, J. A., & Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, *74*, 221-232.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Paley, J. (2008). Positivism. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (Vol. 2, pp. 646-650). Thousand Oaks, CA: Sage.

Ridout, M. S., Demétrio, C. G. B., & Firth, D. (1999). Estimating intraclass correlations for binary data. *Biometrics*, *55*, 137-148.

Ross, S. (1997). *A first course in probability* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Rozzeboom, W. W. (1966). *Foundations of the theory of prediction.* Homewood, IL: Dorsey.

Saumure, K., & Given, L. M. (2008). Rigor in qualitative research. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (Vol. 2, pp. 795-796). Thousand Oaks, CA: Sage.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 81-84.

Seale, C. (1999). Quality in qualitative research. *Qualitative Inquiry, 5*(4), 465-478.

Smith, D. M. (1983). Algorithm AS189: Maximum likelihood estimation of the parameters of the beta binomial distribution. *Applied Statistics*, *32*, 196-204.

Soeken, K. L., & Prescott, P. A. (1986). Issues in the use of kappa to estimate reliability. *Medical Care*, *24*, 733-741.

Stapleton, J. H. (1995). *Linear statistical models*. New York, NY: John Wiley & Sons, Inc.

Stenbacka, C. (2001). Qualitative research requires quality concepts of its own. *Management Decision, 39*(7), 551-555.

Tamura, R. N., & Young, S. S. (1987). A stabilized moment estimator for the beta-binomial distribution. *Biometrics*, *43*, 813-824.

van den Hoonaard, W. C. (2008). Inter- and intracoder reliability. In L. M. Given (Ed.), *The Sage encyclopedia of qualitative research methods* (Vol. 1, pp. 445-446). Thousand Oaks, CA: Sage.

Yamamoto, E., & Yanagimoto, T. (1992). Moment estimators for the binomial distribution. *Journal of Applied Statistics*, *19*, 273-283.