Sex Roles (2011) 64:276–289
DOI 10.1007/s11199-010-9893-0

ORIGINAL ARTICLE

Content Analysis—A Methodological Primer for Gender Research

Kimberly A. Neuendorf

Published online: 20 November 2010

Abstract This article is intended to serve as a primer on methodological standards for gender scholars pursuing content analytic research. The scientific underpinnings of the method are explored, including the roles of theory, past research, population definition, objectivity/intersubjectivity, reliability, validity, generalizability, and replicability. Both human coding and computer coding are considered. The typical process of human-coded content analysis is reviewed, including the steps of unitizing, sampling, measurement, coder training, reliability assessment, and reportage of methods. Numerous applications to research on gender roles and related issues are reviewed. Practical checklists are offered for content analysis preparation and methodological execution.

Keywords Content analysis · Methodological standards · Gender research · Gender roles

Introduction

As an increasingly popular research methodology, quantitative content analysis presents gender researchers with a set of useful tools for comparing messages generated by males and females (e.g., Argamon et al. 2003; Fields et al. 2010), and for studying messages containing information about sex and gender roles. Indeed, perhaps no substantive area has been more thoroughly content analyzed across all media than that of the roles of males and females (Neuendorf 2002). Studies have compared male and female behaviors and attributes for domestic and international content in media ranging from film and television (e.g., Fernandez-Villanueva et al. 2009; Smith 1999) to children's books (Anderson and Hamilton 2005), men's magazines (Ricciardelli et al. 2010), video games (e.g., Martins et al. 2009), radio talk shows (Brinson and Winn 1997), and even postage stamps (Ogletree et al. 1994) and birth congratulatory cards (Bridges 1993). Research on gender images in advertising seems to be particularly popular in recent years (An and Kim 2007; Baker 2005; Fullerton and Kendrick 2000; Ibroscheva 2007; Lindner 2004; Messineo 2008; Odekerken-Schroder et al. 2002; Uray and Burnaz 2003; Valls-Fernandez and Martinez-Vicente 2007).

However, rigorous methodological standards have not always been evident, notably with regard to issues of validity and reliability (Lombard et al. 2002; Neuendorf 2009; Pasadeos et al. 1995), and content analysis has often suffered by comparison with other empirical methodologies. Even contemporary reviews of content analyses find such salient standards as reliability assessment to be lacking in a majority of published studies. For example, a recent systematic analysis of 133 health media content analyses (Neuendorf 2009) found not a single instance of full reliability assessment and reportage (written documentation), with 38% including no reliability assessment whatsoever. This figure is comparable to the 31% found by Lombard et al. (2002) in their review of content analysis in the field of communication. Neuendorf (2009) cited *Sex Roles* as one of a half-dozen journals with better-than-average reliability assessment reportage in articles on health-related content analyses. Nevertheless, in an exploratory review of all 72 quantitative content analyses appearing in *Sex Roles* from 1997 through mid-2010 conducted for the current article, there were still articles

K. A. Neuendorf (✉)
School of Communication, Cleveland State University,
Cleveland, OH 44115, USA
e-mail: k.neuendorf@comcast.net

found that demonstrated instances of poor sampling, unacceptable reliability assessment, and inadequate methodological reportage. In an effort to bolster the methodological rigor of content analysis, this article is intended to serve as a primer on standards for gender scholars pursuing content analytic research.

Assumptions

First, a definition of content analysis is in order, to establish a common understanding of methodological assumptions. Definitions range from Babbie's (2010, p. G2) broadly phrased ”the study of recorded human communications” to the more limiting definition adopted here: Content analysis is a summarizing, quantitative analysis of messages that relies on the scientific method, including attention to objectivity/intersubjectivity, a priori design, reliability, validity, generalizability, replicability, and hypothesis testing. It is not limited as to the types of messages that may be analyzed, nor as to the types of variables that might be measured (Neuendorf 2002, p. 10).

Thus, this definition assumes a quantitative approach. Other valid means of analyzing messages include discourse analysis (Hardy et al. 2004), rhetorical analysis (McCroskey 1993), semiotics (Eco 1976), phenomenological analysis (Johnston and Morrison 2007) and narratology (Lieberman et al. 2009). These methods are often empirical (i.e., based on observations). They may fruitfully serve as complements to content analysis (Neuendorf 2004). While scholarship focusing on message analysis should be committed to the use of a variety of methodologies, this piece will focus on the particular needs of the quantitative content analysis researcher.

Further, this piece will for the most part assume that the researcher plans a content analysis in which measurement is conducted via human coders, the most frequently utilized content analysis technique. Indeed, an inspection of the 72 quantitative content analyses and text analyses published in *Sex Roles* from 1997 through mid-2010 revealed that well over 90% relied solely on human-coding techniques. The alternative, using computer applications to analyze text (called CATA, Computer-Aided Text Analysis), will be addressed here only occasionally (see Gottschalk and Bechtel 2008; Neuendorf 2002; Roberts 1997; West 2001; for a list of CATA programs, see *The Content Analysis Guidebook Online* (Neuendorf and Kane 2010) at http://academic.csuohio.edu/kneuendorf/content).

As noted, within the realm of quantitative content analysis, rigorous standards have not always been met. While the norms vary by journal (Neuendorf 2009), and, fortunately, the quality bar seems to be rising over time, a gap still exists between the rigor required for a content analysis publication and that required for other quantitative methods such as survey or experimental techniques. A number of scholars with wide experience in content analysis and other quantitative research approaches have attempted to close this gap (e.g., Lombard et al. 2002). The first recommendation offered, then, is that all content analysis investigations should be guided by accepted reference texts on methodology, particularly those that take a comprehensive approach to the method and optimally are informed by a wide array of methodological and applied research experiences not limited to a single discipline (e.g., Krippendorff 2004; Neuendorf 2002; Riffe et al. 2005; Weber 1990).

Preparation

As with any systematic empirical investigation, a content analysis should proceed only after adequate planning and preparation. The following six points summarize the major pre-analysis decisions faced by the content analyst. These points also appear in Table 1.

1. *Theoretical and conceptual backing*. Each content analysis must be guided by a theoretical framework. Research in the arena of sex/gender roles seems to exhibit greater commitment to theoretical grounding than does research in many other areas of content analysis, possibly due to gender research's direct derivation from theories of feminism, Marxism, gender role effects, stereotyping, sexism, body image impacts, and the biological bases of sex differences (e.g., Eschholz et al. 2002; Evans and Davies 2000; Harrison and Hefner 2006; Schlenker et al. 1998; Smith 1999).

Some analyses explicitly test hypotheses derived from theory, as in a pair of studies that applied Hofstede's (1980) cultural dimension of masculinity to compare advertising content between nations high and low in cultural-level masculinity (An and Kim 2007; Odekerken-Schroder et al. 2002). Here, the independent variable was a nation's level of masculinity, defined by Hofstede as the degree of gender role differentiation, and dependent variables included a range of gender role portrayals (e.g., female dress shown as demure or seductive; working vs. nonworking role; type of nonworking role (family/recreational/decorative) portrayed by women). The differences predicted by the theory were largely confirmed in An and Kim's comparison of U.S. (low masculinity) and Korean (high masculinity) web ads, but they were not confirmed by Odekerken-Schroder, De Wulf, and Hofstee's study of print advertising in the U.K. and the Netherlands (high and low masculinity, respectively).

On the other hand, most content analyses utilize theory primarily as an underlying rationale for the study of messages (e.g., Ibroscheva 2007; Miller and Summers 2007; Valls-Fernandez and Martinez-Vicente 2007). This often takes the form of an application of theories of message effects. For example, Eschholz et al. (2002) invoked theories of cognitive effects of exposure to gender role portrayals, Dietz (1998) referenced Gerbner's cultivation theory of perceptual effects of media images, and Milburn et al. (2001) cited past research to back the notion that gender stereotyping can affect self-concept, the evaluation of others, and task performance. Other important media effects theories that have served well as bases for content analyses include social cognitive theory (e.g., Cressman et al. 2009), agenda setting and priming (e.g., Balmas and Scheafer 2010), framing (e.g., Pan et al. 2010), cultivation (e.g., Cressman et al. 2009; Martins et al. 2009), and uses and gratifications (e.g., Ebersole 2000). In sum, theory and past research on message effects may serve as the logical basis for content analyses of suspected influential content.

2. *A plan for the scope of the investigation.* In its most basic form, a content analysis may simply be (a) descriptive of message content. However, the scope may be productively expanded by: (b) examining relationships among message variables, (c) combining message data with data about the message source, and/or (d) combining message data with data about the message receiver. These last two approaches to content analysis may be termed "integrative" (Neuendorf 2002), and offer a powerful means of determining the antecedents of message creation (Shoemaker and Reese 1996) and the effects of message reception.

The first type of expansion, that of looking at relationships among the message variables, seems to be a common choice among gender role content analysts. At a minimum, studies often make statistical comparisons between male and female characters or performers in mass media content. For example, Miller and Summers (2007, p. 733) found significant differences between male and female video game characters as presented in video game magazines: Males were more likely to be main characters or heroes, have more abilities, use more weapons, and be more powerful and muscular; females were more likely to be supplementary characters, wear revealing clothes, and be portrayed as attractive, sexy, and innocent. In a second example, Evans and Davies (2000) found males in elementary school reading textbooks to be significantly more aggressive, argumentative, and competitive than females. An example of a more complex statistical analysis of the relationships among content analysis variables is Neuendorf et al.'s (2010) significant prediction of mortality among females in James Bond films using logistic regression with 17 predictor variables.

Second, the combining of message source data with content analysis message data may allow the discovery of factors important to the process of message generation. An interesting example of this type of study is Lauzen et al.'s (2006) investigation of how the involvement of women behind the scenes in the production of reality and scripted prime-time U.S. television programming relates to female representations and portrayals. The presence of women in top creative positions for scripted sitcoms and dramas predicted greater female character representation, and a more egalitarian approach to conflict resolution; these relationships did not emerge for reality programming.

The third option, integrating content analysis message data with message receiver data, affords an opportunity to test message effects theories. For example, Collins et al. (2009) combined content analysis and survey data to develop detailed measures of teens' exposure to specific sexual content on TV, which were then found to predict over time whether the teens initiated sexual intercourse. Weber et al. (2009) examined teenage boys' physiological responses (i.e., heart rate, skin conductance) as outcomes of the content of their first-person-shooter game play. And in a novel linking of content analysis and experimental findings, Franiuk et al. (2008) studied the prevalence of rape myth endorsements in online newspaper headlines about the 2003-2004 Kobe Bryant case, and then conducted an experiment that found male subjects to hold higher rape-supportive attitudes after exposure to myth-endorsing headlines identified via this content analysis.

3. *Review of past research and development of measures.* In anticipation of the development of a content analysis protocol—including measures that constitute a coding scheme or a set of CATA dictionaries—the researcher should conduct an exhaustive search of previous work on the topic.

Scholars may profit by trying to build upon past research in extending the findings of earlier studies to different media, locations, or content forms, or by studying content changes over time. Current findings are occasionally compared to the findings of past studies, but where careful replication of methods (e.g., measures, sampling technique) has not been employed, the comparison is not as meaningful as it might be. Thus, a careful review of past work might provide the key to a more complete, longitudinal research program.

The core of any human coded content analysis is the coding scheme—a combination of codebook and coding form. The codebook contains fully explicated operationalizations for all variables—i.e., "rules" for the coders. The codebook itself looks like a somewhat adapted questionnaire, with carefully explicated definitions of each variable and of all measurement options or categories within each variable.

Unfortunately, no collection of standard codebooks exists (however, see Neuendorf and Kane 2010, for examples). Some content analyses have built on earlier coding schemes (e.g., Ibroscheva 2007), sometimes serving as a replication or extension to another body of content (Schlenker et al. 1998). For example, Domhoff's (1999) study provided further evaluation of the elaborate Hall and Van de Castle coding scheme for the study of dream content, first developed in the 1940s and revised in the 1960s.

The adaptation of measures from other types of research may be considered. For example, Evans and Davies (2000) used the seminal work of Bem (1981) on self-report indicators of femininity and masculinity to extract 16 traits as content analysis measures for their study of elementary school texts (e.g., aggressiveness, argumentativeness, affection, passivity). Eschholz et al. (2002) also adapted elements from the Bem Sex Role Inventory, as well as the Personal Attributes Questionnaire (Spence et al. 1974) to study gender and racial/ethnic roles in contemporary American film. And Lindner (2004), in her analysis of women's images in magazine ads, adapted a set of qualitative criteria from Goffman's classic work on the subtle cues contained within advertising images (Goffman 1979). Uray and Burnaz (2003) provide a veritable model of comprehensive reportage in this regard, presenting fully three tables listing all 22 of their content analysis variables' operational definitions, with scholarly sources for each variable.

When using CATA, decisions must be made regarding how to establish dictionaries (i.e., sets of search terms applied by the computer application to written texts). More than a dozen quantitative CATA programs are available, and most include some pre-set dictionaries (see Neuendorf and Kane 2010). In LIWC (Linguistic Inquiry and Word Count; Pennebaker et al. 2007) there are 84 dictionaries that include such linguistic and semantic concepts as use of first-person pronouns, anger, optimism, reference to home, and reference to motion.

The program Diction 5.0 (Hart 2000), designed to analyze political speech, has 31 pre-set dictionaries, including those intended to measure tenacity, aggression, praise, satisfaction, and complexity. The 31 dictionaries are also combined to form "master variable" scales: Activity, optimism, certainty, realism, and commonality.

The alternative to using pre-set dictionaries is to create one's own custom dictionaries, and most CATA programs allow for this. However, the development of such original dictionaries is demanding both conceptually and logistically. Both content validity and internal consistency reliability are of concern, and custom dictionary development should include a construct validation process that links measured dictionaries with additional indicators of the concepts under investigation.

4. *Defining the population of messages to be analyzed.* The population is the realm of inquiry for an investigation—the set of units (in content analysis, usually messages or message components) to which researchers wish to generalize their findings. The decision as to what messages will constitute the population originates with theory but must be tempered with practical considerations.

The content analyst may choose to take one of two main approaches to determining the population of messages to be studied—an availability-based approach or an exposure-based approach. An availability-based procedure defines the population as the set of messages available to receivers in a given medium at a given time. For example, the population of television content may be defined as all programs appearing on a set of broadcast and cable networks during a specified time period. Kunkel et al. (1995) utilized this approach—which they called a "what's on" method–in constructing composite week samples for their National Television Violence Study. An exposure-based approach defines the population as messages widely attended to by receivers. For example, a television program population may consist of the top rated TV/cable programs (e.g., Fernandez-Villanueva et al. 2009). Xue and Ellzey (2009) chose to study ads in the three top-selling women's and men's magazines, as determined by single-issue sales data provided by the Magazine Publishers of America. Clearly, an availability-based approach is appropriate when content analysts are applying theories of message production, while an exposure-based approach is consistent with theories of message effects. The researchers' theoretical framework may fruitfully guide such decisions of population definition.

The population defined by the researcher may be quite narrow. In such cases, the content analysis may actually constitute a full *census* study of all elements in the population. For example, Neuendorf et al. (2010) were interested in documenting the portrayals of women in James Bond films. The analysis encompassed all 195 major and featured female characters in the 20 Bond films released through 2005. Similarly, Eschholz et al.'s (2002) decision to define their population as the top 50 grossing U.S. films for the year of their study (1996) allowed them to execute a full census of a limited but clearly defined population of messages.

A sampling frame, i.e., a list of the elements in the defined population, does not exist for every population. For example, An and Kim (2007) acknowledged the lack of a perfect sampling frame for their study of web advertising in the U.S. and Korea. Therefore, they chose to sample from lists of top brands prepared by *BusinessWeek* (for the U.S. sample) and the Korean Culture and Information (KCI) database (for the Korean sample). In essence, they refocused their defined *population* as the web sites corresponding to these two "credible" lists. Appropriately, they executed a systematic sample with random start from each list. Another interesting decision in defining a message population is the Milburn et al. (2001) study of clipart—every image containing a humanoid figure (including cartoons and silhouettes) included in the two software packages Microsoft Office 97 and Print Shop Ensemble III was coded, resulting in a census of 3,929 characters in 2,713 pieces of clipart.

Markson and Taylor (2000) first had in mind as their population all U.S. feature films including an actor or actress 60 years or older. Lacking any type of sampling frame to match this, they revised their vision to include all films done after age 60 by every actor and actress nominated for an Academy Award at some time in their life. This resulted in a more limited scope to the research, but one which was fully disclosed and consistently recognized in the report of findings.

Technology changes have made the definition of the population more problematic in some cases. First, multiple delivery systems are often available for a given medium or content type. For example, music videos may be viewed on broadcast TV, via cable, on video or DVD, or online, and this variety of delivery modes makes the definition of the population to which one wishes to generalize a more complicated task. Second, the fluid nature of some communication content may make the population definition problematic. This fluidity may arise from changing content, such as web sites that are frequently updated (McMillan 2000; Weare and Lin 2000) and evolving social networking content (Patchin and Hinduja 2010). Researchers have typically addressed this issue via repeated samplings.

On the other hand, fluidity may also stem from user operation, which may be addressed by taking a sample of users, rather than units of content. This is followed by the recording and analysis of the messages that are received–or created—by these users. For example, some studies of online content have focused on what web users actually attend to, unobtrusively recording their web activity (e.g., Danaher et al. 2006; Jansen and Spink 2006; Mastro et al. 2002). Further, content analyses of video games have adopted the practice of using a set of recorded gaming sessions as the content, rather than a hypothetical population of all potential content for a given game (e.g., Haninger and Thompson 2004; Martins et al. 2009; Weber et al. 2009).

Given the wealth of ways in which researchers might define a population of messages, two important guidelines are evident: (a) Researchers should attempt to establish a population that is consistent with their study's theoretical perspective, and (b) researchers should fully report the nature of their population, so as to clarify their focus and divulge any limitations on generalizability.

5. *Immersion in the message pool.* In addition to reviewing research literature on the topic of interest, the content analyst should also take a practical approach and seek additional clues from a thorough examination of the pool of messages constituting the defined population. This immersion will typically result in the emergence of key variables that might otherwise not have been detected. For example, Knobloch (2008) examined subjects' open-ended responses to a self-administered questionnaire about relational uncertainty in marriage to inductively derive 12 emergent dimensions for further study—some expected, such as children, careers, and finances, and others less predictable, such as retirement, the extended family, and household chores.

6. *Decision on whether to use human coding and/or computer coding (CATA).* For content that is entirely verbal (written or transcribed), researchers have the option of using CATA. A wide variety of programs now provide pre-set dictionaries intended to measure such constructs as optimism, aggression, and emotional tone (Neuendorf and Skalski 2010), and most allow the addition of custom dictionaries by the researcher. Even if a dictionary-based analysis is not desired, CATA programs can operate as simple search tools, assuring that no occurrence of a term such as "smoking" or "cigarettes" is missed. This can reduce coding and recording errors that can depress reliability.

The majority of the examples cited in this article are human coded content analyses. To gain an idea of how CATA might be useful in studies of gender roles, we might examine two applications of the CATA program, LIWC (Pennebaker et al. 2007), to comparisons of texts generated by males and females.

Groom and Pennebaker (2005) applied 74 language dimensions of the LIWC pre-set dictionaries, and one custom dictionary of their own devising (use of third-person singular pronouns) to 1500 internet personal ads. Numerous gender differences were found–e.g., males exhibited higher scores on job-related text, while females showed higher scores on positive emotions, positive feelings, sexuality, and the use of present-tense verbs. And, differences between heterosexual and homosexual sources were also found–e.g., heterosexuals obtained higher scores on overall word count, use of pronouns, and achievement-related texts, while homosexuals showed higher scores on body states, sexuality, and the sensation of sight.

Schmader et al. (2007) used seven of the standard LIWC dictionaries and five more of their own creation to analyze 886 letters of recommendation written on behalf of male and female applicants for either a chemistry or biochemistry faculty position at a U.S. research university. The pre-set LIWC dictionaries used were: Word count, achievement, communication, positive feelings, tentative words, and certainty words. The custom dictionaries were: Standout words (e.g., superb, outstanding, supreme), ability words (e.g., talent*, brilliant*, competent), grindstone words (e.g., conscientious, reliab*, methodical), teaching words (e.g., mentor, colleague, lectur*), and research words (e.g., study, scholarship, publish*). Results revealed more similarities than differences between letters written for men and women candidates; among the differences–letters written for men used significantly more "standout" words.

CATA procedures offer speed, standardization, and guaranteed reliability. However, they typically rather blindly count dictionaried words, without full accounting of ambiguity, negation, or other contextual factors, making the validity of the measures a critical question. Further, pre-set CATA dictionaries may not match the needs of the researcher, and the development of custom dictionaries is demanding. Compared to CATA, human coding allows much more in-depth, nuanced measures, but of course is held to the high standard of intercoder reliability and is time-consuming.

Methodology Concerns

The practice of content analysis research should be approached with attention to detail and rigor. Key areas of methodological concern, summarized in Table 2, are examined below.

1. *Unitizing—decisions, training, and another stage for reliability assessment.* Often, a major challenge is the identification of clearly defined message units to which the measures will be applied. There may be a definitive set of rules for the identification of units, as in Uray and Burnaz's (2003, p. 80) study of characters in Turkish television commercials: "Adult male and female characters that appeared on camera either speaking or having prominent exposure for at least 3s formed the database of this study. . . a maximum of two characters were accepted as being primary figures for each advertisement. In cases where there were more than two characters, the two most dominant characters were selected as primary characters."

Often, coders may be unitizing "live" as they code, as in the case of coding each instance of cigarette smoking as it occurs in a feature film (Dozier et al. 2005). Whenever researchers or coders are required to identify message units, a separate layer of reliability assessment is in order—the reliability of unitizing. Unfortunately, a standard has not been set for statistical assessment of unitizing reliability. One that has been proposed, Guetzkow's agreement statistic, *U*, (Guetzkow 1950) assesses only the comparative number of units that coders identified, not whether the actual units were the same. More recent calls for unitizing reliability have noted the need to identify specific units in common, but do not provide a statistical test beyond simple percentage agreement on these common units (Cissna et al. 1990; Garvin et al. 1988). Krippendorff (2004) does provide an adaptation of his alpha coefficient for this task, but only for instances where unitizing involves the segmentation of a time continuum.

In general, unitizing should be done in such a concrete fashion that coders do not have to make decisions during the coding process. Reliability is compromised whenever coders have difficulty in identifying units. For example, coding each discrete instance of smoking in a linear narrative will surely be a less reliable process than coding each character's smoking behavior overall (e.g., whether the character smoked or not), or coding whether smoking occurred in each five-minute interval.

Often, multiple units of analysis are employed in a content analysis project. For example, an analysis of feature films involved coding (a) at the whole-film level, (b) with each lead, major, or medium character as the unit, and (c) production techniques and motifs measured with a five-minute time interval as the unit of data collection (Janstova et al. 2010). This is in essence *three* different content analyses, with three separate coding schemes.

2. *Sampling.* When the researcher needs to select a subset of units from the population, s/he has two options: Probability and nonprobability sampling. Probability sampling, intended to provide a representative subset, is essential if generalization to the larger population of messages is desired. The two main choices for probability sampling are simple random sampling and systematic sampling with a random start, which involves taking every k-th element from a sampling frame. But as noted above, a valid sampling frame that lists the entire population is not always available, and the use of such nonprobability sampling techniques as convenience, purposive, or quota might be necessary. Further, the size of the sample should be established with accepted statistical practices (see Riffe et al. 2005).

The particular medium in which the messages are carried will clearly affect the sampling process (e.g., availability of sampling frame, units of sampling) as it does the population definition. For example, for content analyses of web sites, it is typical that a "snapshot" approach is used for collecting the sample (Norris 2003). Curtin and Gaither (2003) downloaded entire web sites, collecting their content twice, 1 month apart, in order to capture the "dynamic nature of the web" (p. 12). This freezing of the content is essential to reliability.

3. *Measurement–Codebook construction and dictionary definition*. The operationalization of concepts derived from theory, past research, and immersion in the message pool results in a coding scheme (i.e., a codebook and coding form) or in a set of dictionaries (for CATA). In general, the measurement of content analytic variables should be viewed as not substantially different from other quantitative measurement approaches. Measures should be evaluated in much the same way as measures in surveys and experiments–each variable needs to have options that are exhaustive and mutually exclusive, and should be measured at the highest possible level of measurement. Attention should be paid to individual variables' variances and distributions, and statistical transformations of the variables should be executed as needed. Measures with good characteristics (e.g., with a reasonable amount of variance, and a distribution that is normal) are more likely to result in reliable and valid outcomes. Variables may be combined into scales, for which internal consistency reliability may be assessed (e.g., with Cronbach's alpha).

This notion of multiple measures might relate to the distinction between latent and manifest content (Gray and Densten 1998). Broad, latent constructs such as assertiveness, nurturing, compassion, and submissiveness are common concerns in gender studies, and so the consideration of latent vs. manifest content is particularly appropriate to the field. Manifest content may be defined as elements that are present and directly identifiable, while latent content constitutes the deeper meaning, i.e., that which is not directly observable. Based on Freud's interpretation of dreams (Gregory and Zangwill 1987), the delineation of latent and manifest content is a rather contested approach within content analysis (Potter and Levine-Donnerstein 1999; Riffe et al. 2005; Shapiro and Markoff 1997). Further, some scholars propose that variables be situated on a continuum (Neuendorf 2002; Riffe et al. 2005) rather than placed in one category or the other. Regardless, what is important about the distinction between manifest and latent for gender studies content analysts is the question of how one might measure constructs that are by nature not directly observable.

The examination of clearly latent constructs in messages is more often the province of qualitative textual analyses. For example, Laird et al. (2007), examining MEDLINE-indexed abstracts containing reference to Muslims or Islam, utilized a two-researcher consultative technique in which the question "What does this text convey to the reader about Islam or Muslims?" was used for the identification of latent (implicit) themes. As Riffe, Lacy, and Fico note (2005, p. 126), measurement of latent constructs is subjective, relying "on coder interpretation of content meaning." However, with considerable codebook definition (preceded by substantial qualitative work; e.g., Clarke and Everest 2006) and in-depth training, quantitative content analysis may achieve direct measurement of latent constructs. Indeed, coders have reliably measured such subjective constructs as "defamation" (Simon et al. 1989) and journalistic framing (Jones and Himelboim 2010).

As with surveys and experiments, latent content in content analysis is often measured with multiple indicators of manifest characteristics that together indicate a latent state (e.g., Radwin and Cabral 2010), such as Ghose and Dou's (1998) 23 manifest indicators representing the latent construct "interactivity" for web sites, and Kinney's (2005) factor analytic extraction of four latent patterns from a set of 11 manifest CATA measures of word use in seven U.S. newspapers.

The validity of measures must be a strong consideration, with content analysis as with any empirical, quantitative research enterprise. Reliability is a necessary but not sufficient condition for the establishment of validity (Potter 2009); thus, additional criteria beyond intercoder reliability should be engaged. Unfortunately, formal assessment of the validity of content analysis measures is uncommon. Options include researcher inspection of the measures for basic face validity (Janis 1965) and for content validity, i.e., the extent to which one or multiple measures fully reflect a specified domain (Carmines and Zeller 1979). Further, construct validity examines whether a measure relates to other measures in ways that are consistent with theoretically derived hypotheses (McAdams and Zeldow 1993). A measure has high construct validity when it correlates with other measures of the same construct (convergent validity) and does not correlate with measures of dissimilar constructs (discriminant validity) (Weber 1990, p. 19). An example of convergent construct validation is Gottschalk and colleagues' long-term efforts to link CATA dictionary findings to the outcomes of clinical diagnostic procedures applied to the sources of the messages analyzed (Gottschalk and Bechtel 2008). The inherent difficulties in assessing the validity of measures applied to content that is often far removed from the source have long been recognized (Janis 1965; Potter and Levine-Donnerstein 1999). Nevertheless, some scholars have attempted this process, including George's (1959) ex post facto efforts to validate World War II propaganda content analyses with documents seized after the war (see also Krippendorff 2004).

One common question is whether codebook instructions should include examples in addition to the carefully worded concept explications (e.g., Miller and Summers 2007). While conventional wisdom finds this practice to increase reliability, emerging evidence indicates that the inclusion of examples may prove a threat to validity by materially changing the codebook—specifically, coders may be less likely to code the presence of an attribute/behavior/etc. when an example is included. The inclusion of an example seems to limit the coder's vision as to the application of the variable.

When devising a codebook, it is worth considering the particular medium in which the target messages are carried. There may be critical medium-specific (form) variables that moderate the presentation of content in that medium. A clear example is the study of MTV's portrayals of aggressive acts and cues that found that females were no more likely than were males to be the victims of aggression; however, when victims were females, they were significantly more likely to be shown in closeup, and for a longer period of time (Kalis and Neuendorf 1989). The critical form variables of shot type and shot length provided additional information about the presentation of the aggressive content that was essential to a full understanding of its potential audience reception.

4. *Training*. A unique characteristic of human-coded content analysis measures, one that distinguishes them from measures in other types of studies such as surveys and experiments, is their reliance on trained individuals as part of the coding protocol—that is, human coders are an integral part of the measurement device. Thus, nothing is more important to the valid and reliable measurement process than coder training. Training should involve both a full discussion of the coding scheme and a series of group coding sessions, during which the coding team members become calibrated to one another. During the training process, the codebook may undergo changes. Ultimately, reliability checks and final coding should be conducted independently by the trained coders. The minimum number of coders is two, to allow for a reliability test, but more may be employed as needed.

Generally, in order to assure replicability, the assumption is that nearly any individual may serve as a coder, and the selection of coders should not be based on some prior expert knowledge or skill. The particular skills necessary for the coding protocol should be developed during training, and should be fully reflected in the codebook, thus documenting all information needed to replicate the protocol by other coding teams at other times.

The issue of "blind" coding has gained some attention in the literature. It is proposed that coders be kept ignorant as to the true intent of the research so as to minimize coder bias (Kolbe and Burnett 1991; Pollock and Yulis 2004). Lindner (2004) employed blind coding in her content analysis of magazine ads, and Knobloch (2008) also employed "judges who were blind to the goals of the study" (p. 474).

5. *Reliability*. For human coding, reliability is essential. Optimally, at least two reliability subsamples will be selected for a given content analysis. One will serve as the content for a pilot reliability test before full coding commences; this pilot provides one last chance to change the coding scheme to maximize reliability. The second will provide material for the final reliability test, conducted throughout the process of full coding and reported with the study's findings. A number of options exist for the selection of reliability subsamples. The most common technique is to randomly select a subset of the main content analysis sample, usually about 10-20% of the full sample. Just as the full sample typically is representative of a larger defined population of interest, the reliability subsample is viewed as representative of the sample.

However, another option exists, similar to the choice of testing hypotheses in experiments by using only the extreme high and low groups. This second option is to select a reliability sample that maximizes the variance on key dimensions of interest (e.g., Potter et al. 1998), which might be thought of as a "rich range" sample (Neuendorf 2009). This option is particularly appealing in cases where many of the variables under examination are "rare event" measures, in which the targeted activity occurs in only a small proportion of the cases. The option calls for, in essence, oversampling for these rare events, (a) providing more opportunity for coders to become skilled at identifying these instances, and (b) producing variables that have greater variance within the reliability data set. Such "rich range" sampling is also well suited to the selection of examples of the content for *training* (e.g., Hubbell and Dearing 2003).

Intercoder reliability statistics may be categorized as indicators of (a) agreement, (b) chance-corrected agreement (agreement beyond chance), and (c) covariation. Generally, it is *not* acceptable to present only indicators of agreement with no correction for chance (i.e., percent agreement or Holsti's coefficient). Chance-corrected agreement is appropriate when a measured variable is categorical (i.e., nominal), while covariation is appropriate to a variable that is measured via a metric (i.e., interval/ratio level of measurement).

There is ongoing debate over the merits of the various intercoder reliability statistics currently available (e.g., Hayes and Krippendorff 2007; Krippendorff 2004; Lombard et al. 2002; Lombard et al. 2004; Neuendorf 2009; Potter and Levine-Donnerstein 1999; Zwick 1988). We may explore the different assumptions of an "intercoder reliability" approach vs. the "interrater reliability" approach more commonly found in clinical applications. The latter treats the raters more as experts, and acknowledges and allows for disagreements among them—indeed, their differences are sometimes valued and closely examined (Goodwin 2001). Additionally, the development of new reliability statistics might be considered. For example, problems with achieving an acceptable level of reliability with "rare event" variables have been noted (Janstova et al. 2010). Such problems follow from existing nominal-level coefficients' reliance on marginal probabilities that may be skewed or unbalanced, and correlational statistics' sensitivity to low variance and truncated range.

At present, it is recommended that researchers use some of the more widely accepted reliability coefficients; those statistics with a fuller "track record" provide us with greater bases for comparison with past work, and allow a more standard shared statistical language for discussion among scholars. Many frequently used statistics (e.g., Scott's pi, Cohen's kappa) are calculable via PRAM, a PC-based program available in a trial version (Skymeg Software 2009; see http://academic.csuohio.edu/kneuendorf/content/reliable/pram.htm). PRAM produces the following heuristics and intercoder reliability coefficients: Percent agreement, Scott's pi, Cohen's kappa, Fleiss' multi-coder version of Cohen's kappa (Fleiss 1981; Fleiss et al. 2003), Spearman rho, Pearson correlation, Lin's concordance coefficient (Lin 1989), and Krippendorff's alphas (the four alphas—one for each of the four levels of measurement, i.e., nominal, ordinal, interval, ratio–have not been fully validated in PRAM). PRAM appears to be the only program that handles multiple coders and multiple variables simultaneously. It uses an Excel spreadsheet database format that is compatible with SPSS.

Current recommendations for the selection of reliability statistics may be summarized as follows:

> *For nominal data*—Agreement controlling for chance is the contemporary required standard. While simple percent agreement or Holsti's coefficient (based on percent agreement) may serve as an heuristic for researchers, these statistics do not take into account the impact of chance agreement and are therefore not acceptable as the sole indicator of intercoder reliability. Cohen's kappa, Scott's pi, and Krippendorff's alpha (nominal) all control for chance and/or chance agreement between coders; initial Monte Carlo tests reveal only minor differences in the performance of these coefficients, and all possess similar advantages and disadvantages. All result in low (often unacceptable) values when applied to a variable that shows a "rare event" distribution and there is even moderate disagreement between the coders as to this "rare" occurrence. These characteristics point to the need for the development of alternative statistical tests. Cohen's kappa (or multi-coder kappa; Cohen 1960; Fleiss 1971) is a widely used coefficient for nominal/categorical data (Lombard et al. 2002). The threshold of acceptability for the coefficient is a point of disagreement. The most liberal criteria are provided by Banerjee et al. (1999), who hold that a kappa of .75 or higher indicates excellent agreement beyond chance, .40 to .75, fair to good agreement beyond chance, and below .40, poor agreement beyond chance. Most scholars recommend a minimum of at least .60, with some recommending .80. It is recommended here that a chance-corrected agreement coefficient of at least .60 be achieved.
>
> *For ordinal data*—Covariation or agreement controlling for chance is recommended. For the assessment of covariation with ordinal data, the main choices are Spearman rho and Krippendorff's alpha (ordinal). The threshold of the coefficients' acceptability, according to Krippendorff (2004), is .80, with coefficients between .667 and .80 allowing for only tentative conclusions. Both of these statistics measure a type of covariation that relies on rank ordering of cases, and are therefore particularly appropriate for data that reflect that technique. Often, researchers with ordinal data that instead reflect ordered categories (e.g., low, medium, high) opt for using Cohen's kappa or multi-coder kappa in order to assess agreement controlling for chance, which is recommended here.
>
> *For interval/ratio data*—While some past research has used the Pearson correlation coefficient to assess reliability, that statistic gives an indication of covariation without regard to correspondence of values (thus, a Pearson r of 1.0 may be achieved with even great disagreement, if one coder systematically codes much higher than another), and is therefore not recommended. Alternatives are Krippendorff's alpha (interval or ratio), and the ICC (intra-class correlation coefficient). These statistics are based on a variance-partitioning model, rather than a covariation model, and may or may not meet the needs of a given researcher (Shrout and Fleiss 1979). Covariation with some credit given for a greater degree of near-agreement or correspondence is recommended here. Lin's concordance coefficient (Lin 1989) is designed for this task. This coefficient emulates the Pearson correlation coefficient, but with the correlation line forced to extend through the origin and to have a defined slope of 1. A standard

threshold level for this statistic has not been established; this author currently uses a squared Lin's coefficient as a measure of shared variance (like a coefficient of determination), with a minimum of .50 (this corresponds to 50% shared variance, and a minimum Lin's coefficient of approximately .70).

A newer approach to reliability assessment is the acknowledgement that reliability statistics may be used as diagnostics, rather than simply provide a "thumbs up/ thumbs down" assessment for each measured variable. Such a diagnostic application may identify problematic variables, problematic coders ("rogue" coders), and problematic variable/coder interactions (e.g., Neuendorf 2009). Reliability assessment may result in the collapsing of categories within a single variable, or the combining of multiple variables into a scale. Again, an initial pilot reliability test gives the researcher an opportunity to conduct any desired diagnostics, and change the coding scheme as needed before final coding commences.

Regardless of the selection of a particular reliability statistic, one mathematical truism holds—reliability coefficients are more likely to be acceptable for a variable that has a reasonable amount of variance. There are several ways this might be achieved: (a) select variables that past work has indicated hold good variance in the population under examination; (b) if selecting a set of indicators that measure the same general construct, be prepared to combine these indicators in order to achieve good variance, and (c) be prepared to collapse categories within an individual variable in order to achieve a better distribution on that variable. For example, in a study of film techniques, we measured whether various colored and diffusion filters were used. Due to rare occurrence of each filter type, we combined these measures in an additive scale (Janstova et al. 2010).

A number of future scholarly endeavors may help provide all content analysts with more guidance in the selection of their "tools" for reliability assessment. The aforementioned set of tests of reliability statistics' characteristics, including Monte Carlo tests (Mooney 1997) and/or bootstrapping techniques (Hayes and Krippendorff 2007), could also provide new information on the statistics' sampling distributions (e.g., Petersson et al. 2002) and viable methods for establishing confidence intervals and tests of statistical significance for reliability statistics. For example, Shrout and Fleiss (1979) have presented confidence intervals for six different forms of the ICC, and Hayes and Krippendorff (2007) provide a demonstration of the construction of a confidence interval via bootstrapping for one version of Krippendorff's alpha. Additionally, it is hoped that the available statistics be examined and compared with regard to their response to changes in such conditions as number of coders, number of cases, level of measurement, precision of measurement, presence of missing data, and distributional characteristics of a variable (variance, skew, etc.). With the recent availability of programs that make the calculation of reliability coefficients much quicker and easier, reliability assessment may be more clearly viewed as a *process* of improving the content analysis coding scheme rather than a rigid post-hoc indicator of success or failure.

6. *Reportage.* Perhaps the most common methodological offense in content analysis research is poor documentation. Often, research articles fail to report the defined population, the method of sampling, or the precise nature of the measured variables in the codebook. Even more prevalent is the tendency to under-report the reliability assessment process—many times, it is impossible to discern whether or not reliability even has been assessed. Very frequently, reliability coefficients are not reported for each variable separately, but rather a single overall or "average" coefficient is reported. This is an unacceptable practice that may obscure the poor performance of one or more variables in the study.

Although tedious, it is appropriate to report reliability coefficients separately for each measured variable. Scharrer (2001) admirably takes the time to report intercoder reliability coefficients variable-by-variable in an endnote.

Table 1 A methodological checklist for content analysis–preparation

1. Theory—Has the role of theory been explicated fully? Is theory tested directly by the study? Or, does some theory about message sources or message impacts on receivers motivate and guide the investigation?
2. Scope—What is the scope of the data collection? Is it limited to message content, or are source and/or receiver variables also measured?
3. Past research and measurement—Has past research on the topic been fully reviewed? If previously developed coding schemes are available for use, have they been considered for adoption or revision? Have other standard measures (e.g., a self-report scale for gender roles) been considered for adaptation?
4. Population—What exactly is the defined population of messages that will be examined?
5. Immersion—Have the researchers immersed themselves in the message pool? What concepts have been derived from this immersion?
6. Human coding vs. CATA—Will human coding and/or computer coding (CATA) be utilized?

Table 2 A methodological checklist for content analysis—methodological decisions

Overall—has an up-to-date, accepted, general (not discipline-specific) content analysis methodology reference work been consulted?

1. Unitizing—what are the units of data collection (those messages or message components to which the content analysis measures are applied) and how are these units identified? If researcher or coder unitizing is required, is the reliability of this process assessed?
2. Sampling—will a sample be taken, or will the study constitute a census of the population of messages? What type of sampling will be used? Is it a representative sample?
3. Measurement—Have measures for the variables been established that are appropriate, with regard to exhaustiveness, mutual exclusivity, and level of measurement? Has careful consideration been given to measures of manifest and latent content? After data are collected, have the variables' distributions been examined, and appropriate data transformations executed? Has adequate assessment of the validity of measures—from face validity to construct validity (e.g., Carmines and Zeller 1979)—been conducted whenever possible?
4. Training—Have the coders been fully trained, via a process that includes multiple coding sessions, with possible codebook revision? Has the full protocol of training and codebook been documented, so as to assure replicability? Have reliability coding and final coding been conducted independently by the coders?
5. Reliability—Have at least two reliability subsamples been employed, one at the piloting stage and one for the final reliability check? How have these subsamples been selected? Have appropriate intercoder reliability statistics been selected? And have the minimums for these coefficients been met?
6. Reportage—Has the content analysis process been fully reported, including the theoretical backing, the origins of variables or scales, the definition of the message population, and the method of sampling? Has intercoder reliability assessment been fully reported, including reliability coefficients for each variable separately?

Conclusion

Methodological standards for quantitative content analysis seem to have lagged behind those for other research techniques. This article provides an overview of guidelines and recommendations for reviewers and researchers in the field of gender studies, in an attempt to provide support for increased rigor in the execution of content analyses. While the reader should consult a full-length content analysis methods reference work before conducting a study, this article may alert the reader to key issues that have emerged as problematic in content analysis research—including issues of the establishment of a theoretical framework, population definition, sampling, validity, reliability, and reportage.

References

An, D., & Kim, S. (2007). Relating Hofstede's masculinity dimension to gender role portrayals in advertising. *International Marketing Review, 24*, 181–207.

Anderson, D. A., & Hamilton, M. (2005). Gender role stereotyping of parents in children's picture books: The invisible father. *Sex Roles, 52*, 145–151.

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. Retrieved from http://lingcog.iit.edu/doc/gendertext04.pdf

Babbie, E. R. (2010). *The practice of social research* (12th ed.). Belmont: Wadsworth Cengage.

Baker, C. N. (2005). Images of women's sexuality in advertisements: A content analysis of black- and white-oriented women's and men's magazines. *Sex Roles, 52*, 13–27.

Balmas, M., & Scheafer, T. (2010). Candidate image in election campaigns: Attribute agenda setting, affective priming, and voting intentions. *International Journal of Public Opinion Research, 22*(2), 204–229.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics, 27*(1), 3–23.

Bem, S. (1981). *Bem sex role inventory professional manual*. Palo Alto: Consulting Psychologists Press.

Bridges, J. S. (1993). Pink or blue: Gender-stereotypic perceptions of infants as conveyed by birth congratulatory cards. *Psychology of Women Quarterly, 17*(2), 193–205.

Brinson, S. L., & Winn, J. E. (1997). Talk shows' representations of interpersonal conflicts. *Journal of Broadcasting & Electronic Media, 41*, 25–39.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Newbury Park: Sage.

Cissna, K. N., Garvin, B. J., & Kennedy, C. W. (1990). Reliability in coding social interaction: A study of confirmation. *Communication Reports, 3*(2), 58–69.

Clarke, J. N., & Everest, M. M. (2006). Cancer in the mass print media: Fear, uncertainty and the medical model. *Social Science & Medicine, 62*, 2591–2600.

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

Collins, R. L., Elliott, M. N., & Miu, A. (2009). Linking media content to media effects: The RAND Television and Adolescent Sexuality Study. In A. B. Jordan, D. Kunkel, J. Manganello, & M. Fishbein (Eds.), *Media messages and public health: A decisions approach to content analysis* (pp. 154–172). New York: Routledge.

Cressman, D. L., Callister, M., Robinson, T., & Near, C. (2009). An analysis of profanity in US teen-oriented movies, 1980-2006. *Journal of Children and Media, 3*(2), 117–134.

Curtin, P., & Gaither, K. (2003). *Public relations and propaganda in cyberspace: A quantitative content analysis of Middle Eastern government websites*. Paper presented at the annual meeting of the International Communication Association, San Diego, CA.

Danaher, B. G., Boles, S. M., Akers, L., Gordon, J. S., & Severson, H. H. (2006). Defining participant exposure measures in web-based health behavior change programs. *Journal of Medical Internet Research*, *8*, Article 3.

Dietz, T. L. (1998). An examination of violence and gender role portrayals in video games: Implications for gender socialization and aggressive behavior. *Sex Roles, 38*, 425–442.

Domhoff, G. W. (1999). New directions in the study of dream content using the Hall and Van de Castle coding system. *Dreaming, 9*, 115–137.

Dozier, D. M., Lauzen, M. M., Day, C. A., Payne, S. M., & Tafoya, M. R. (2005). Leaders and elites: Portrayals of smoking in popular films. *Tobacco Control, 14*(1), 7–9.

Ebersole, S. (2000). Uses and gratifications of the web among students. *Journal of Computer-Mediated Communication*, *6*(1). Retrieved from http://jcmc.indiana.edu/vol6/issue1/ebersole.html

Eco, U. (1976). *A theory of semiotics*. Bloomington: Indiana University Press.

Eschholz, S., Bufkin, J., & Long, J. (2002). Symbolic reality bites: Women and racial/ethnic minorities in modern film. *Sociological Spectrum, 22*, 299–334.

Evans, L., & Davies, K. (2000). No sissy boys here: A content analysis of the representation of masculinity in elementary school reading textbooks. *Sex Roles, 42*, 255–270.

Fernandez-Villanueva, C., Revilla-Castro, J. C., Dominguez-Bilbao, R., Gimeno-Jimenez, L., & Almagro, A. (2009). Gender differences in the representation of violence on Spanish television: Should women be more violent? *Sex Roles, 61*, 85–100.

Fields, A. M., Swan, S., & Kloos, B. (2010). "What it means to be a woman:" Ambivalent sexism in female college students' experiences and attitudes. *Sex Roles, 62*, 554–567.

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

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken: Wiley-Interscience.

Franiuk, R., Seefelt, J. L., & Vandello, J. A. (2008). Prevalence of rape myths in headlines and their effects on attitudes toward rape. *Sex Roles, 58*, 790–801.

Fullerton, J. A., & Kendrick, A. (2000). Portrayal of men and women in U.S. Spanish-language television commercials. *J&MC Quarterly, 77*, 128–142.

Garvin, B. J., Kennedy, C. W., & Cissna, K. N. (1988). Reliability in category coding systems. *Nursing Research, 37*(1), 52–58.

George, A. L. (1959). *Propaganda analysis: A study of inferences made from Nazi propaganda in World War II*. Westport: Greenwood.

Ghose, S., & Dou, W. (1998). Interactive functions and their impacts on the appeal of Internet presence sites. *Journal of Advertising Research, 38*(2), 29–43.

Goffman, E. (1979). *Gender advertisements*. New York: Harper and Row.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science, 5*(1), 13–34.

Gottschalk, L. A., & Bechtel, R. J. (Eds.). (2008). *Computerized content analysis of speech and verbal texts and its many applications*. New York: Nova Science.

Gray, J. H., & Densten, I. L. (1998). Integrating quantitative and qualitative analysis using latent and manifest variables. *Quality & Quantity, 32*, 419–431.

Gregory, R. L., & Zangwill, O. L. (Eds.). (1987). *The Oxford companion to the mind*. Oxford: Oxford University Press.

Groom, C. J., & Pennebaker, J. W. (2005). The language of love: Sex, sexual orientation, and language use in online personal advertisements. *Sex Roles, 52*, 447–461.

Guetzkow, H. (1950). Unitizing and categorizing problems in coding qualitative data. *Journal of Clinical Psychology, 6*, 47–58.

Haninger, K., & Thompson, K. M. (2004). Content and ratings of teen-rated video games. *JAMA, 291*(7), 856–865.

Hardy, C., Harley, B., & Phillips, N. (2004). Discourse analysis and content analysis: Two solitudes? *Qualitative methods: Newsletter of the American Political Science Association Organized Section on Qualitative Methods, 2*(1), 19–22.

Harrison, K., & Hefner, V. (2006). Media exposure, current and future body ideals, and disordered eating among preadolescent girls: A longitudinal panel study. *Journal of Youth and Adolescence, 35*, 153–163.

Hart, R. P. (2000). *The text-analysis program: Diction 5.0*. Austin: Digitext.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77–89.

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Beverly Hills: Sage.

Hubbell, A. P., & Dearing, J. W. (2003). Local newspapers, community partnerships, and health improvement projects: Their roles in a comprehensive community initiative. *Journal of Community Health, 28*, 363–376.

Ibroscheva, E. (2007). Caught between East and West? Portrayals of gender in Bulgarian television advertisements. *Sex Roles, 57*, 409–418.

Janis, I. (1965). The problem of validating content analysis. In H. D. Lasswell, N. Leites, et al. (Eds.), *Language of politics* (pp. 55–82). Cambridge: MIT.

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management, 42*(1), 248–263.

Janstova, P., Neuendorf, K. A., & Lieberman, E. A. (2010). *Empirical testing of auteur theory via content analysis*. Manuscript in preparation, Cleveland State University.

Johnston, C. A. B., & Morrison, T. G. (2007). The presentation of masculinity in everyday life: Contextual variations in the masculine behaviour of young Irish men. *Sex Roles, 57*, 661–674.

Jones, J., & Himelboim, I. (2010). Just a guy in pajamas? Framing the blogs in mainstream U.S. newspaper coverage (1999-2005). *New Media & Society, 12*, 271–288.

Kalis, P., & Neuendorf, K. A. (1989). Aggressive cue prominence and gender participation in MTV. *Journalism Quarterly, 66*, 148–154, 229.

Kinney, N. T. (2005). Engaging in "loose talk": Analyzing salience in discourse from the formulation of welfare policy. *Policy Sciences, 38*, 251–268.

Knobloch, L. K. (2008). The content of relational uncertainty within marriage. *Journal of Social and Personal Relationships, 25*, 467–495.

Kolbe, R. H., & Burnett, M. S. (1991). Content-analysis research: An examination of application with directives for improving research reliability and objectivity. *Journal of Consumer Research, 18*, 243–250.

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

Kunkel, D., Wilson, B., Donnerstein, E., Linz, D., Smith, S., Gray, T., et al. (1995). Measuring television violence: The importance of context. *Journal of Broadcasting & Electronic Media, 39*, 284–291.

Laird, L. D., de Marrais, J., & Barnes, L. L. (2007). Portraying Islam and Muslims in MEDLINE: A content analysis. *Social Science & Medicine, 65*, 2425–2439.

Lauzen, M. M., Dozier, D. M., & Cleveland, E. (2006). Genre matters: An examination of women working behind the scenes and on-screen portrayals in reality and scripted prime-time programming. *Sex Roles, 55*, 445–455.

Lieberman, E. A., Neuendorf, K. A., Denny, J., Skalski, P. D., & Wang, J. (2009). The language of laughter: A quantitative/qualitative fusion examining television narrative and humor. *Journal of Broadcasting & Electronic Media, 53*, 497–514.

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

Lindner, K. (2004). Images of women in general interest and fashion magazine advertisements from 1955 to 2002. *Sex Roles, 51*, 409–421.

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

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2004). A call for standardization in content analysis reliability. *Human Communication Research, 30*, 434–437.

Markson, E. W., & Taylor, C. A. (2000). The mirror has two faces. *Ageing and Society, 20*, 137–160.

Martins, N., Williams, D. C., Harrison, K., & Ratan, R. A. (2009). A content analysis of female body imagery in video games. *Sex Roles, 61*, 824–836.

Mastro, D. E., Eastin, M. S., & Tamborini, R. (2002). Internet search behaviors and mood alterations: A selective exposure approach. *Media Psychology, 4*, 157–172.

McAdams, D. P., & Zeldow, P. B. (1993). Construct validity and content analysis. *Journal of Personality Assessment, 61*, 243–245.

McCroskey, J. C. (1993). *An introduction to rhetorical communication* (6th ed.). Englewood Cliffs: Prentice Hall.

McMillan, S. J. (2000). The microscope and the moving target: The challenge of applying content analysis to the World Wide Web. *Journalism and Mass Communication Quarterly, 77*, 80–98.

Messineo, M. J. (2008). Does advertising on Black Entertainment Television portray more positive gender representations compared to broadcast networks? *Sex Roles, 59*, 752–764.

Milburn, S. S., Carney, D. R., & Ramirez, A. M. (2001). Even in modern media, the picture is still the same: A content analysis of clipart images. *Sex Roles, 44*, 277–294.

Miller, M. K., & Summers, A. (2007). Gender differences in video game characters' roles, appearances, and attire as portrayed in video game magazines. *Sex Roles, 57*, 733–742.

Mooney, C. Z. (1997). *Monte Carlo simulation*. Thousands Oaks: Sage.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks: Sage.

Neuendorf, K. A. (2004). Content analysis: A contrast and complement to discourse analysis. *Qualitative methods: Newsletter of the American Political Science Association Organized Section on Qualitative Methods, 2*(1), 33–36.

Neuendorf, K. A. (2009). Reliability for content analysis. In A. B. Jordan, D. Kunkel, J. Manganello, & M. Fishbein (Eds.), *Media messages and public health: A decisions approach to content analysis* (pp. 67–87). New York: Routledge.

Neuendorf, K. A., & Kane, C. L. (2010). The content analysis guidebook online. Retrieved from http://academic.csuohio.edu/kneuendorf/content

Neuendorf, K. A., & Skalski, P. D. (2010). *Content analysis: Description, prediction, and explanation*. Paper presented to the Social Science and Social Computing Workshop, University of Hawaii, Honolulu, HI.

Neuendorf, K. A., Gore, T. D., Dalessandro, A., Janstova, P., & Snyder-Suhy, S. (2010). Shaken and stirred: A content analysis of women's portrayals in James Bond films. *Sex Roles, 62*, 747–761.

Norris, P. (2003). Preaching to the converted? Pluralism, participation and party websites. *Party Politics, 9*(1), 21–45.

Odekerken-Schroder, G., De Wulf, K., & Hofstee, N. (2002). Is gender stereotyping in advertising more prevalent in masculine countries? *International Marketing Review, 19*, 408–419.

Ogletree, S. M., Merritt, S., & Roberts, J. (1994). Female/male portrayals on U.S. postage stamps of the twentieth century. *Communication Research Reports, 11*, 77–85.

Pan, P.-L., Meng, J., & Zhou, S. (2010). Morality or equality? Ideological framing in news coverage of gay marriage legitimization. *The Social Science Journal, 47*, 630–645.

Pasadeos, Y., Huhman, B., Standley, T., & Wilson, G. (1995). *Applications of content analysis in news research: A critical examination*. Paper presented to the Communication Theory and Methodology Division of the Association for Education in Journalism and Mass Communication, Washington, DC.

Patchin, J. W., & Hinduja, S. (2010). Trends in online social networking: Adolescent use of MySpace over time. *New Media & Society, 12*(2), 197–216.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic inquiry and word count (LIWC2007)*. Austin, TX: www.licw.net.

Petersson, H., Gill, H., & Ahlfeldt, H. (2002). A variance-based measure of inter-rater agreement in medical databases. *Journal of Biomedical Informatics, 35*, 331–342.

Pollock, J. C., & Yulis, S. G. (2004). Nationwide newspaper coverage of physician-assisted suicide: A community structure approach. *Journal of Health Communication, 9*, 281–307.

Potter, W. J. (2009). Defining and measuring key content variables. In A. B. Jordan, D. Kunkel, J. Manganello, & M. Fishbein (Eds.), *Media messages and public health: A decisions approach to content analysis* (pp. 35–52). New York: Routledge.

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research, 27*, 258–284.

Potter, J., Linz, D., Wilson, B. J., Kunkel, D., Donnerstein, E., Smith, S. L., et al. (1998). Content analysis of entertainment television: New methodological developments. In J. T. Hamilton (Ed.), *Television violence and public policy* (pp. 55–103). Ann Arbor: The University of Michigan Press.

Radwin, L. E., & Cabral, H. J. (2010). Trust in Nurses Scale: Construct validity and internal reliability evaluation. *Journal of Advanced Nursing, 66*, 683–689.

Ricciardelli, R., Clow, K. A., & White, P. (2010). Investigating hegemonic masculinity: Portrayals of masculinity in men's lifestyle magazines. *Sex Roles, 63*, 64–78.

Riffe, D., Lacy, S., & Fico, F. G. (2005). *Analyzing media messages: Using quantitative content analysis in research* (2nd ed.). Mahwah: Lawrence Erlbaum.

Roberts, C. W. (Ed.). (1997). *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts*. Mahwah: Lawrence Erlbaum.

Scharrer, E. (2001). From wise to foolish: The portrayal of the sitcom father, 1950s–1990s. *Journal of Broadcasting & Electronic Media, 45*, 23–40.

Schlenker, J. A., Caron, S. L., & Halteman, W. A. (1998). A feminist analysis of *Seventeen* magazine: Content analysis from 1945 to 1995. *Sex Roles, 38*, 135–149.

Schmader, T., Whitehead, J., & Wysocki, V. H. (2007). A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles, 57*, 509–514.

Shapiro, G., & Markoff, J. (1997). A matter of definition. In C. W. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* (pp. 9–34). Mahwah: Lawrence Erlbaum.

Shoemaker, P. J., & Reese, S. D. (1996). *Mediating the message: Theories of influences on mass media content* (2nd ed.). White Plains: Longman.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Simon, T. F., Fico, F., & Lacy, S. (1989). Covering conflict and controversy: Measuring balance, fairness and defamation in local news stories. *Journalism Quarterly, 66*, 427–434.

Skymeg Software. (2009). *Program for reliability assessment with multiple coders (PRAM)*. Unpublished manuscript, Cleveland, OH.

Smith, A. M. (1999). *Girls on film: Analysis of women's images in contemporary American and "Golden Age" Hollywood films*. Masters thesis, Cleveland State University, Cleveland, OH.

Spence, J., Helmreich, R., & Stapp, J. (1974). The Personal Attributes Questionnaire: A measure of sex-role stereotypes and masculinity-femininity. *JSAS Catalog of Selected Documents in Psychology, 4*, 43.

Uray, N., & Burnaz, S. (2003). An analysis of the portrayal of gender roles in Turkish television advertisements. *Sex Roles, 48*, 77–87.

Valls-Fernandez, F., & Martinez-Vicente, J. M. (2007). Gender stereotypes in Spanish television commercials. *Sex Roles, 56*, 691–699.

Weare, C., & Lin, W. Y. (2000). Content analysis of the World Wide Web—Opportunities and challenges. *Social Science Computer Review, 18*, 272–292.

Weber, R. P. (1990). *Basic content analysis* (2nd ed.). Newbury Park: Sage.

Weber, R., Behr, K.-M., Tamborini, R., Ritterfeld, U., & Mathiak, K. (2009). What do we really know about first-person-shooter games? An event-related, high-resolution content analysis. *Journal of Computer-Mediated Communication, 14*, 1016–1037.

West, M. D. (Ed.). (2001). *Theory, method, and practice in computer content analysis*. Westport: Ablex.

Xue, F., & Ellzey, M. (2009). What do couples do? A content analysis of couple images in consumer magazine advertising. *Journal of Magazine and New Media Research, 10*(2), 1–17.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103*, 347–387.