7

Quantitative Content Analysis and the Measurement of Collective Identity

Kimberly A. Neuendorf and Paul D. Skalski

CONTENT ANALYSIS INTRODUCED

Content analysis, simply put, is the quantitative investigation of message characteristics. Most definitions are a bit more specific than this, often delineated by the degree to which a scientific method is assumed.[1] The following definition is employed 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) and is not limited as to the types of variables that may be measured or the context in which the messages are created or presented. (Neuendorf 2002: 10)

In content analysis, as in all quantitative investigations, the quality of a measure is dependent on several factors. First, there must be a clear conceptualization of the construct of interest, for it is the congruence between conceptualization and operationalization (measurement) that constitutes basic internal validity (Babbie 1998).[2]

First, We Conceptualize

With the construct of collective identity, conceptualization can be problematic, in that theoretic approaches abound. Abdelal et al. (2006: 695) refer to the

[1] Some definitions allow for direct inferences from message characteristics to source or receiver characteristics, but this is a point of debate (Berelson 1952; Neuendorf 2002; Riffe, Lacy, and Fico 2005; R. Weber 1990).

[2] Other key criteria must be met – external validity (the generalizability of the measure), and reliability (something that, when human coders are engaged, is essential to the content analytic measure). Additionally, the measures ought to be as precise and accurate, and at as high a level of measurement as possible.

"definitional anarchy" of identity research and, as Bruland and Horowitz (2003: 1) note, "the existence of identity as a universal but largely implicit concept makes it difficult to isolate and understand its use." It is not the purpose of this chapter to exhaustively cover all theoretic or conceptual approaches to the study of collective identity,[3] as this has been addressed elsewhere in this volume, most clearly in the four types of identity delineated by Abdelal et al. In this chapter, various conceptual definitions have been utilized as the bases for quantitative content analyses, and a specific definitional assumption is presented for each example given in this chapter. Thus, we present a variety of content analysis measurement techniques, based on portions of the wide array of possible conceptualizations for collective identity.

Previous Research

In the past, content analysis has rarely been used explicitly to measure identity of any type. But content analysis has been used to measure an incredible range of message characteristics. Scholars have charted the violent content of television (Wilson et al. 1997), the gender role portrayals of film characters (A. M. Smith 1999), the structure of news Web sites (Scharl 2004), and even categories of restroom graffiti (Schreer and Strichartz 1997). Other analyses have used the content analysis of naturally occurring speech to pinpoint patterns indicative of psychological pathologies (Gottschalk 1995). And contrary to some authors' opinions, content analysis is *not* limited to simple counting of messages. Some analyses have been layered and complicated, some have charted complex changes over time, and some have identified significant statistical relationships between

[3] For reviews of the use of identity in the social scientific literature, see Ashmore, Deaux, and McLaughlin-Volpe 2004 and Gleason 1983. Here, a brief summary of how identity has been conceptualized might be useful. On a basic level, Horowitz (2002), in his study of the use of a multitude of identity constructs in the academic literature of international relations, has delineated between definitions that are "essentialist" (i.e., preestablished or primordial), socially "constructed," or a blend of the two. Additionally, a review of the relevant literature finds four nonmutually exclusive types: (1) self-identification (i.e., the individual decides), as demonstrated in work on cultural identity (Brass 1991), perceptions of self (Bem 1972; Berger and Luckmann 1966; Fiske and Taylor 1991), and self-concept (Burke and Tully 1977; D. M. Taylor 2002; J. Turner 1985); (2) attribution by others, as exemplified by work by John Turner (1985); (3) identity as defined by the roles one plays, pioneered by Mead (1934) and further explored in political and other contexts (Brewer and Gardner 1996; Monroe, Hankin, and van Vechten 2000), and referred to as "identity theory" by some scholars (Stryker 1987); (4) one's identity as defined by one's position in a larger aggregate, variously referred to as collective identity (Ashmore, Deaux, and McLaughlin-Volpe 2004), group affiliation (Brubaker and Cooper 2000; J. Turner 1985), social identity (Abrams and Hogg 1990, 1999; Reykowski 1994; Tajfel and Turner 1986), and acculturation (Berry 1980; Chun, Organista, and Marin 2003; Padilla 1980). These four ways of viewing identity, derived from a review of the literature, are compatible with those presented by Abdelal et al (2006).

message content and social movements – for example, as in work indicating that news coverage precedes, rather than follows, public opinion on critical topics (Hertog and Fan 1995; Willnat and Zhu 1996; also see Neuendorf 2002 for examples of the complexities to which content analysis may be applied).

The application of content analysis to political science arenas has been rather limited, although the methodology seems to be growing in popularity, particularly when used in concert with other, more qualitative methods.[4] For example, Yoshiko Herrera's (2005) book on contemporary Russian regionalism freely integrates original content analyses with more traditional historical, economic, and discourse analyses of the context in which regionalism has developed. In one of her two quantitative content analyses, she examined Sverdlovsk local newspaper articles whose content was relevant to the sovereignty movement for a Urals Republic, finding evidence that the movement was characterized in the press by "negative interpretations of economic conditions, and, in particular, concerns over constitutional inequality and economic autonomy" (Herrera 2005: 10). She also found key differences in whether the articles contained arguments favoring or opposing the Urals Republic, with local communicators more favorable than those in Moscow.

National identity was the subject of a content analysis by Laitin (1998), who examined the frequency of identity term use in Russian-language newspapers. Laitin compared articles from Kazakhstan, Estonia, Latvia, and Ukraine to get a sense of how identities had been formed within the post–Soviet Union republics. To do this, he developed a number of analysis categories specifically tailored to Russian populations (e.g., references to Russian-speakingness) that could be used to classify and discriminate along identity lines. Though it has a strong focus on Russian identity, Laitin's research suggests several possibilities for using content analysis to study national identity in general. Importantly, his content analysis is one part of a well-integrated track of study that includes survey and experimental work, as well as more qualitative ethnographic and discourse analyses.

In one of the best examples of content analysis by political scientists, Richard Merritt examined markers of common identity in colonial newspapers as part of his book *Symbols of American Community,* 1735–1775 (1966). Building on the pioneering work of his adviser Karl Deutsch (1953), who used materials such as first-class letters and phone calls as indicators of emerging national identity in his

[4] Discourse analysis, a more qualitative and "inductive" method of analyzing messages (Hardy, Harley, and Phillips 2004), has enjoyed greater popularity in the political science literature. The pros and cons of discourse analysis and content analysis were explored in a special symposium in *Qualitative Methods: Newsletter of the American Political Science Association Organized Section on Qualitative Methods* in 2004 (see Hardy, Harley, and Phillips 2004; Hopf 2004; Lowe 2004a; Neuendorf 2004a).

seminal work on national identity, Merritt developed a richly complex and carefully delineated coding scheme for his newspaper analysis. In what he called "symbol analysis research," Merritt's focus was on place-name words that were not solely geographic, such as colony names (e.g., Virginia, Pennsylvania), British symbols (e.g., British Isles, Irish), and symbols of common identity (e.g., Americans, British Americans). He was able to chart the growing trend for newspapers to refer to Americans as a single group, and an increasing use of terms identifying the colonists as Americans rather than members of a British political community (Merritt 1966: 180). Merritt, ahead of his time, executed a basic intercoder reliability assessment and explicated his entire coding scheme as an appendix (resulting in high replicability).

In a rare example of work that looks at individual identity measured via content analysis, Stone (1997) reports on proprietary research that identified four ways that individuals talk about a topic: goals and gratifications, rules and responsibilities, feelings and emotions, and unique outlook and ways of understanding. Typically applied to consumers' reactions to products and services, this typology of identity, Stone notes, could be applied in nonconsumer contexts, such as one's own national identity as an American.

And in a unique application of basic content analysis methods to nation-level analyses, George Barnett and Han Woo Park (2004) have collected and analyzed data on international hyperlinks. Barnett (2004: 2) suggests that, because of increased communicative contact via the Internet, differences among national cultures will diminish, resulting in the formation of a single "transnational identity." In earlier work, Barnett (2002) tracked international telephone calls over the period 1978 to 1999, concluding that during the latter part of the period, a trend toward decentralization was exhibited, with East Asian nodes becoming less integrated with North American and Western European nodes. These studies show how the sheer *volume* of communication might be studied to draw conclusions about the connectedness of different nations and therefore their unique or increasingly interdependent "identities."

CONTENT ANALYSIS OPTIONS

There are several ways in which content analysis methodologies appropriate to the task of identity research might be considered. First, there are basic choices for execution of a content analysis.

Human Preset Coding

The historically standard method of executing content analysis is via "human" coding, the application of a set of written rules for measurement to a set of

messages by trained individuals. This a priori development of a coding scheme is the norm in classical content analysis – a researcher devises measures based on theory, past research, and, often, pilot work that includes immersion in the message pool under investigation. The coding scheme is made up of a code book (instructions to the trained coders) and a coding form (a form or questionnaire-type sheet or spreadsheet into which coders write or enter their assessments). It should be noted that such human coding is not limited to the analysis of text; images, emblems, and symbols may all serve as the messages one might analyze for the measurement of collective identity. With all human-coded projects, intercoder *reliability* is vital, and variables for which it is not achieved should be dropped from the analysis.[5] In content analysis, the emphasis is on the coding scheme as the critical instrument, rather than on the observer's or coder's expertise.

Computer (CATA) Preset Coding

In recent decades, there have been numerous advances in the development of automated "machine" coding of text messages – that is, computer-assisted text analysis (CATA). Usually, CATA coding involves the use of preset dictionaries, that is, lists of words and/or word combinations that are counted via a computer application. These dictionaries may be provided in the software or created by the researcher. In the case of programs such as the General Inquirer and Diction 5.0, numerous measures are already built into the program, some with mathematical algorithms that go beyond simple word counts. Other programs, such as VBPro and Yoshikoder, require that the researcher establish dictionaries. Some, such as Diction 5.0 and WordStat, allow for both built-in and researcher-devised dictionaries.[6]

Devising one's own dictionary is typically a long and painstaking process. First, words consistent with the conceptual definition of that construct must be

[5] Reliability criteria vary, but all content analysis methodology sources agree – for human coding, intercoder reliability should be measured by an appropriate statistic and reported for each measured variable separately (Krippendorff 2004; Lombard, Snyder-Duch, and Bracken 2002; Neuendorf 2002). Appropriate reliability coefficients include Cohen's kappa, Krippendorff's alpha, and Lin's concordance coefficient, but not simple percent agreement, which fails to remove the effect of chance agreement. Reliability assessment requires that a representative subset of the messages under study be coded independently by two or more trained coders.

[6] Nearly a score of other CATA packages are available. See Neuendorf 2002; Skalski 2002 and the corresponding Content Analysis Guidebook Online (http://academic.csuohio.edu/kneuendorf/content); Alexa and Zuell 2000; Popping 2000; Lowe 2004b or the Text Analysis Info Web site (http://www.textanalysis.info) for more options and comparisons. There are CATA programs that attempt higher-order functioning such as grammatical parsing; however, these are uniformly weak in their summative powers – that is, they tend to act more simply as aids to human perusal than as true content analysis engines. The bottom line is that none of the CATA programs have been specifically designed or adapted for the measurement of identity.

exhaustively identified. Then, variations on those root terms need to be added (e.g., if "pleasure" is a root word, then "pleasurable," "pleasure-seeking," "pleasured" and others might be added). Checks for inappropriate variations and for words too ambiguous to be validly included need to be made (e.g., "pleasant" may be deemed to be too far from the conceptual origins of the root word). Some CATA programs and related adjunct softwares include dictionary-building facilities. For example, WordStat is programmed to allow the addition to a base dictionary of antonyms, synonyms, similar terms, hypernyms, hyponyms, holonyms, and several other classes of words. And, WordStat and other programs allow for "wild card" specifications of root words (e.g., including "pleasur*" would capture all words beginning with "pleasur"). But, the researcher needs to be intimately involved in the decision to include or exclude each term; there is nothing automatic about dictionary construction.

Computer (CATA) Emergent Coding

This type of coding is less universally accepted among content analysts, because of its apparent deviation from the accepted positivist, a priori assumptions of the method. In "emergent coding," dimensions or patterns of text are derived from the data at hand (i.e., the texts under investigation), without any preset dictionaries.[7] Such programs as CATPAC, TextSmart, and TextAnalyst are well designed to allow a visual representation of the frequency of occurrence, co-occurrence, and/or correspondence of words and text segments through cluster analysis, multidimensional scaling, and neural networking. We present this non-traditional option for two reasons: advances in computer text content analysis have made some emergent techniques more objective and replicable; and, increasingly, we and others have found emergent techniques to be useful tools in the early stages of a content analysis project. That is, dimensions derived from emergent coding may be used in a second-stage, ordinary preset coding process.

THE SELECTION OF DATA RELEVANT TO THE MEASUREMENT OF COLLECTIVE IDENTITY

The Pragmatics of Content Analysis

In all content analyses, one must decide what "data" will be collected – messages, to be sure. But, what types of messages? And is the goal to simply describe the messages, or make inferences to the sources of those messages?

[7] Note that we do not address the possible option of "emergent human coding" as we view this as simply pilot work, an essential part of the process of developing a preset coding scheme.

Early content analysis scholars (e.g., Berelson 1952) argued that content analysis was a gateway to inferences that would illuminate the intentions of speakers. Even some contemporary authors (e.g., Krippendorff 2004; R. Weber 1990) allow for ready inferences from message to source intentionalities or receiver impacts. But elsewhere (Neuendorf 2002) we argue against the inferential approach to the conduct of content analysis,[8] presenting instead what is called an "integrative model of content analysis," which calls for the collation of message-centric data with other available empirical information regarding source, receiver, channel, or other contextual state, whenever possible.

An alternative to trying to draw logical or statistical inferences from message characteristics is to focus on the messages themselves. This follows nicely from the seminal work of Watzlawick, Bavelas, and Jackson (1967), who introduced the framing of the "pragmatics of human communication," wherein the focus is on the behavioral effects of communication for both receiver and source. Following Gregory Bateson's (1958: 175–176) definition of social psychology as "the study of the reactions of individuals to the reactions of other individuals," Watzlawick, Bavelas, and Jackson (1967) focus on the message content rather than the traits and states of source or receiver. However, they extend their pragmatic approach to *all* aspects of message exchange, including nonverbals and paralinguistic cues of all types, making the complete study of communication rather challenging.

Here, we adopt an essentially pragmatic approach to content analysis with regard to collective identity. When conceptual definitions of identity include message attributes, then direct measurement via content analysis is appropriate. When conceptual definitions of identity are focused more on internal states (e.g., cognitive structure) or motivations, then content analysis might not be appropriate as a central measurement technique.

Individual versus Aggregate Messages

The messages to be analyzed via content analysis may be either at the individual or the aggregate level – that is, generated by an individual or forwarded by

[8] The only case in which a type of inference from message data alone becomes a viable option is when the linkages between message data and source or receiver data have become well established through replication via numerous research investigations over time. For example, the works of Gottschalk (1995) and certain thematic content analysis coding schemes (e.g., Veroff's power motivation scheme; 1992) have established well-worn paths between psychological traits and states and message characteristics indicative of their status. These researchers have validated their schemes with multiple datasets over time. Such validation between message characteristics and source attitudes, cognitions, and/or behaviors takes the form of construct validity (testing the measure's relationship to other constructs that theory would predict would relate) or criterion validity (testing the measure's relationship with a relevant behavior or action that is external to the measure) (Carmines and Zeller 1979).

an institution representing a collective entity.[9] The collection and analysis of individual-level messages are somewhat straightforward. Original speech or text may be collected either as it occurs naturally (e.g., transcripts of groups discussions; chat room postings; letters to the editor) or in response to prompts (i.e., using content analysis as applied to essays or open-ended responses to a questionnaire or interview protocol). When attempting to access aggregate (e.g., nation-level) messages, the task provides many options. Cultural products providing the grist for content analysis at the cultural or national level are varied.[10] They might include[11] official codifications (e.g., constitutions, laws; see, e.g., Stratigaki 2004), official news releases, news stories about a nation (either internal or external, if one adopts an attributional perspective to identity; see, e.g., Chang 1998), official Web sites, or other cultural products such as folk ballads, the visual arts, television programming, commercial film releases, and architecture (Carley 1994; Chon, Barnett, and Choi 2004; Corn 1999; Custen 1992; Sirgy et al. 1998; A. D. Smith 1993). They might also

[9] It should be noted that comparisons are possible – for example, we may compare individual-level messages about the self with those representing the collective in order to assess the "fit" of the individual into the larger collective identity. This may match a conceptual definition of collective identity that focuses particularly on the individual's position within a larger social construction.

[10] National identity seems uniquely suited to measurement via content analysis of cultural products. As noted by Ringrose and Lerner (1993: 1), "the concept of the nation is one of the most powerful and ubiquitous organizing principles of modern times." National identity (Hooson 1994) may be viewed as a particular application of identity, discrete from the construct of nationalism (most commonly referring to political movements; see Goldmann, Hannerz, and Westin 2000; Lazarus 1999; Oshiba, Rhodes, and Kitagawa Otsuru 2002). In the social scientific literature, national identity has been mainly studied ad hoc, with anecdotal evidence for a particular nation at a particular point in history. Specific pieces of evidence include the use of language in border territories, French wine, Argentinean sports, and Brazilian cinema. However, national identity also may be conceptualized in the range of manners explored here for identity in general – as a psychological construct, as attributions, as role structures, and as collectivities. Clearly, identity is a multifaceted construct and may be measured via a wide variety of variables. In the study of national identity, numerous scholars have acknowledged a probable weakening of national identities in the wake of cultural imperialism (Lazarus 1999), media imperialism (Nordenstreng and Schiller 1993), and Internet penetration (Chinn and Fairlie 2004). Others have examined specifically the growth of world consumer culture (Costa and Bamossy 1995; Neuendorf, Blake, and Valdiserri 2003), the transnational nature of ethnicity or cultural identity (Featherstone 1990; Goldmann, Hannerz, and Westin 2000; A. D. Smith 1990), the "globalization" of political, economic, and social spheres (Yamada 2002), the growth of a "global identity" (Crawford 2005), and the growing cultural hegemony of American mass media (Gitlin 1979; Kellner 1990). Importantly, these trends might readily be tested via content analysis of representative cultural products.

[11] Please note that none of these studies of cultural products have included explicit measures of identity.

be official political communications such as party platforms (Budge and Hofferbert 1996), speeches and debates (Satterfield 1998), governmental annual reports (Andersson et al. 2003), national leaders' vision statements (Oliver 2004), and other political documents (Anheier, Neidhardt, and Vortkamp 1998; Beriker and Druckman 1996).[12] Other representations might include such quasi-official sources as history textbooks (Gordy and Pritchard 1995; Holt 1995).

Schwartz (1994) has taken on the challenge of measuring constructs (in his case, cultural values) at both the individual and aggregate levels, attempting an exhaustive coverage of essential cross-cultural values. He concludes that at the aggregate level, one must rely on cultural products, making this task ideal for the method of content analysis. Similarly, Inglehart and Baker (2000: 19) contest that from cross-cultural differences derive national cultures, which are then transmitted by educational and mass-media institutions. Anthony Smith (1993) sees the nation as an "imagined community," whose members will never know most of their fellow members, a construction made possible by the technologies of communication (e.g., the newspaper).

A MODEL OF CONTENT ANALYSIS OPTIONS FOR THE MEASUREMENT OF COLLECTIVE IDENTITY

If we consider both the type of execution of the content analysis and the type of "data" collected for the measurement of collective identity, we may develop a useful typology of options for such measurement. We note a three-type breakdown of the nature of the messages to which one might apply measures of collective identity:

1. Response-based messages (not naturally occurring; individuals generate messages in response to assigned identity-related tasks or prompts)
2. Naturally occurring messages that one might *assume* to constitute identity messages (e.g., individual-level messages such as personal ads or speeches, or aggregate-level cultural products such as films, news stories, or literature that might represent a culture or society)
3. Naturally occurring messages from which identity messages might be *extracted* (i.e., the messages are not wholly identity-related and must be inspected and partitioned, and then the identity messages may be

[12] It should be noted that nearly all extant content analyses of official national or political messages have examined European communications.

analyzed; this is a two-step process, involving inspection or extraction according to stated rules, followed by coding).

An intersection of the three types of execution and the three types of data produces the following typology (see Table 7.1). We hold that the intersection of emergent coding and "extracted" identity messages is logically inconsistent. Thus, there are eight possible options for content analysis coding of identity measures. In the sections that follow, we present a brief example of each – some are original small analyses, and others are examples of previously conducted research.

Option 1: Human Preset Coding, Response-Based Messages

A small, original exemplar dataset served as the raw material for several examples, including this one. The data were open-ended responses to prompts that asked respondents to describe themselves "as an American" and to describe the "typical American." Twenty individuals associated with the communication program at Cleveland State University (i.e., graduate and undergraduate students, staff) completed unrestricted written responses to these national identity prompts.

An example coding scheme for human-coding measurement was constructed. The "National Identity Pilot Coding Scheme" was devised to demonstrate key options for human coding and is shown in the Appendix. The scheme was applied to the responses to the "self as American" prompt by two coders. The analysis focused on each sentence in the verbal descriptions collected, counting the number of sentences that contained mention of each of the series of content dimensions. The measures were derived from several sources. Some of these measures were established from a simple review of the various ways in which collective identity has been conceptualized and measured in

TABLE 7.1. *A Model of Content Analysis Options for the Measurement of Collective Identity*

	Human Preset Coding	**Computer (CATA) Preset Coding**	**Computer (CATA) Emergent Coding**
Response-based messages	Option 1	Option 2	Option 3
"Assumed" identity messages (naturally occurring)	Option 4	Option 5	Option 6
"Extracted" identity messages (naturally occurring)	Option 7	Option 8	Not applicable

survey, experimental, and qualitative work – the ''macro'' variables of self-identification, role identification, and collective identification. Four measures were based directly on Stone's (1997) conceptualization of identity as manifested in four distinct ways: goals and gratifications, rules and responsibilities, feelings and emotions, and a unique outlook and way of understanding. Other measures were derived from elements of Ashmore, Deaux, and McLaughlin-Volpe's (2004) model for collective identity. Their model includes a compilation from several earlier theoretic treatments,[13] plus their own additions. Finally, from pilot work that examined the respondent essays with regard to both self as American and typical American, using simple visual inspection and a word-frequency analysis via CATPAC, several measures were added: physical characteristics, ascriptive social categories, rights, freedom, power, opportunity, patriotism, and worldview. Across the sets of measures, those for which a logical negative was possible had the added option of negation (e.g., ''The typical American is *not* very religious''). These various measures represent common choices for coding scheme development – basing measures on existing conceptualizations and models, adapting measures from other contexts, and developing measures from a pilot inspection of the ''raw'' data.

As with every human coding scheme, revisions were made after coder training and trial coding. A ''wide net'' was cast with regard to attempted measures, with the anticipation that some measures would ''fail'' because of infrequency of occurrence or poor intercoder reliability. Typically, some measures were indeed found wanting – that is, in order to explore the viability of applying Abdelal et al.'s (2006) delineation of content versus contestation, the dimension of contestation was adapted for some measures.[14] This dimension was dropped when it was found that it was rarely invoked in the content under examination and when intercoder reliability was found to be unsatisfactory.[15]

[13] Ashmore, Deaux, and McLaughlin-Volpe (2004) combine elements from Tajfel and Turner's (1986) social identity theory, J. Turner et al.'s (1987) self-categorization theory, Stryker's (1987, 2000) identity theory, and Cross's (1991; as cited by Ashmore, Deaux, and McLaughlin-Volpe 2004) nigrescence model in a comprehensive model of social identity.

[14] The following language was included in the code book: ''Additionally, when the number of sentences in (a) is greater than 0, then in (c) record whether there is contestation reported for the sentences referred to in (a). Contestation refers to language that contests, or debates, the particular target topic; this contestation may be first person (e.g., 'I'm not certain that this is typical for Americans . . .') or third person (e.g., 'Most people would not view the typical American as very isolationist, but I . . .').''

[15] Additionally, a fair number of other measures did not exhibit sufficient range in the small pilot dataset for meaningful, statistical reliability assessment (Neuendorf 2002). Thus, the coding scheme should be considered ''under construction,'' presenting a range of possibilities for researchers but not a final coding scheme.

The human coding measures summarized in Table 7.2 show that the majority of the descriptive content (76.3 percent of sentences, on average) used a self-identification mode, with about 10 percent of sentences including reference to some type of social role. References to collective identification were less frequent, at 6.2 percent. With regard to Stone's (1997) template of identity, the most common framing for one's national identity seems to be via feelings and emotions, with an average of 21.7 percent of statements containing such a reference. This is followed by 13.9 percent of statements with reference to goals and gratifications, 5.6 percent containing reference to rules and responsibilities, and 4.7 percent having some reference to a unique outlook or ways of understanding. The measures derived from Ashmore et al. (2004) showed 20.9 percent of a respondent's statements to be positive in tone and 17.4 percent negative, with, on average, 6.9 percent of a respondent's sentences containing reference to behavioral involvement and 5.3 percent to ideology. Finally, with regard to "novel" measures devised particularly for this analysis, ascriptive social categories were the most prevalent mode of description, contained in 13.1 percent of a respondent's sentences. Other framings were less prevalent –

TABLE 7.2. *Option 1: Findings for Human Coding of "Self as American" Essays (average percent of sentences)*

Number of Sentences	**Mean = 10.5 per Essay**
Self-identification	76.3
Role identification	9.9
Collective identification	6.2
Goals and gratifications	13.9
Rules and responsibilities	5.6
Feelings and emotions	21.7 (2% negated)
Unique outlook or ways of understanding	4.7
Positive	20.9 (5% negated)
Negative	17.4 (3% negated)
Ideology	5.3 (13% negated)
Behavioral involvement	6.9
Physical characteristics	0.0
Ascriptive social categories	13.1
Rights	8.0 (8% negated)
Freedom	8.6
Power	0.9 (50% negated)
Opportunity	2.8
Patriotism	8.2 (46% negated)
Worldview	8.5 (26% negated)

8.6 percent of sentences contain reference to freedom, 8.5 percent to worldview, 8.2 percent to patriotism, 8.0 percent to rights, 2.8 percent to opportunity, and only 0.9 percent to power.

Some of the advantages and disadvantages of human coding are apparent. More nuanced measures are possible than with computer coding, including disambiguation and the recognition of negation.[16] However, human coding is labor-intensive and completely dependent on the ability of the researcher to create a coding scheme and training procedure that will result in reliable measures.

Option 2: Computer (CATA) Preset Coding of Response-Based Messages

For this example of CATA preset coding, the same national identity exemplar dataset was used. At present, no standard CATA dictionaries have been developed for the express purpose of measuring identity. Two computer applications with preset built-in dictionaries were selected as most relevant to the measurement of identity: the General Inquirer (used in this example) and Diction (to be used in example 5).

The venerable General Inquirer (Buvac and Stone 2001) includes more than 180 dictionaries and subdictionaries measuring a wide variety of attributes, with the goal of what is called "thematic content analysis" – the attempted measurement of psychological constructs via the analysis of messages. Composed principally of constructs from the Harvard IV-4 dictionary and the Lasswell value dictionary, it continues to be adapted for a variety of purposes. A number of constructs measured by the General Inquirer might meet the needs of a researcher whose conceptual definition of national identity includes expressions of affective states (e.g., positive tone, negative tone, emotionality), role identifications (e.g., religion, political, work, family, academic), or expressions of attachment to larger social units (e.g., race).

The General Inquirer system demands that each message to be analyzed be contained in a separate text file. Thus, the sample dataset of twenty responses to the two prompts resulted in forty separate files, submitted "batch" to the PC version of the program. The dataset that was generated included both basic dictionary counts and percentage of words figures.

[16] A pilot analysis reported by Neuendorf (2004b) compared CATA measures with comparable human-coded versions for both Diction and the General Inquirer computer applications. Importantly, human coding found far more "political" references than did General Inquirer; human coding identified fewer than half as many references to "centrality" than did Diction, and 15 percent of the human-coded references were instances of negation, a central concern for the use of CATA.

Sample results for this analysis are presented in Table 7.3. Here we see a comparison between texts generated by male and female respondents for selected General Inquirer measures. Differences are not major, although we see a tendency for the male respondents to use criteria related to strength (i.e., the dictionaries for "strong" and "weak") when describing their own national identity more often than do females, and for males to use language referring to things "academic" less frequently than do females.

We see some of the typical advantages and disadvantages of preset CATA coding – a large amount of text is analyzed very quickly; however, CATA is insensitive to (i.e., cannot measure) nuances such as negation and colloquial speech, which can be tapped by trained human coders, and generally cannot disambiguate (e.g., note the difference among a "fine" for a traffic violation, "fine" linens, and feeling "fine"). Additionally, although the ready-made measures provided by the General Inquirer are attractive for their ease, the researcher's needs may not be met precisely by the particular dictionaries available.[17]

Option 3: Computer (CATA) Emergent Coding of Response-Based Messages

Once again, the exemplar open-ended text data were used for this sample analysis, focusing on descriptions of the "typical American." Several CATA programs allow for some type of emergent coding. CATPAC is a particularly intriguing option for discovering dimensions of discourse and concept differentiation as emergent from text. For our purposes, it might allow us to discover the most common terminology used by individuals to describe national identity, and to see the patterns of word co-occurrence that could reveal clusters of concepts and/or dimensions of concept differentiation. Here, the implicit conceptualization of identity is, quite simply, the manner in which one verbally identifies the "typical American."

[17] Ready-made dictionaries included with CATA programs should be scrutinized before being used, to assure their quality and appropriateness for a particular project. And some CATA dictionaries are simply not intended to be used for anything other than demonstration purposes, such as the "seeking" dictionary packaged with WordStat. This dictionary includes several dubious coding categories – a "sports" category, for example, includes only the terms aerobics, baseball, boxing, bowling, skating, skiing, soccer, sport, and swimming. In other words, it is missing key sports such as basketball, football, hockey, and many others. Clearly, this dictionary would miss important sports references in messages, making it virtually worthless for serious scientific inquiry. In fairness to the author of WordStat, an otherwise stellar program, this dictionary was created merely to demonstrate the features of the program. But other dictionaries may suffer from similar deficiencies, so users should exert caution before jumping on the premade dictionary bandwagon.

TABLE 7.3. *Option 2: Findings for CATA Preset Coding of "Self as American" Essays – Male-Female Comparisons for Selected General Inquirer Measures*

Dictionary	Males		Females		Total	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Academics	0.66	0.78	1.39	1.69	1.07	1.38
Politics	7.72	2.68	7.67	3.33	7.69	2.97
Religion	0.12	0.23	0.06	0.20	0.09	0.21
Work	1.71	0.87	1.43	1.33	1.55	1.12
Collectivities	2.75	1.79	2.71	1.59	2.73	1.63
Strong	9.82	3.57	8.03	2.73	8.83	3.17
Weak	1.95	1.51	0.50	0.54	1.14	1.28
Emotional	0.52	1.69	0.96	0.53	0.77	0.70

Note: All means are average percentages of words fitting the General Inquirer dictionary. For example, on average, 9.82% of a male respondent essay's words fit the "strong" dictionary. On average, 8.03% of a female respondent essay's words were "strong."

The texts for "typical American" were submitted in a case-delimited fashion to the CATPAC program. A typical "negotiated" process of examining the most frequently occurring words and adding nonmeaningful words to the default "exclusion" list (i.e., those words not included in the analysis) in several passes through the program resulted in a twenty-five-word analysis.[18] Using a seven-word moving window, CATPAC assessed the co-occurrence of these words and provided a dendogram (cluster analysis output, not shown) and multidimensional space coordinates (where proximity in space indicates frequent co-occurrence).

Figures 7.1 and 7.2 show the latter, an example of the type of "emergent" coding possible with CATPAC. The output for the "typical American" text descriptions has been graphed via SPSS's Interactive Graph function. Figure 7.1 displays the typical three-dimensional output; Figure 7.2 presents the first two dimensions only and is given to assist in discriminating among closely placed concepts. We see a clear clustering of the practical considerations of money, time, things, goals, and better, which seems to indicate a coalescence of concepts related to practical and secular aspects of life. On the other side of the space, we find such concepts as religion, culture, good, trying, children, and work. This may indicate a divergence of discourse about the typical American, with practical-oriented concepts as quite separate from other modes of description. Thus, in the future, we may wish to add a measure or two to the human

[18] Forty-seven words were added to the exclusion list, including such words as seen, will, seem, however, down, along, and anything.

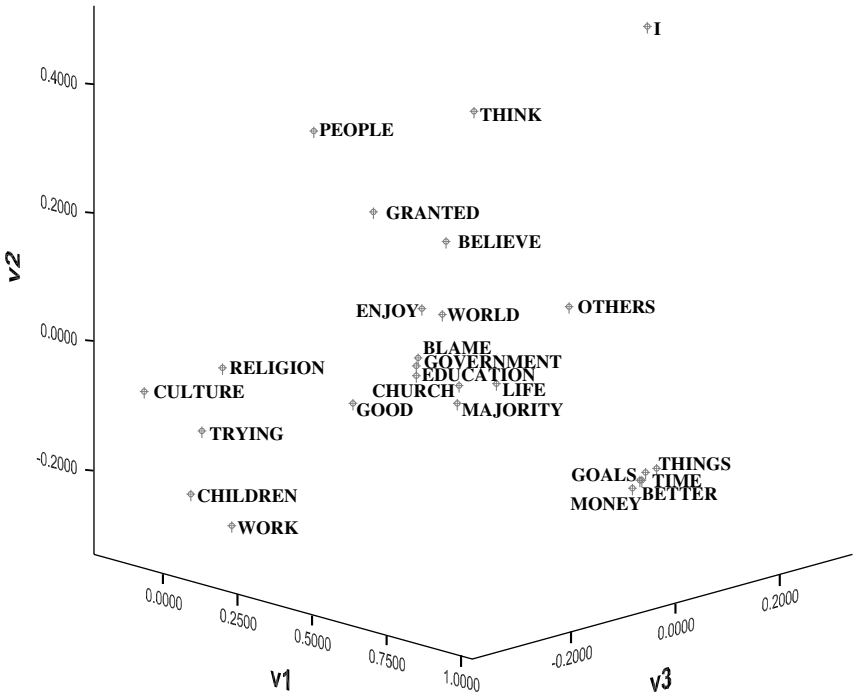

FIGURE 7.1. Option 3: Findings for CATA Emergent Coding of "Typical American" Essays – CATPAC Output in Three Dimensions

coding scheme, or develop appropriate dictionaries for CATA analyses, that look at monetary-based definitions of national identity.

To date, only qualitative comparisons among CATPAC solutions are possible. However, the current work of Hsieh (2004) is aimed at providing a quantitative method of convergence of two or more CATPAC solutions, allowing an empirical basis for pinpointing concepts that differ significantly from one solution to another. Thus, one might assess whether the constellation of discussion about the "self as American" differs from the constellation for the "typical American," for example.[19]

The advantages of emergent coding include the "fresh" look provided by the unrestricted analysis. However, it is not standard procedure to consider emergent coding results as the final outcome in a content analysis but rather as a guide to subsequent, more concerted a priori analyses.

[19] Such discrepancy-congruence analysis might serve as an appropriate measure for those forwarding a conceptual definition of national identity as the degree to which one "fits" with the standard or aggregate national profile.

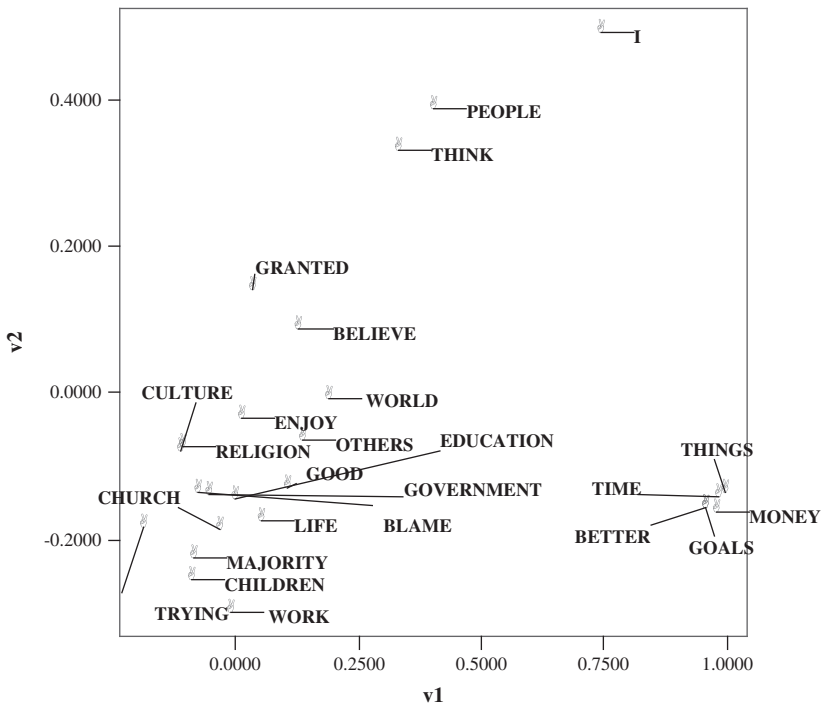

FIGURE 7.2. Option 3: Findings for CATA Emergent Coding of "Typical American" Essays – CATPAC Output in Two Dimensions

Option 4: Human Preset Coding of "Assumed" Identity Messages

The research of Jacques Hymans (2006) into the national identity conceptions (NICs) of world leaders provides a useful example of the application of human coding to "assumed" identity messages, in this case the speeches of selected prime ministers from four nations – France, Australia, Argentina, and India – over a sixty-year period. The work is concerned with leaders' nuclear policy decisions and how NICs might relate to them. Hymans (2006: 2) has conceptualized each leader's NIC as "how the leader understands the natural positioning of the nation with respect to its key comparison other(s) . . . along two basic dimensions . . . of 'solidarity' and 'status.'" He then *assumes* that their major public speeches contain representations of these dimensions.

Hymans first identified the frequently occurring "comparison others" (i.e., any human community not primarily based inside the speaker's national boundaries) for each prime minister.[20] Then, he coded these references in one

[20] Any "other" that was referenced twenty or more times was retained for further analysis.

particular way – whether the comparison other included the speaker's nation (inclusive references – e.g., the United Nations, Europe [for France only], the free world) or not (exclusive references – e.g., Germany, the Palestinians). From this simple content analysis measure, Hymans created two new indicators: (1) a measure of solidarity that was the ratio of the number of exclusive references to the number of inclusive references, ranging from 0 (entirely sportsmanlike) to 1 (entirely oppositional), and (2) a measure of status that was the ratio of paragraphs containing exclusive references only to the number of paragraphs containing exclusive references (with or without inclusive references in the same paragraph), ranging from 0 (entirely subaltern) to 1 (entirely nationalist).

Fruitfully, Hymans then used the prime ministers' ratings on these two dimensions to develop country-by-country typologies that categorized prime ministers as follows:

Sportsmanlike Nationalist	Oppositional Nationalist
Sportsmanlike Subaltern	Oppositional Subaltern

This typology, created from quantitative analyses, was then compared with qualitative, in-depth analyses of the political contexts in which the various prime ministers operated. This analysis is what Neuendorf (2002) would call a "second-order" integrative model of content analysis, combining source data and message data across selected time points.

Hymans (2006) provides an excellent example of how content analysis might be integrated into an overall investigation that also includes a strong qualitative component. This linking of qualitative and quantitative information is to be applauded; as Budge and Hofferbert (1996: 83) point out, "the information we can get is often richest and most revealing when we can put both together." In Hymans's case, the quantitative and qualitative analyses prove to be mutually supportive, each aiding in the interpretation of the other.

Option 5: Computer (CATA) Preset Coding of "Assumed" Identity Messages

Rod Hart's development of the Diction 5.0 program followed a process similar to that of the General Inquirer. As a political communication scholar, he became interested in the objective and reliable tracking of political speech – in the form of debates, speeches, and the like. He first devised a coding scheme for human coders, then later the computer program, measuring forty variables with standardized dictionaries. Over the years, he has expanded the constructs measured by Diction and has added a variety of "comparative" indicators for other types of

communication (thirty-six types, in fact, including such normed categories as campaign speeches, emails and telephone conversations, student essays, poetry, and newspaper editorials). However, the system's forte remains political speech.

For this example, Diction was used to analyze the first State of the Union Address by each of nine recent U.S. presidents, from Dwight D. Eisenhower to William J. Clinton, along with the second State of the Union Address by George W. Bush, which took place shortly after the 9/11 terrorist attacks. These texts were considered examples of assumed or naturally occurring identity messages because of the strong emphasis State of Union Addresses typically place on establishing a national identity. The second George W. Bush speech was chosen to compare past presidential addresses and other political policy speeches to the first State of the Union Address in the current "Age of Terror." The addresses were obtained online from Kenneth Janda's PoliTxt Digital Archive (http://janda.org/politxts/index.html), which includes a complete collection of State of the Union texts. Once the addresses were obtained, analyzing them in Diction was fast and easy, with the program's included dictionaries providing an instant array of results.

Some of the preset Diction dictionaries purport to measure the constructs of accomplishment, communication, (group) centrality, cooperation, and (social) exclusion, all of which relate to conceptualizations of national identity that include interaction, role behaviors, and a collective orientation. In our case, we see in Table 7.4 results from the analysis of George W. Bush's 2002 State of the Union Address compared to other political public policy speeches. The first section of the table illustrates the standard dictionary measures and analyses in Diction, while the second part shows those for the master variables. As the first part of the table shows, Bush's address was above the normal range on the constructs of satisfaction, inspiration, and cooperation, indicating greater emphasis on these than in typical political speeches, which makes sense following 9/11. The cooperation score is particularly interesting from an identity standpoint and could be indicative of a desire to shift U.S. national identity toward greater internal and external cooperation in an effort to more effectively fight terrorism.

The Diction 5.0 CATA program provides five "master variable" indexes that sum multiple dictionaries. The numerical values of these master variables are meaningful only in comparison to the Diction-provided "normal ranges" for various text sets that have been previously analyzed by Hart (2000). Table 7.5 shows the scores for this speech on the five master variables. Only one master variable, optimism, is out of range. This score of 57.22 indicates greater emphasis on optimism than in most political speeches, which again makes sense given the post 9/11 desire to reassure the American people.

TABLE 7.4. *Option 5: Sample CATA Preset Coding via Diction of George W. Bush's 2002 State of the Union Address – Standard Dictionary Totals*

Variable	Frequency	Percentage of Words Analyzed	Normal Range		Standard Score
			Low	High	
Numerical terms	6.42	1.28	0.3	15.04	−0.17
Ambivalence	7.57	1.51	6.49	19.21	−0.83
Self-reference	7.8	1.56	0	15.1	0.1
Tenacity	23.69	4.74	23.32	39.76	−0.95
Leveling terms	7.69	1.54	5.02	12.76	−0.31
Collectives	13.57	2.71	4.04	14.46	0.83
Praise	8.99	1.8	2.77	9.59	0.83
Satisfaction	16.67	3.33	0.47	6.09	4.77
Inspiration	12.01	2.4	1.56	11.1	1.19
Blame	1.59	0.32	0.06	4.16	−0.25
Hardship	9	1.8	1.26	10.48	0.68
Aggression	8.59	1.72	1.07	9.79	0.73
Accomplishment	18.04	3.61	4.96	23.78	0.39
Communication	5.79	1.16	2.21	11.79	−0.25
Cognition	7.01	1.4	4.43	14.27	−0.48
Passivity	6.78	1.36	2.1	8.08	0.56
Spatial terms	14.49	2.9	4.17	19.85	0.32
Familiarity	111.34	22.27	117.87	147.19	−1.45
Temporal terms	18.45	3.69	8.36	21.82	0.5
Present concern	9.33	1.87	7.02	16.6	−0.52
Human interest	31.3	6.26	18.13	45.49	−0.04
Concreteness	27.52	5.5	10.7	28.5	0.89
Past concern	1.8	0.36	0.97	6.19	−0.68
Centrality	3.17	0.63	1.19	7.54	−0.37
Rapport	2.84	0.57	0.42	4.26	0.26
Cooperation	8.73	1.75	0.36	8.44	1.07
Diversity	1.69	0.34	0.07	3.81	−0.14
Exclusion	0.72	0.14	0	4.31	−0.65
Liberation	3.62	0.72	0	4.72	0.57
Denial	3.09	0.62	2.57	10.35	−0.87
Motion	2.27	0.45	0.17	4.35	0

The State of the Union Addresses were also scanned for interesting overall trends. Diction can analyze multiple texts at once and allows for quick interactive switching of results from one text to another. This exploratory technique revealed an interesting and perhaps counterintuitive difference in the use of "collectives" in State of the Union Addresses by Republican and Democratic presidents. Republican presidents (Nixon, Ford, Reagan, Bush I, and Bush II) all used collective language above the normal range, while Democrat presidents

TABLE 7.5. *Option 5: Sample CATA Preset Coding via Diction of George W. Bush's 2002 State of the Union Address – Master Variables*

Master Variable	Score	Normal Range		Out of Range
		Low	High	
Activity	50.1	46.74	55.48	
Optimism	57.22	46.37	52.25	*
Certainty	49.96	46.9	51.96	
Realism	49.75	46.1	52.62	
Commonality	51.18	46.86	52.28	

(Kennedy, Johnson, Carter, and Clinton) all used collectives within the normal range. It should be noted that these analyses were done using the default Diction setting of analyzing only the first 500 words of a text. The program can also be set to analyze longer texts by averaging 500 word "chunks," which is what was done for the analysis of George W. Bush's address reported here. Finally, it should be noted that the data from multiple text Diction analyses, such as a collection of State of the Union speeches, could be imported into statistical programs such as SPSS for further analysis.

The preset coding technique offered by programs such as Diction can be very useful to identity scholars when the program's coded constructs tap relevant identity constructs. Though there are some limitations to preset coding, as mentioned elsewhere in this chapter, this CATA technique benefits from its ability to quickly and reliably elucidate interesting features of texts.

Option 6: Computer (CATA) Emergent Coding of "Assumed" Identity Messages

Some scholars have assumed that news media descriptions of issues constitute important "images" or "identities" of these issues (e.g., Chang 1998); this conceptualization of "collective identity" is one of simple attribution by widely available institutional information sources. For this example, we developed a small original analysis of recent U.S. newspaper coverage of the Northern Ireland political party Sinn Fein, using TextAnalyst software. Recent events have brought renewed attention to Sinn Fein, and the changing color of the discourse about the political entity makes it uniquely suited for emergent content analysis.

For the collection of news stories, we used the LexisNexis database, perhaps the largest single information database in the world. The search was of major U.S. newspapers as defined by LexisNexis (of which there are fifty-three), for all articles with the term "Sinn Fein" in the headline, appearing between January 1

and March 15, 2005. A total of twenty-five articles met the search criteria, and the body of these articles served as the raw material for an exploratory analysis via TextAnalyst. The articles had to be individually stripped of extraneous information (e.g., dateline, name of publication) and saved as an aggregated text file for ready presentation to the CATA application.

TextAnalyst is designed primarily for managing texts, rather than true content analysis, but its neural networking application can provide an interactive look at how news articles discuss or "identify" our targeted topic, Sinn Fein. While not divulging the algorithms used to do so, the program documentation claims that TextAnalyst determines "what concepts – word and word combinations – are most important in the context of the investigated text" (Froelich 2000: 2).

Figure 7.3 provides a screen shot of TextAnalyst results for the Sinn Fein articles. The tree-branching structure we see at the left indicates the strength of

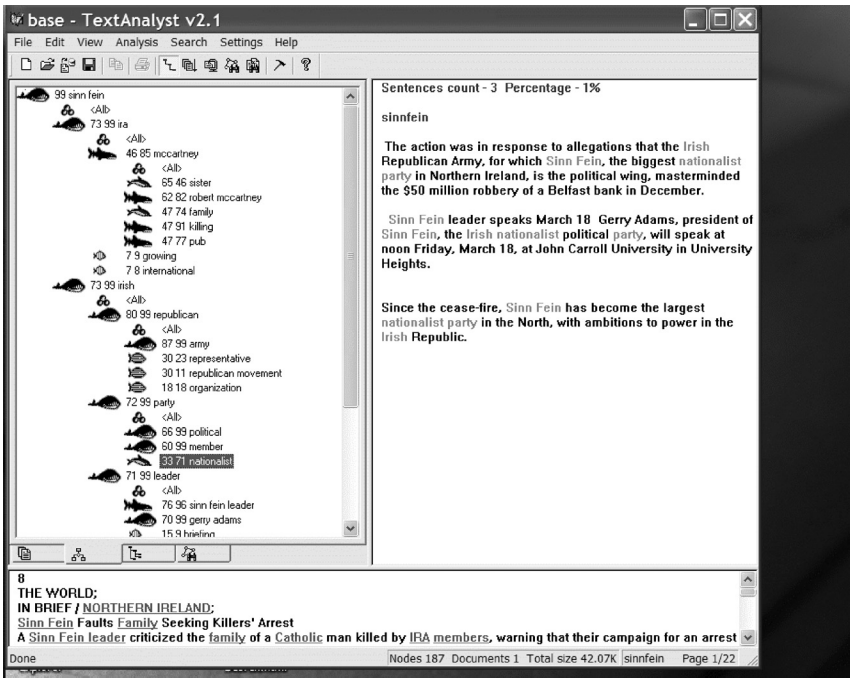

FIGURE 7.3. Option 6: Findings for CATA Emergent Coding of New Stories about Sinn Fein, TextAnalyst Output. *Note:* The TextAnalyst output is presented in three windows: (1) the View Pane (*at left*) provides the branching "topic structure" that is the main focus of our example; (2) the Results Pane (*at right*) displays all sentences in which the selected concept (here, "nationalist," as selected and highlighted in the View Pane) occurs; and (3) the Text Pane (*at bottom*) provides the original text in full length.

co-occurrence of terms. The program has identified "Sinn Fein" as the most central substantive term in the collection of texts – it appears at the top of the outline. It is linked most strongly with the terms "IRA" and "Irish." Each of these terms is shown to be linked with other terms – for example, "IRA" is most strongly linked with "McCartney," and "McCartney" with "sister," "Robert McCartney," "family," "killing," and "pub."[21] The numbers indicate relative "semantic weights" for each concept, ranging from 0 to 100.[22] The first number reflects the strength of the semantic relationship of the concept to the parent concept above (e.g., "McCartney" is related to "IRA" with a strength of 46); the second, the semantic weight of the concept in the entire text (e.g., "McCartney" is related to the text as a whole with a strength of 85). While the numbers do not have an objective meaning beyond their 0–100 range, a number of researchers have developed cutoff values after inspection of their semantic network solutions. For example, Bourret et al. (2006) used a criterion for co-occurrence semantic weights of 50 or 75, depending on the number of nodes in the analysis. Also, J. Adams and Roscigno (2005) have established a cutoff score of 30 in their study of identity, interpretational framing of cause and effect, and political efficacy in the text of white supremacist Web sites.

TextAnalyst has a very interactive interface, and it is difficult to do it justice in this static presentation. For example, if we were able to click on "killing," more terms linked to that word would appear.

This "snapshot" of contemporary communication about Sinn Fein in major U.S. newspapers can provide us with ways of thinking about how the discourse is being shaped in the popular media. Although not truly a content analysis in the traditional sense, it may help us devise a suitable coding scheme or begin to construct an appropriate set of dictionaries.

Option 7: Human Preset Coding of "Extracted" Identity Messages

A study by Eilders and Luter (2000) examined the ways in which media assessed Germany's first participation in military action since World War II – the Kosovo war of 1999. They studied editorials in the top five German newspapers over a three-month period, utilizing a qualitative framing identification

[21] This section of the semantic analysis is clearly referencing a cluster of news coverage of the January 2005 murder of activist Robert McCartney outside a Belfast pub, purportedly by members of the IRA.

[22] These figures may be interpreted as indicators of the probability that the concept is important in the studied text and are determined by the concept's frequency of occurrence, and by its co-occurrence with other concepts. The TextAnalyst Web site indicates that the program, using internal (proprietary) algorithms, determines the semantic weights via a neural networking procedure (www.megaputer.com/tech/wp/tm.php3).

process (Goffman 1974) to locate frames within editorials, finding 364 distinct frames in 190 editorials. A total of 25.8 percent of these framings were deemed to constitute "identity frames" (another 29.1 percent were decided to be "diagnostic framing" and 45.1 percent to be "prognostic framing"). Their conceptualization of an "identity frame" was communication that addresses the national self-image, particularly in this case of German involvement in the Kosovo war; they looked for a motivational framing that provides an answer to "Why should *we* become involved?" Their assumption was that war approval was likely to be based on a "convincing construction of collective identity" (Eilders and Luter 2000: 417). This process is one example of what we might call the "extraction" of identity messages from a larger pool of messages – that is, many messages were screened, and only a subset was deemed to be identity-related and was retained for further analyses regarding identity. Eilders and Luter (2000) then content-analyzed the 94 identity framings as to type, finding 30 percent to fit a category of "Germany in a moral dilemma," 16 percent as "Germany as a loyal NATO member," 15 percent "Germany as part of Europe," 10 percent "Nazi experience calls for defence of human rights," and 29 percent "other." Although the piece did not report reliability figures for the extraction process or the coding of type, the research presents a clear option in the straightforward content analysis of public communications that may contain relevant "identity" messages.

Option 8: Computer (CATA) Preset Coding of "Extracted" Identity Messages

A CATA program that combines elements of preset coding and emergent coding is WordStat, a companion to the statistical package SimStat (www.simstat.com). WordStat allows for the ready creation of researcher-established dictionaries, with some basic example dictionaries provided[23] and synonyms suggested via the program's dictionary builder. And finally, WordStat provides cluster analysis and multidimensional scaling options for the dictionary results. Thus, although the dictionaries are preset by the researcher, their configuration and dimensions of discrimination emerge from the analysis.

For this example, the Schwartz (1992, 1994) cultural values dimensions were used as the basis for the development of a WordStat set of dictionaries that could be applied to autobiographical texts from Project Gutenberg Online.

[23] WordStat's example dictionaries are Appearance, Arts, Communication, Education, Family, Finance, Humor, Nightlife, Outdoor, Sexuality, Spirituality, Sports, and Work. They were designed for research on personal ads. Each includes only a small number of words, ostensibly as "seeds" for further development via the dictionary builder by the researcher.

These autobiographies were specially selected for this section from the vast Project Gutenberg collection based on a scan of each for their relevance to the study of identity, making this an example of "extracted" identity messages. The texts selected were considered to contain insight into the identity of each subject, in part by revealing their cultural values as an important component in their cultural identity. In sum, six complete autobiographical texts were selected for this section, three by famous Caucasian Americans (Benjamin Franklin, Theodore Roosevelt, and "Buffalo Bill" Cody) and three by famous African Americans (Booker T. Washington, Sojourner Truth, and Frederick Douglass). Texts were chosen in this manner so that Caucasian American and African American identities could be compared.

The dictionary development began by taking the ten key Schwartz cultural values and their subdimensions and coming up with synonyms that seemed to appropriately tap each value dimension of interest. To validate these choices, the dictionaries were checked and further refined using the WordStat Dictionary Builder (discussed earlier). Once the ten dictionaries were settled on,[24] the texts were analyzed in WordStat separately for Caucasian and African Americans.

WordStat output includes a full array of analyses of interest to the content analyst, including basic word counts, cross tabulations ("crosstabs") of dictionary categories with other variables of interest (e.g., gender, race), key-words-in-context (KWIC), and the more advanced analyses mentioned at the start of this section.

Table 7.6 shows an example of a crosstab analysis from the Schwartz values–biography text analysis, with each row representing a cultural value and the columns comparing the Caucasian biography texts with those by African

[24] Dictionaries routinely include dozens, possibly hundreds, of search terms, depending on the construct being measured. In our case, we tried to stay close to Schwartz's original measures, limiting the number of terms. For example, Schwartz's dimension of "STIMULATION" includes three questionnaire items to which respondents are instructed to respond using an eight-point scale of how important the cultural value is to them: "An Exciting Life (stimulating experiences)," "A Varied Life (filled with challenge, novelty, and change)," and "Daring (seeking adventure, risk)." Our measure includes the root terms and synonyms, many of which are included as wild cards (e.g., "excit*"), expanding the number significantly (i.e., 12 terms, 9 of which are wild cards). The other Schwartz dimensions resulted in the dictionaries of the following sizes:

STIMULATION: 17 terms, 12 of which are wild cards
HEDONISM: 12 terms, 6 of which are wild cards
ACHIEVEMENT: 14 terms, 10 of which are wild cards
POWER: 16 terms, 7 of which are wild cards
SECURITY: 10 terms, 6 of which are wild cards
CONFORMITY: 13 terms, 11 of which are wild cards
SPIRITUALITY: 10 terms, 7 of which are wild cards
BENEVOLENCE: 13 terms, 11 of which are wild cards
UNIVERSALISM: 14 terms, 11 of which are wild cards

TABLE 7.6. *Option 8: Findings for CATA Preset Coding of Autobiographical Texts Using Researcher-Created Dictionaries of Schwartz's Cultural Values, Applied with WordStat*

	Caucasian Americans	African Americans	Chi-Square	p (2-tailed)
Achievement	8.60%	8.10%	2.63	0.27
Benevolence	24.10%	19.40%	0.72	0.70
Conformity	1.40%	2.00%	8.80	0.01
Hedonism	5.60%	6.90%	16.89	< 0.01
Power	9.40%	12.70%	45.12	< 0.01
Security	11.20%	9.90%	1.17	0.56
Self-direction	8.80%	11.90%	42.30	< 0.01
Spirituality	4.30%	14.90%	315.90	< 0.01
Stimulation	6.20%	3.90%	5.39	0.07
Universalism	20.50%	10.20%	50.63	< 0.01
TOTAL	100.00%	100.00%		

Note: A total of 517,974 words were processed in these analyses, 6,923 of which were dictionary terms. Broken down by race, the Caucasian American texts, which had 364,771 total words, contained 4,598 terms identified as fitting one of the Schwartz dictionaries, and the African American texts, which had 153,203 total words, contained 2,325 dictionary terms.

Americans. The percentages represent the percent of total value mentions in each category (Caucasian or African American) accounted for by each individual value. So, for example, in the African American–generated texts, 19.4 percent of value word occurrences tapped the "benevolence" dimension. This type of analysis can reveal how often different values or other identity terms appear in relation to one another, and how frequently they appear in one set of messages compared to another set.

As an example of the latter, consider the "spirituality" row of Table 7.6. This suggests that the African American authors studied here were more likely to have spirituality as an expressed cultural value, relative to other values, than the Caucasian Americans were, with a percentage difference of 14.9 percent versus 4.3 percent. WordStat also allows statistical testing of differences and the results for this comparison (as assessed though a chi-square test) are indeed significant. Other significant differences shown in Table 7.6 include greater reference to conformity, hedonism, power, and self-direction in the African American–generated texts, and more references to stimulation and universalism in the Caucasian-generated texts.

As discussed throughout this chapter, the overall merit of these types of analyses to identity researchers stems from an assumption that the developed dictionary terms tap into identity conceptualizations of interest. In the present

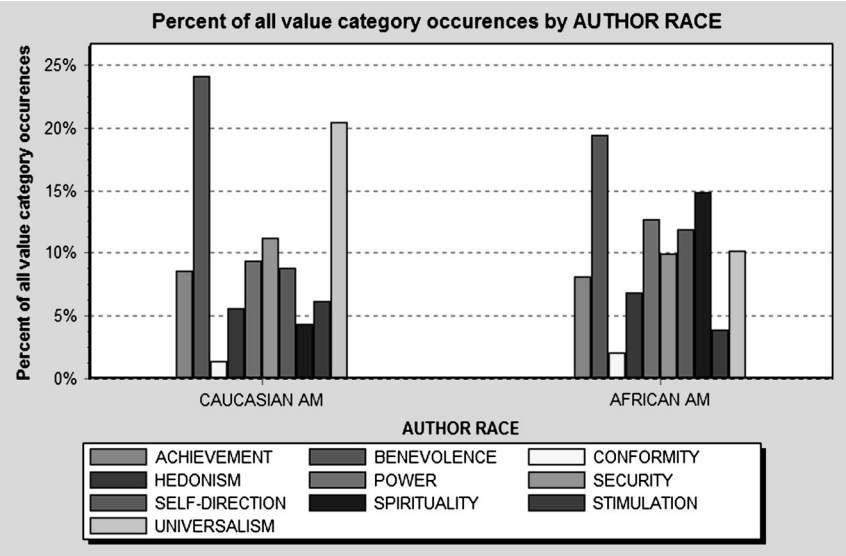

FIGURE 7.4. Option 8: Findings for CATA Preset Coding of Autobiographical Texts Using Researcher-Created Dictionaries of Schwartz's Cultural Values, Applied with WordStat

example, one might view the Schwartz dictionary categories as indicators of different types of collective (cultural) identity, and the biography texts as messages containing information about collective identity.

DEVELOPING CONTENT ANALYSIS MEASURES FROM EXISTING SURVEY OR OTHER MEASURES OF COLLECTIVE IDENTITY

Our examples thus far have focused on national and cultural identities, as most relevant to political discourse. However, other collective identities might certainly be assessed via content analysis. No standard coding schemes have been devised to date. One tack that can prove fruitful is to adapt standard measures of collective identity for content analysis purposes, as shown earlier in option 8 (using an adaptation of Schwartz's survey instrument for cultural values) and as exemplified by many of the General Inquirer indexes.

Other existing models and measures of identity might serve as the base for content analysis measures. For example:

1. Stephan and Stephan's (2000) roster of self-identification measures for racial and ethnic identity could serve as the basis for dimensions of content analysis indicators.

2. In Chapter 2 of this volume, Brady and Kaplan provide a wealth of conceptualizations *and* possible measures for ethnic identity.
3. Phinney's 14-item MEIM (Multigroup Ethnic Identity Measure) has three dimensions that might be informative – ethnic identity achievement, affirmation and belonging, and ethnic behaviors (Richard M. Lee and Yoo 2004) – as well as individual survey items that might be adaptable to content analysis applications (e.g., "pride in ethnic groups," "participation in cultural practices").
4. In addition to the work in cultural values by Schwartz (1992, 1994), the substantial body of work on cultural and national values by Triandis (1994), Inglehart and Baker (2000), and Kabanoff and Nesbit (1997) would be useful starting points (see also the World Values Survey 2004).
5. Hoffman's (2001) summary of the history of the measurement of masculinity and femininity gives us many ideas as to how gender identity might be measured, from the early Bem Sex-Role Inventory to the recently developed Hoffman Gender Scale; we might also examine the Worthington et al. (2002) model of heterosexual identity development at the individual and social levels.
6. Puddifoot's (1995) dimensions of community identity are concrete enough to provide easy guidance for the construction of content analysis measures – for example, residents' perceptions of community boundaries; evaluations of quality of community life including friendliness, sense of mutuality, and cooperativeness; and evaluations of community functioning including leisure services, health services, and material quality of life.
7. Postme's (2003) approach to measuring social identity in organizations provides many clues to original measures, including nonverbal indicators such as clothing and the public expression of group goals.
8. Marcia's (1993) handbook for the study of ego identity provides a raft of open-ended interview measures related to understanding of self and various social roles one plays.
9. In Chapter 12 of this volume, McDermott reviews the conceptualization and measurement of identity from a psychological perspective.

LIMITATIONS AND POTENTIALITIES OF CONTENT ANALYSIS AS A TOOL TO MEASURE IDENTITY

Content analysis is dependent on a clear conceptualization of the construct(s) under examination. The examples presented here cover a wide range of implicit

conceptual definitions – from national identity as a collection of ascriptive characteristics (e.g., social categories) to expressed emotional or affective orientations (e.g., patriotism, ethics, worldview) to general life outlooks (e.g., goals and gratifications, opportunity) to structural states (e.g., rights, freedoms). The utilization of content analysis for measuring identity makes the assumption that we take a pragmatic approach to the study of communication and identity – that is, that we are willing to conceptualize identity as something that is constituted from communication behavior.

For the sample results presented for options 1–3, it must be recalled that the dataset – that is, the set of verbal identity descriptions – is small and from a nonprobability sample. Therefore, firm conclusions should not be reached from the findings. Rather, the results are presented as options for the types of analyses that might be done with content analysis. There may be additional debate concerning the assumption or conceptualization that "identity" may be reflected in responses to prompts. However, open-ended interviewing and surveying are commonly used to tap ethnic identity, cultural identity, and acculturation (Arends-Tóth and van de Vijver 2004; Berry 1980; Chun, Organista, and Marin 2003; Stephan and Stephan 2000).

Other novel analyses presented as examples here are also limited in their generalizability and statistical power. However, we believe that these examples present a full range of options for the identity researcher, and we have placed them in a model that encourages comparisons among options for content analysis execution (i.e., CATA vs. human coding, preset vs. emergent coding) and for types of communications to analyze (i.e., responses to prompts, naturally occurring communication).

CATA procedures offer easy reliability and standardization. However, CATA techniques are what we might call "knee-jerk"–they blindly count dictionaried words, regardless of context, negation, or, in general, ambiguation.[25] Most important, a program-provided dictionary may not match a definition of the construct it purports to measure. The development of identity-focused dictionaries is certainly possible but will require a long and careful process.

When compared with CATA, human coding is amenable to much more in-depth, nuanced measures, and is more flexible. With human coding, our measures will be sensitive to various grammatical operations (e.g., negation, as in "I do not feel an affinity for flag-waving Americans"). And we may attempt measures of content and contestation. But it is held to the high

[25] Diction 5.0 does attempt disambiguation, assigning fractional words to dictionaries on the basis of common usage. For example, the word "well" might be divided among positive mentions (e.g., "I don't feel well"), physical objects (e.g., "He drew water from the well"), and nonfluencies (e.g., "Well . . ."), among other usages.

standard of intercoder reliability. If we fail to develop measures – a coding scheme and a training procedure – that can be reliably applied by nonexpert coders, all is for naught. The human-coded examples presented here do not emphasize intercoder reliability to the extent that we would like because of the pilot nature of some examples and the limited reporting by other researchers for other examples. Any full reportage of a human-coded content analysis must include variable-by-variable reporting of appropriate reliability coefficients (see note 5).

Abdelal et al. (2006) introduce the notion of identity as content and contestation. The various human and CATA coding examples in this chapter present a wide array of content types, and the frequency of occurrence of each type has been assessed in various ways.

Contestation proved to be problematic as a dimension for human coding in option 1, occurring very rarely and thus revealing that the sample respondents tended not to *discuss* identity in terms of its contested nature. However, contestation may perhaps be more directly assessed by the *variances* of the measures. For example, the General Inquirer measure "strong" showed moderate variance across respondent essays for the description of the typical American (with a standard deviation/mean ratio of .51). But, the General Inquirer measure "economic" showed greater variance across respondents (with a ratio of 1.14). We might interpret this difference to indicate greater contestation with regard to whether the typical American is defined in economic terms rather than whether the typical American is defined by strength.

The procedures outlined here have focused on the analysis of verbal (i.e., word-based) communication. It should be remembered that content analysis may also include an analysis of nonverbal and pictorial variables. Indeed, Watzlawick, Bavelas, and Jackson's (1967) pragmatic approach to communication would demand such inclusion. Once again, the conceptualization of the construct of collective identity must set the stage. If one were to subscribe to a conceptual definition of identity as including nonverbal behaviors such as (literally) flag waving or choice of clothing (Oshiba 2002), then measures of such behaviors would be quite properly executed via content analysis.

The example procedures presented here are intended to initiate a dialog among researchers about the viability of using standard and novel content analysis techniques for the measurement of collective identity. The conceptual definition is the vital origin for every content analysis, and the arena of collective identity seems marked by significant definitional contestation that must be further explored if standardized measurement techniques are to be developed for the study of collective identity via content analysis. Abdelal et al. (2006) have begun the process of definitional clarification. In the meantime, the palette of

options provided by human and computer content analysis procedures offers much to the identity researcher.

APPENDIX: NATIONAL IDENTITY PILOT CODING SCHEME FOR HUMAN CODING

National Identity Pilot Coding Scheme Codebook (Coding Instructions)
Unit of data collection: Each essay/open response will serve as the unit of data collection. For each essay/response, fill out one coding form.

I. General Information

1. Coder#: Record the coder ID #
   1 K.N.
   2 R.O.
2. Essay #: Record the essay # found above each essay, which refers to the *referent* for that essay, as follows:
   1 Self as American
   2 Typical American
3. ID#: Record the respondent ID # found at the top of the essay set
4. Number of sentences: Count and record the number of sentences contained in the essay, as delimited by standard punctuation (i.e., period, question mark), whether or not the sentence is grammatically complete.

II. Macro variables

For variables 5–7, record the number of sentences that contain at least some reference to each of the following. The term ''referent'' means :

5. Self-identification: This type of statement provides some identifying information about the referent (see #2 above), phrased in the first person (e.g., ''I am a religious person,'' ''I believe that America is headed in the wrong direction'').
6. Role identification: This type of statement provides some identifying information about the referent relevant to social roles that the referent performs (or is constituted of, for referent #5, America) (e.g., ''America is the protector of Third World countries,'' ''I'm close to my two sisters'').

7. Collective identification: This type of statement provides some identifying information about the referent relevant to an in-group, or out-group, status (e.g., "I'm typical of most Americans in that I enjoy sports," "Americans are much ruder than other nationalities").

NOTE: Items 5–7 are *not* mutually exclusive; that is, a sentence may be counted as more than one type.

III. Constructs from Stone (1997)

For variables 8–11, (a) record the number of sentences that contain at least some reference to each of the following. Then, (b) record the number of those sentences that contain an explicit verbal negation for the target topic after the "N" (e.g., "I do not feel a responsibility . . .).
NOTE: For each measure, the number for (b) is a subset of the number recorded for (a).

8. Goals and gratifications – references to hopes, intentions, objectives or motives for acting, and/or recompense, reward, or type of satisfaction sought by the actor
9. Rules and responsibilities – references to rules, guidelines, laws, norms of behavior, and/or the acknowledgment of responsibility, answerability, duty, or obligation toward other individuals or toward a collective
10. Feelings and emotions – references to internal and subjective affective states (e.g., anger, joy, sorrow, pity, passion, love, grief, etc.)
11. Unique outlook/ways of understanding – references to any cognitive activities (beliefs, knowledge, etc.) and/or perspectives that are unique, one-of-a-kind, and/or distinctive

IV. Additional Constructs from Ashmore, Deaux, and McLaughlin (2004)

NOTE: Other Ashmore et al. aspects are covered by dimensions above (e.g., Self Identification, Collective Identification). Once again, the same coding procedure as for 8–11 is to be used.

12. Positive evaluation – contains reference to a positive outlook or favorable judgment
13. Negative evaluation – contains reference to a negative outlook or unfavorable judgment

14. Ideology – contains reference to beliefs about a group's experience, history, and position in society
15. Behavioral involvement – contains reference to *actions* that directly implicate the collective identity category in question

V. Additional Constructs Derived from the Data

Again, the same coding procedure should be used; variables 16–17 are exceptional, not allowing for a negation tally.

16. Physical characteristics – contains reference to observable human physical characteristics, including hair color and type, and skin tone (but NOT race/ethnicity)
17. Ascriptive social categories – contains reference to such demographic and social categories information as race, ethnicity, and gender. Rights–contains reference to rights of the individual or group
18. Freedom – contains reference to freedoms of the individual or group
19. Power – contains reference to power status or relationship of the individual or group
20. Opportunity – contains reference to positive opportunities available to the individual or group
21. Patriotism – contains reference to an aspect of pride in country as expressed at the individual or group level
22. Worldview – contains reference to a recognition of or comparison to groups/nations/individuals outside of the referent nation (i.e., America)

National Identity Pilot Coding Scheme 2005
Coding Form

1. Coder#:_____
2. Item #: _____
3. ID#:_____
4. Number of sentences: _____

Macro variables: Number of sentences that include:

5. Self-identification _____
6. Role identification _____
7. Collective identification _____

Stone (1997):

8. Goals and gratifications _____ (N_____)
9. Rules and responsibilities _____ (N_____)
10. Feelings and emotions _____ (N_____)
11. Unique outlook/ways of understanding _____ (N_____)

Ashmore et al. (2004):

12. Positive _____ (N_____)
13. Negative _____ (N_____)
14. Ideology _____ (N_____)
15. Behavioral involvement _____ (N_____)

From the data:

16. Physical characteristics _____
17. Ascriptive social categories _____
18. Rights _____ (N_____)
19. Freedom _____ (N_____)
20. Power _____ (N_____)
21. Opportunity _____ (N_____)
22. Patriotism _____ (N_____)
23. World view _____ (N_____)