This chapter deals with the content coding of written or verbal responses to open-ended questions. Responses to open-ended questions can range from the completion of sentences to extended narratives. The content may be derived from the respondent's imagination, opinions on various topics, or autobiographical memory. The amount of and variability in the material generated about thoughts, feelings, and behavior in response to an open-ended question can be staggering. Yet this material may arguably be the richest information available to personality psychologists. As more and more researchers turn to these techniques to investigate complex personality processes, it is important for the field to establish a set of guidelines for conducting content coding of open-ended responses. This chapter attempts to bring to light the unique advantages and challenges of this method. It also seeks to stimulate the reader's imagination about the possible uses of content coding in research. Moreover, it warns against treating data yielded by content scoring techniques as though they were data derived from questionnaires and argues for a specific set of rules governing issues of reliability and validity of content-coded data that can maintain both methodological rigor and the richness of the material. Thus, the chapter presents (1) information on the history and development of open-ended scoring techniques, followed by (2) a review of the basic types of scoring systems, (3) a step-by-step guide to using these scoring techniques, and (4) a case example of developing a hypothesis-driven content scoring system. The final section discusses briefly the complexities involved in determining reliability and validity of data collected with open-ended scoring techniques.

**History and Development of Open-Ended Scoring Methods**

Throughout the history of psychology, open-ended-style questions have been a common
choice for researchers interested in complex, subjective experiences, including intentions, patterns of reasoning, and attempts to find meaning in personal experiences. In the late 19th century and early 20th century, the most influential psychologists of the day, including Galton, William James, Wundt, Freud, G. Stanley Hall, and Piaget, made extensive use of narrative accounts of personal experiences in their early explorations of the inner workings of the mind (Smith, 2000). These researchers were concerned with the broad issues of how people think, reason, and make use of their experiences to understand themselves and the world around them. Even as the behaviorists rejected the subjective, introspective, and mentalistic aspects of the person and limited their psychological investigations to observable behavior, some personologically and clinically oriented psychologists, such as Murray and Allport, continued to explore the inner life of the person using a variety of open-ended techniques.

When interest in cognition and affect returned to mainstream psychology during the 1950s and 1960s, a foundation had been laid to explore these topics with increased experimental rigor (Smith, 2000). For some researchers this meant developing reliable measures in the form of fixed-response instruments such as inventories and questionnaires. Some investigators found that these methods did not capture the complexity of human thought, feelings, and behavior, and they turned to content and narrative analysis as an alternative that provided less structure and greater insight into the richness of the subjective experience of the individual. The research programs that attempted to assess personality motives via storytelling techniques, such as the work on the achievement motive (McClelland, Atkinson, Clark, & Lowell, 1958), put forth the notion that indirect, open-ended measures of personality can be highly revealing of a person’s preferences and enduring concerns (see Schultheiss & Pang, Chapter 19, this volume).

Since the mid-1980s, there has been a burgeoning of interest in open-ended responses that involve content-analytic techniques. Notably, McAdams (1985) put forth a model of identity as being rooted in one’s personal life story. According to this view, an individual’s life story is an internalized and evolving narrative of the self that serves to integrate disparate roles and to bring together the reconstructed past, perceived present, and anticipated future. Moreover, the groundbreaking work of Pennebaker (1997; Pennebaker, Mehl, & Niederhofer, 2003) demonstrated the importance of narratives in making meaning of traumatic and difficult personal experiences. Thus, personal narratives have been shown to provide a person’s life with some degree of unity, purpose, and meaning.

Interest in open-ended responses reached a new high when Smith (1992) devoted an entire book to issues of thematic content analysis that included 14 different scoring systems for open-ended responses for motives, attributions and cognitive orientations, and psychosocial orientations. The topics investigated by the various content scoring systems in the book included personality change and development, stress and coping, goal setting, physical health and illness, and societal trends. This book provided practical information on using content-analytic methods and helped to make the techniques more accessible to researchers interested in open-ended responses.

Nowadays, it is quite common for personality psychologists interested in complex psychological processes to employ open-ended response techniques. Many researchers have found that by asking people to write or to express orally their thoughts and reactions, it is possible to discover unique patterns of expression that would be extremely difficult to capture in the form of fixed-response questionnaires. Thus, despite its labor-intensiveness and complications involving reliability and validity (discussed below), many researchers find content coding of open-ended responses to be an invaluable tool.

A unique, and perhaps most important aspect of open-ended responses is that they offer individuals freedom of expression. Some personality psychologists have argued that individual differences are most easily detected in situations that are loosely structured so that the person cannot rely on external social cues on how to act, think, and feel. Open-ended techniques can create such ambiguity. Participants may find that the open-ended format gives them more freedom to explore their thoughts, feelings, and reactions in ways that they had not done previously. So the qualitative material generated from open-ended questions may reveal innermost thoughts, frames of reference, emotional reactions, and cultural assumptions that may or may not be accessible by other methods.
In contrast, when developing a fixed-response format, researchers use their expertise to come up with typically four or fewer alternatives from which the participant must select one that best represents his or her thoughts, feelings, and reactions to the question. The procedure obviously limits the participants’ choices and restricts their options to what the researcher views as most probable or relevant. Participants may endorse responses that do not at all represent their true reactions—even if they are the most representative ones available. It may be that a question is simply irrelevant to the individual. For example, if a researcher is interested in the characteristic of “zaniness” and develops a zaniness questionnaire, it may have utility only with people who have the characteristic of zaniness and offer no predictive utility for those who lack this characteristic (cf. Baumeister & Tice, 1988).

Open-ended responses are more likely to be personally relevant and important to the individual because they are generated by the individual—often spontaneously. Because responses to open-ended questions are generated without a great deal of prompting, they may be more likely to reveal the individual’s most candid and least self-critical thoughts, feelings, and reactions. For this reason, response biases such as social desirability effects, which often plague questionnaire methods (see Paulhus & Vazire, Chapter 13, this volume), should be of less concern for responses to open-ended questions. The content of open-ended responses may reveal more about people than they want to disclose, or even more than they know about themselves.

Another related advantage of open-ended responses is that they can often bring to light issues that the researcher did not think to inquire about directly. There may be aspects of the phenomenon in question that are extremely important to the participant but go undetected owing to preconceived ideas or lack of awareness of the researcher. For instance, in a pilot study that my colleagues and I conducted (Wang, 2004) on how young women balance their commitments to work and relationships, we asked fixed-response questions about the number of mentors and how much social support the participants received from the mentors. We found that the participants generally reported few mentors and very little social support. It was later discovered via the responses to open-ended items, such as “Describe someone in your life who is helpful to you,” that the words mentor and social support had very negative connotations. For the young women in the sample, these words implied subordinate relationships and lack of ability to cope. By asking them open-ended questions about helpful people in their lives, the participants were able and quite willing to express their experiences with mentors and social support, but in their own words. If we had not asked the questions in an open-ended way, we would never have known about these supportive relationships that were central to the study. Open-ended responses may help lessen unknown researcher biases. Thus, the open-ended response method offers a unique set of advantages and challenges.

Types of Open-Ended Responses

It is beyond the scope of this chapter to inventory every established content scoring system created to analyze open-ended response data (see Smith, 1992, 2000, for a more extensive list). Three major types of open-ended responses are discussed here: (1) sentence completion, (2) verbal material from essays or archives, and (3) personal narratives, stories, and diaries. It is intended that the examples of each type be given in enough detail for the reader to appreciate the breadth of content that can be studied with these methods.

Sentence Completion Tests

Sentence completion tests generally involve providing participants with a set of sentence stems each from which they must complete to form a complete thought. Participants are usually instructed to respond quickly or to go with their initial reactions to the stem. Sentence completion tests are often employed to measure motivation, as well as to assess moral and cognitive reasoning. The idea is to obtain a sample of the participants’ thoughts, reasoned opinions, and/or preoccupations in response to a minimal prompt, such as “I feel good when . . .,” “It is important to . . .,” “I am trying to . . .,” and “Most people are . . .”

Three examples of sentence completion scoring systems are discussed below. The first two coding systems are excellent examples of the operationalization of complex psychological
that the words very negative women in the inordinate rela-
cope. By asking out helpful peo-
were able and expe-
ions with at in their own he questions in lld never have e relationships
Open-ended re-
researcher response method
ages and char-
characteristics derived from well-known theo-
ories of personality development. The final ex-
ample is a very successful attempt to measure and to classify personal goals. In examining the responses to the sentence stems, one often finds that no two responses are alike. Yet these clas-
sification systems can reliably identify the com-
mon characteristics of complex psychological
processes. Numerous studies have validated the
utility of these tests.

The first example of a sentence completion
test measures human motivation based on
Maslow’s (1954) hierarchy of needs. Aronoff
(Aronoff & Wilson, 1985) developed the
sentence completion test to measure three of the
needs from Maslow’s well-known hierarchy.
On the basis of his knowledge of Maslow’s the-
ory, Aronoff developed scoring categories that
were articulated in enough detail for indepen-
dent coders to reliability identify the responses.
For instance, sentence completions such as “I
feel good when I have enough money” and
“Most people are rude” demonstrate a desire
for safety, whereas responses like “Most people
are potential friends” and “I feel good when I
am loved” indicate a need for affiliation. Re-
sponses like “I feel good when I am doing my
best” and “It is important to make a lasting
contribution” represent the need for esteem.

Individuals complete a series of sentence stems
that are then scored for these motives by inde-
pendent judges. Aronoff conducted a series of
studies both cross-culturally and with college
students in the United States that offered convi-
cing evidence for the validity of the scoring
system. Research on developmental anteced-
ents and group behavior validated the system’s
power in distinguishing individuals concerned
with safety and esteem needs (Aronoff & Wil-
son, 1985).

Loevinger’s model of ego development
(1976) has been especially influential in per-
sonality psychology because it allows the ego
stages to be measured via a standardized sen-
tence completion test, called the Washington
University Sentence Completion Test for Ego
Development (WUSCTED; Loevinger &
Wessler, 1970). The work is theoretically
grounded in a cognitive-developmental para-
digm, influenced by Piaget and concepts from
the psychoanalytic tradition. The WUSCTED
is used to study how people make sense of and
synthesize experiences as they move across the
lifespan. As in Maslow’s theory, development is
viewed as a progression through hierarchical

stages. Earlier stages must be mastered before
subsequent stages can be approached. In gen-
eral, moving from lower to higher stages, one
becomes less governed by immediate impulses
and more flexible, operating according to inter-
nalized standards of conduct. Interpersonally,
the person moves from egocentrism through
conformity to relative autonomy and mutual
interdependence. The WUSCTED is composed of
a series of sentence stems, such as “The thing
I like about myself . . .” and “I am . . .” For
each stem, the person writes an ending to the
sentence. Each response is classified into one of
Loevinger’s stages, ranging from impulsive to
integrated, according to carefully designed
scoring manuals. The scores are then tabulated,
and a final ego-state score is derived according
to a numerical formula. The test has generated
a considerable amount of research that, to
a great extent, has shown the validity of this clas-
sification system. For instance, researchers
interested in personality development over time
have used this system to develop a tripartite
typology of personality characteristics and
have demonstrated coherent patterns over time
(e.g., John, Pals, & Westenberg, 1998; York

A sentence completion format can also be
used to classify the content of personal goals.
Notably, Emmons (1986, 1992) developed the
concept of personal strivings, defined as “char-
acteristic, recurring goals that a person is trying
to accomplish” (Emmons, 1992, p. 292). The
participant is asked to list 10–15 different per-
sonal strivings, or what he or she is “currently
trying to do.” Examples used to complete the
stem “Trying to:” include “avoid anything that
upsets me,” “be a kinder person,” “attain posi-
tions of leadership,” and “not eat between
meals.” Emmons (1992) developed a 12-cate-
gory coding system to capture the idiographic
nature of these strivings and demonstrated that
they are related to other important motiva-
tional constructs, as well as quality-of-life mea-
sures. For example, strivings related to inti-
macy, generativity, and spirituality are corre-
related with well-being, whereas avoidance
and power strivings are negatively correlated
with life satisfaction (Emmons, 1992). It has
also been possible to examine the interrela-
tionship between and among different strivings
(Emmons & King, 1988) and how these goals
may be related to different levels of motivation
that can predict action (Emmons & McAdams,
Paragraphs from Essays, Archives, and Other Verbal Material

Content-analytic methods have also been used to identify patterns of content and structure in essays written in response to specific questions, as well as in archival material such as letters, diaries, and speeches, which may provide a rich source of information about a person's inner life. Archives may be the only source of information about persons who are deceased, unavailable, or uncooperative. Written or oral records may provide the only feasible way of studying individuals in depth, and at the same time take into account historical periods, cohort, and social-cultural influences on the individual (see Cramer, Chapter 7, this volume, for a more detailed discussion).

The length of such material can range from a paragraph to many pages of text. Therefore, to analyze this type of material, the researcher must first decide on a meaningful unit of analysis. This may be the whole response, regardless of length, or the unit may be arbitrarily decided by using some factor, such as the number of words or paragraphs, that is unrelated to the content of the material. Another strategy is to develop a criterion for meaningful passages that lend themselves to further analysis and discard the other material that does not meet the criterion. Once the unit of analysis is identified, a scoring system may be created to quantify the characteristics of interest. Below are two examples of such scoring systems.

One well-known scoring system used with lengthy material of this sort is the Content Analysis of Verbatim Explanations (CAVE; Peterson, Schulman, Castellon, & Seligman, 1992), which measures the explanatory style that individuals use to make sense of events. The CAVE technique involves extracting casual explanations from verbal material and then rating these explanations on 7-point scales according to their internality, stability, and globality. The system was developed to approximate the Attributional Style Questionnaire (ASQ; Peterson et al., 1982), and make it possible to research "all manner of interesting subjects, including those inaccessible with the ASQ—the quick, dead, famous, belligerent, sensitive or remote" (Peterson et al., 1992, p. 380). The CAVE technique has been used reliably by independent judges unaware of the research hypotheses. There is a good deal of validation data supporting the CAVE technique as a measure of explanatory style. Studies show that highly internal, stable, and global causes preceded increased depression and that highly external, unstable, and specific causes preceded decreased depression in psychotherapy transcripts coded with the use of CAVE. Peterson, Seligman, and Vaillant (1988) used the CAVE technique to analyze data from a 33-year longitudinal study of the psychological precursors of physical illness. They found that a pessimistic explanatory style found in the written responses around age 25 predicted poor health from ages 45 through 60, even when initial mental and physical health were held constant, thereby lending a good deal of support for the powerful influence of explanatory style.

A second scoring system used extensively to analyze text is the Integrative Complexity Coding Manual (Baker-Brown et al., 1992). This system has been developed from successive versions of theory that focus on the complexity of information processing and decision making. Complexity includes two components: differentiation and integration. Differentiation refers to the perception of different dimensions within a stimulus domain, and to the taking of different perspectives when considering the domain. In this scoring system, differentiation is a necessary but not sufficient prerequisite for integration, which is the development of conceptual connections between differentiated dimensions or perspectives. Examples of such connections include references to tradeoffs between alternatives, a synthesis between them, and a reference to a higher-order concept that subsumes them. Integrative complexity is scored on a 1–7 scale, where 1 indicates no evidence of either differentiation or integration, 3 indicates moderate to high differentiation but no integration, 5 indicates moderate to high differentiation and moderate integration, and 7 indicates both high differentiation and high integration. A great deal of research has demonstrated the utility of the system. Nonarchival research has been directed toward the general topics of social perception, attitude and attitude change, and attribution; preparation of speeches, organization problems, attitudes concerning social policy decisions, and moral dilemmas. Archival work has addressed the prediction of international crises and their outcomes, the effects of social and political roles, the succession and duration of leader careers, and the relation between political ideology and political climate (Suedfeld, Tetlock, &
Personal Narratives, Stories, and Diaries

In recent years, personality psychologists have become increasingly interested in the stories people tell about their own lives (e.g., Pillemer, 1998; Singer & Salovey, 1993; Thorne, 2000). Autobiographical essays and stories may hold special significance because they provide valuable information about a person's self-perceptions, including self-identity, life scripts, and beliefs in how the world works pertaining to the person's destiny. Moreover, research suggests that personal narratives promote healing from personal traumas and other difficult life experiences and can aid in integrating life experiences into a stable sense of identity (McAdams, 1985). The telling of personal stories allows individuals to make sense of difficult or discordant events and to integrate these experiences with the self.

Most notably, the work of Pennebaker (1997) suggests that translating personal traumas into words has long-term health benefits. This basic writing paradigm, used in a large number of these studies, involves having participants write about assigned topics for 3–5 consecutive days, 15–30 minutes each day. The writing is generally done in the laboratory with no feedback given. Participants assigned to the control condition are typically asked to write about superficial topics, whereas participants in the experimental condition are asked to “[Write about the] very deepest thoughts and feelings about an extremely important emotional issue that has affected you and your life” (Pennebaker, 1997, p. 162). Common themes found in these emotional essays include lost loves, deaths, incidents of sexual and physical abuse, and tragic failures. In order to study the cognitive changes associated with writing about traumatic experiences, Pennebaker and Francis (1996) created the Linguistic Inquiry and Word Count (LIWC), a computer program that analyzes essays in text format. The LIWC was developed by having judges evaluate the degree to which more than 2,000 words were related to several different categories. These categories included negative emotion words (sad, angry), positive emotion words (happy, laugh), causal words (because, reason), and insight words (understand, realize). The LIWC program allows for the quick computation of the percentage of the total words that are represented in each of these categories.

Three linguistic factors have been found to reliably predict improved physical health. First, the more positive words people used, the better their subsequent health. Second, moderate use of negative emotion words was associated with better health, whereas the use of very high or very low numbers of negative emotion words correlated with poorer health. And perhaps most important, an increase in both causal and insight words over the course of the writing was strongly associated with improved health (Pennebaker, Mayne, & Francis, 1997). That is, people who benefited most from writing began with poorly organized descriptions and progressed to a coherent story by the end of the writing assignment.

The life story model of personality is another highly influential narrative approach (McAdams, 1985). According to this model, identity itself is the life story that an individual begins constructing in late adolescence and young adulthood. McAdams created a life story interview in which participants divide their lives into chapters and provide a plot summary for each. They are then asked to describe eight key episodes (e.g., low and high points, turning point, earliest memory). Participants then identify and describe their greatest life challenges, the most influential people in their lives as the main characters, future plot, personal ideology, and life theme.

Obviously, much information is gleaned from these extensive life narrative interviews. McAdams has developed two different scoring techniques to capture the variability and common themes among these highly idiographic accounts. First, by examining the thematic lines of the stories, it was found that many characters strive for love or power or both. Love and power reflect Bakan’s (1966) themes of agency and communion, which he argues are the two fundamental modalities of experience. Agency refers to the individual’s efforts to expand, assert, perfect, and protect the self, to separate
the self from others, and to master the environment. Communion refers to the individual’s efforts to merge with other individuals, to join together with others in bonds of love, intimacy, friendship, and community. McAdams, Hoffman, Mansfield, and Day (1996) developed and validated a system that breaks down agency and communion into four main subthemes that can be reliably coded in narrative accounts of key episodes in a life story, such as high points or turning points. For agency, life story episodes can be reliably coded for the subthemes of self-mastery, status/victory, achievement/responsibility, and empowerment. For communion, life story episodes can be reliably coded for subthemes of friendship/love, dialogue, care/help, and unity/togetherness. Research suggests that these themes are consistently related to people’s motivational tendencies (McAdams et al., 1996; Woike, 1994, 1995; Woike, Gershkovich, Piorkowski, & Polo, 1999; Woike & Polo, 2001).

Second, McAdams’ more recent work looks at the life stories of midlife adults. A general theme from these life narratives is the commitment story, similar to Tomkins’s (1987) notion of a commitment script. In the commitment story, the protagonist (1) enjoys an early family blessing or advantage, (2) is sensitized to the suffering of others at an early age, (3) is guided by a clear and compelling personal ideology that remains relatively stable over time, (4) transforms or redeems bad events into good outcomes, and (5) sets goals for the future to benefit society and its institutions. Commitment life stories are common among adults committed to promoting the well-being of youth and the next generation, or what Erikson referred to as generativity. For both midlife adults and students, the tendency to construct life stories in which bad events are ultimately redeemed into good outcomes is associated with greater psychological well-being (McAdams & de St. Aubin, 1992; McAdams, de St. Aubin, & Logan, 1993).

Many of the aforementioned scoring techniques can be applied to smaller snippets of data, such as short diary entries and data gathered from experience sampling techniques (see Conner, Feldman Barrett, Tugade, & Tennen, Chapter 5, this volume). For instance, Mehl, Pennebaker, Crow, Dabbs, and Price (2001) developed the Electronically Activated Recorder (EAR), a device for sampling naturalistic daily activities and conversations. The EAR tape-records for 30 seconds once and every 12 minutes for 2–4 days. It is also possible to examine daily diary entries by adapting the scoring criteria to smaller units of analysis. For instance, Woike (1995; Woike & Polo, 2001) converted McAdams’s agentic and communal scoring criteria to score narratives of significant life events into a scoring system for daily diary entries.

Thus, content-analytic methods are used to measure a wide range of personological phenomena. By giving people simple guidelines for open-ended questions, it is possible to tap into their most important aspirations, hopes, and fears—as well as their unique capacity of understanding themselves and the world around them. Evidence shows that these systems can be used reliably, by independent judges, to discern and to categorize a wide array of responses. A large body of validation data demonstrates that these techniques reveal complex personality processes, including motivations, beliefs, and even probable courses of action.

A Step-by-Step Guide to Using Content Coding Techniques

The basic steps involved in content coding of open-ended responses have been identified (Smith, 2000). Each of these steps is discussed here in terms of general issues and decisions that must be made at each respective point in the research process.

- Step 1: What is the research question? What is to be identified, described, or measured? In personality psychology, open-ended response data may be used to test three basic types of hypotheses, which are not necessarily mutually exclusive. First, the method may be employed to identify a personality characteristic that is assumed to be relatively stable over time, such as exploratory style or achievement motivation. Second, open-ended responses may be used to assess a person’s understanding of a certain event, such as in Pennebaker’s (1997) technique for retelling traumatic experiences. This allows the researcher to ask questions pertaining to how a person reasons or makes sense of particular events and issues. There is the potential to uncover numerous patterns from these data. The technique may also allow researchers to uncover how describing the experience itself influences the individual—for ex-
ample, how writing about an event many times changes the person's understanding of that event over time. Finally, coding techniques for open-ended responses may be employed as outcome or dependent measures—for instance, studies that show that experimental manipulations such as accountability (Tetlock, 1983) influence the complexity of a person's narrative opinion of various issues. In addition, open-ended scoring procedures may valuable in post hoc explorations and as unobtrusive manipulation checks.

- **Step 2:** Decide whether content analysis will provide the needed information, either by itself or in conjunction with another method. It is often preferable to employ some other measure for cross-validation of the scoring system for open-ended responses, particularly if it is a newly developed system. However, it is often the case that the reason a content- analytic technique is employed is that the phenomena under investigation cannot be properly studied via other methods. It is also important to be mindful of the associations, or lack thereof, between self-report questionnaires and data coded from open-ended responses, inasmuch as lack of association may be due to method variance.

If the open-ended response system is used to assess a stable personality characteristic, such as the need for achievement, the results of the study itself may verify the validity of the open-ended measure. If open-ended responses are to be used as dependent variables, it may be especially desirable to obtain multiple measures derived from more than one assessment technique. Demonstrating an effect with converging methods offers convincing evidence for the hypotheses and validates the open-ended scoring technique at the same time.

- **Step 3:** Decide what type of qualitative material will best provide the information needed and how to obtain it. There are some important issues for consideration at this step in the research process. First, what questions should be asked? The task is to create enough structure to make meaningful comparisons and for meaningful differences to be detected, if they do in fact exist. A researcher would probably not want to approach an elderly participant and simply say, “Tell me the story of your life.” Insufficient structure would probably introduce too much variability. However, the questions should not be so structured that they resemble those of a self-report questionnaire. Should similar questions be asked more than once as a way to obtain a more stable measure? This technique works well in the scoring systems of measures such as sentence completion tests, in which many stems are provided, each response is scored, and then the total scores are examined to derive the participant’s score. Other methods, such as interviews, do not provide a way to obtain repeated measures, but asking questions in sequence provides continuity and some degree of structure across participants.

Should response time or number of words be limited in order to create some degree of standardization? How might this influence the responses? It may not be practical to allow participants an unlimited amount of time to tell their personal accounts. Yet imposing a time or word limit can create demand characteristics and limit spontaneity. Should the questions be asked via computer, written format, or orally? A computer format may be most efficient if the data will be analyzed using a computer-assisted scoring procedure. However, it is necessary that participants know how to use a computer and feel comfortable in responding this way. The written format may be the most common form of collecting open-response data, as it is easy for the participant to see the question or stem and then construct an answer in a standard format created by the researcher. Oral responding may be more appropriate for participants less familiar with a writing format, such as children, elderly individuals, and persons with learning disabilities.

Should responses be anonymous or given in face-to-face interviews? Anonymity is often preferable, as it believed to increase honesty and candidness. However, it may be necessary to conduct face-to-face interviews if participants are unfamiliar with the procedure or need more guidance to maintain their interest and increase the appropriateness of their responses, such as with children, elderly people, individuals with learning disabilities, or those who speak a different native language. Of course, if the researcher is analyzing archival data, he or she must take what is available in the format in which it is found. It may still be worthwhile to consider how the format may have influenced the responses—as some open-ended methods may be more susceptible to context effects than questionnaires.

- **Step 4:** Determine the unit of analysis to be coded. Depending on the decisions made in
the previous steps, the length of open-ended response material can range from a phrase, to a sentence, a paragraph, or many pages of text. Whatever the amount, a central task is deciding on a meaningful unit of analysis. Coding units used in prior research include (1) the entire text unit, such as an essay, (2) linguistically defined segments, such as meaningful words, clauses, or sentences, (3) response segments, such as a response to an interview question, (4) physically defined segments, such as a set number of words, and (5) temporal segments, such as number of minutes of a verbal response (Smith, 2000). Once the unit of analysis has been identified, a scoring system may be created to quantify the characteristics of interest.

- **Step 5: Select or develop a content coding system.** A number of scoring systems for open-ended responses have been developed in prior research. Therefore, reviewing the literature relevant to the research question is an important task. An example of how to develop a new scoring system for open-ended responses is presented in a case study in the next section of this chapter. In deciding on a scoring system, it is important to consider what types of scores are most desirable for understanding the data. There are three kinds of scores that may be generated from the content coding of open-ended responses: categories, ratings, and frequencies.

A categorization system is typically used to describe and identify broad orientations that are assumed to be related to personality structures and processes. The researcher develops a description of the characteristics that must be present for the responses to be categorized. The description must be general enough to apply to the various responses generated from the open-ended questions, as well as specific enough that coders can use the system reliably.

Ratings systems are similar to categories, except that they are incremental, and therefore it is not possible to score a unit of analysis for more than one rating. The psychosocial phenomenon being rated necessarily has a valenced quality. For example, the scale may range from positive to negative or from simple to complex. As with a categorization system, the researcher develops a description of the characteristics that must be present for the responses to be given a certain rating. The description must be general enough to apply to the various responses but specific enough to be used reliably. As in the integrative complexity scoring system (Baker-Brown et al., 1992) and Loevinger and Wessler's (1970) ego development coding system, the response to each open-ended question is given a rating. In this way, an average rating can be computed from each category. A potential problem with ratings is that there must be enough open-ended response questions so that when the ratings are used as a dependent measure they are relatively stable—that is, not based on one data point. This issue may be particularly problematic if specific meaningful units are being selected and certain cases contain only a smaller number of these units. Another problem can be obtaining enough variability among scores in the sample.

Frequency coding involves developing criteria for meaningful units of the response and recording the number of instances of these units in the data. Frequency coding can be quite labor-intensive, especially if responses are dense and lengthy. However, frequencies are a good choice for data in which the meaningful units occur rather infrequently and the researcher has multiple scoring criteria that can be discerned reliably from the data.

Another issue is whether it is most efficient to score the data using trained coders or whether the data can be scored via computer. There are obvious advantages to computer scoring in terms of efficiency and consistency. However, computers may miss important variability and nuances in the data that trained judges can detect easily. Computer-assisted scoring may be most useful for voluminous text in which simple sentences and phrases are to be counted, such as in the LIWC. But complex sentence structures and themes that may have an infinite number of semantic manifestations in open-ended responses do not lend themselves easily to computer-assisted techniques.

- **Step 6: Obtain pilot data to test and to refine the coding system.** An important consideration is whether the coding system will be (or has been) developed using an a priori approach, in which the criteria are specified before the material is examined, or an empirical approach, in which the criteria emerge from material to be analyzed (Smith, 2000). If the system is developed a priori, the researcher must compare the coding criteria with the pilot data and consider whether the phenomena of interest are meaningfully represented in the data. If not, then the researcher must consider whether the criteria can be modified to better
suit the data or whether a different set of a priori criteria should be used. If the a priori criteria are in fact used in the pilot data, then the researcher must decide if the definitions of the coding system are specific enough for independent judges, unaware of the hypotheses, to detect this content using the specified criteria.

If the empirical approach is employed, researchers can provide independent judges, unaware of the hypotheses, with broad definitions of the phenomena of interest and see what they find. Alternatively, researchers can examine the pilot data themselves and come up with scoring criteria based on the patterns in the pilot data. Then the researcher must develop the definitions of the coding system to be specific enough for a new set of independent judges, unaware of the hypotheses, to reliably detect the content of interest.

- Step 7: Train coders and ensure that intercoder agreement is satisfactory. Essential requirements for training coders are a clear coding manual, ample practice material, and an opportunity to discuss coding decisions with an expert scorer (Smith, 2000). The rules and procedures for the coding system must be explicitly defined in a scoring manual. These manuals typically include a general orientation to the phenomena of interest, then specific definitions of the content to be identified, as well as how it is to be identified, (e.g., categories, ratings, or frequencies). Many established scoring systems also include practice exercises for coders to hone their scoring skills. The practice exercises can be used to help coders achieve interrater agreement and to give researchers a good estimate of how accurately each coder is able to score the data. In general, it is a good idea for an expert coder to meet regularly with the group of coders to discuss their progress in learning the scoring criteria and to discuss any difficulties or ambiguities in the scoring system. It is most desirable for all data to be scored by at least two coders who are unaware of the hypotheses. In some cases, it may be acceptable for the data to be scored by one expert scorer and one naive scorer. The scoring systems mentioned in this chapter should be appropriate for advanced undergraduates or graduate students. It may be desirable to train more coders than needed and use only those coders who demonstrate the highest aptitude on practice or pilot data (Smith, 2000).

Intercoder agreement is computed differently, depending on whether the data are scores or categories. When coding systems yield scores, a correlation coefficient can be used to assess agreement. Assessing intercoder reliability with categories can be more complex. The simplest and most frequently reported category index is the percentage of agreement between two or more coders in classifying materials into two or more categories. For example, two coders may agree 75% of the time. There are two problems with this measure. First, it is affected by the frequency with which the category is present. Second, it does not take into account chance agreement. The first problem is most evident with low or high frequency of occurrences—for example, if achievement imagery is present in 2 out of 100 imaginative stories. If a coder misses the two instances and codes achievement imagery as absent in 100 out of 100 stories, then the coder's agreement with the practice exercises would be 98%. To correct the percentage of agreement for the frequency with which the category occurs, the following index was developed by McClelland and colleagues (1958):

\[
\frac{2 \times (\# \text{ of agreements between coders on presence of category})}{(\# \text{ scored present by coder 1}) + (\# \text{ scored present by coder 2})}
\]

If the scoring system involves several categories, it may be useful to compute the index for each category, both to assess agreement and to discover categories that may need to be eliminated or revised.

To correct for chance agreement, Cohen's (1960) kappa is often employed when the coding categories are independent and mutually exclusive. A weighted kappa may be used to evaluate agreement with ordinal categorical data. Kappa is an index that characterizes agreement in applying a coding scheme that varies from 0 (indicating no agreement) to 1 (indicating perfect agreement). It is a summary statistic derived from a matrix of the proportion of agreement actually observed and the proportion of agreement expected. For more details concerning computation and available programs, see Bakeman (2000).

What should be considered an acceptable level of intercoder reliability? Published research has tended to regard 85% agreement, or an interscorer correlation of .85 or more, as acceptable (Smith, Feld, & Franz, 1992), although a kappa score of .80 or more can be
viewed as satisfactory, depending on the data (see McDowell & Acklin, 1996). Once agreement is calculated and an acceptable level of intercoder reliability is achieved, coders can discuss their disagreements. In the final analyses, disagreements can be resolved through a consensus discussion or the disparate scores can be averaged.

• **Step 8: Obtain final material to be analyzed.** The final material should be collected by researchers unaware of the hypotheses. Ideally, the data should be collected by a group of researchers different from those who will score the open-ended response data. As in all data collections, it is important that all participants receive the same instructions and are tested in the same context under the same conditions. Open-ended responses can have a good deal of variability, and the experimenter needs to be careful not to inadvertently introduce more while collecting the responses.

• **Step 9: Code the material with identifying characteristics removed, and determine inter-rater reliability; or perform computer-assisted content analysis.** Ideally, handwritten data should be transcribed. Coders who have achieved an acceptable level of reliability on practice materials or pilot data, and who are unaware of the hypotheses, should work independently to score the data. The scores should then be compared; the data must meet acceptable levels of reliability, as discussed in Step 7, before any data analysis or interpretation is performed on the data.

• **Step 10: Analyze the data; carry out cross-validation if appropriate.** An exhaustive discussion of data analytic strategies is beyond the scope of the chapter. However, some important issues for consideration when using categories, ratings, and frequencies are discussed here. First, if responses may be classified into a fixed number of mutually exclusive categories, then they must be treated as categories in the analyses. If hypotheses are to be tested with inferential statistics, it is essential that there is a large enough sample size for each category. If the other variables in the study are also categorical, then the chi-square statistic may be most appropriate. For instance, Woeke, Candel, and Osier (1996) investigated the relationship between college men’s attachment styles (secure, avoidant, anxious) and the violent imagery they used to write imaginative stories about relationships. Criteria for scoring violent imagery developed by Pollack and Gilligan (1982) were adopted, and the discrete categories of violent imagery (e.g., verbal threats, murder) were combined into one violent imagery category because the individual categories occurred in low frequency. A chi-square analysis was conducted using the three attachment categories and the presence or absence of violent imagery. The frequency of anxious men who wrote stories with the presence of violent imagery was significantly greater than expected by chance.

If the other variables in the study are continuous, it is possible to compare the means of these variables between categories using parametric statistics. For instance, Woeke, McLeod, and Goggin (2003) asked people to describe two memories, which were then categorized as either agentic or communal following McAdams’s (1985) scoring criteria. The prediction was that people would be more likely to recall a memory that was congruent with their motivation (either achievement/power or intimacy). To test the prediction, t-tests were performed on the motive scores, which were continuous variables between the presence and absence of the categories. For instance, the investigators compared the average of the achievement/power motives scores for participants who recalled a memory categorized as agentic with the scores of those who recalled memories that were not categorized as agentic. It was found that those who recalled an agentic memory had higher achievement motivation scores than those who did not recall an agentic memory, and the same pattern was found for intimacy motivation and communal memories.

Second, the data yielded from scoring open-ended responses may be ratings on a continuum. In this case, it is important to first examine the descriptive properties of the data. Is there a normal distribution? In many cases, rating systems for variables reflecting psychological or intellectual maturity, such as the scoring systems for ego development and integrative complexity, tend to be skewed or restricted in range on the lower end of the scale, particularly in college populations (Woeke, 1994). Thus, it is important to obtain a sufficient sample size and ratings that represent a number of responses per participant. It may be useful to correct for range restriction. If there is a broad range of skewed scores, it may be useful to convert ratings to standard scores. After taking these issues into consideration, the ratings can be used to test a range of hypotheses with most inferential statistics.

Finally, the scoring of open-ended responses may yield frequency data. As with ratings, it is
important to examine the distribution of frequency data. Another very important consideration is how the frequencies relate to the size of the response, that is, how do frequencies relate to the number of words used in the response? Many researchers argue that all frequency data should be corrected for word count (see Schultheiss & Pang, Chapter 19, this volume, for more details about correction techniques). However, it may behoove the researcher to consider alternatives. If we are scoring for specific themes that by necessity need words to be presented, it is the mere presence of these themes—not the number per number of words—that is important. For instance, if an individual's written recollection of a traumatic event reflects a greater insight and understanding based on the frequency of certain specified words, the actual written recollection may be longer than one that does not reflect such insight and understanding. It would be an inaccurate representation of the response to calculate the "insight per word" ratio, as it would lose all external validity.

- **Step 11:** Interpret results. Compare your findings with norms, if available. The interpretation and generalizability of the findings depends largely on the research questions. Open-ended response data will provide some corroboration of past findings and current hypotheses and add new information to the psychological phenomena under question.

### Case Example: Using and Developing Hypothesis-Driven Content-Analytic Scoring Systems

My research deals with how personality motivation influences social perceptual-cognitive processes, most specifically autobiographical memory. In doing this work, my colleagues and I have used several open-ended response scoring systems. Therefore, the work seems suitable for a case demonstration of the development and use of open-ended response scoring techniques. The steps outlined above are used to organize the case study, and in some instances the methods of multiple studies (e.g., Woike, 1994, 1993; Woike & Aronoff, 1992; Woike et al., 1999; Woike, Mcleod, & Goggins, 2003; Woike & Polo, 2001) are combined for ease of presentation.

- **Step 1:** What is the research question? What is to be identified, described, or measured? The empirical question was, How does personality motivation influence autobiographical memory—specifically, what is remembered and how is it remembered? To conduct the study, we needed a measure of personality motivation, a measure of memory content (what) and a measure of memory process or structure (how). In considering how to measure two broad and contrasting types of personality motivation (achievement/power and intimacy), we decided to use a version of the Thematic Apperception Test (TAT; see Chapter 19, this volume). For gathering the memory data, open-ended questions were useful to test the hypothesis involving what, that is, What are people's most important memories about? Specifically, we predicted that the content of people's most important memories would be thematically related to their motivations. Therefore, we needed to identify memory content related to these two broad motivational categories, agency (or task-related concerns) and communion (or socially related concerns). Because we were also interested in how—that is, How does one explain an autobiographical memory?—open-ended data were used to examine patterns in the written recollections of these experiences. It was predicted that the structure of the memories would be related to motivations as well. We predicted that achievement/power-motivated individuals would focus on differences as a way to gauge their performance, whereas intimacy-motivated individuals would focus on similarities as a way to relate to others (for a more thorough discussion of this reasoning, see Woike, 1994). Thus, we needed a way to identify the perception of differences and similarities in the memory data.

- **Step 2:** Decide whether content analysis will provide the needed information, either by itself or in conjunction with another method. In order to test the hypotheses, we needed separate measures of personality motivation and autobiographical memories. To obtain a measure of autobiographical memory, we determined that asking an open-ended question or series of questions would be the best way to gather information on the variability in content and in structure of autobiographical material. Because our research question did not hinge on accuracy or corroboration of the events, we determined that an autobiographical memory written by each participant from his or her own point of view would be an appropriate and sufficient measure of autobiographical memory. In considering how to measure personality moti-
vation, we decided to use an open-ended scoring technique, a version of the TAT, that was already well established in the literature. These measures do not correlate with self-report measures of motivation (see Woike, 1995), so no other personality measures were employed for cross-validation. Thus, we assessed motive and memory variables independently—but because both assessment procedures involved open-ended responses that were content scored, the measures were at the same level of analysis and therefore did not introduce unnecessary method variance.

- **Step 3. Decide what type of qualitative material will best provide the information needed and how to obtain it.** For the measure of autobiographical memory, we thought that one memory of an important or significant event in the participant's life would suffice. Participants were instructed to think for a few moments about their life experiences and then chose one experience that they felt was significant and important. They were told that they had about 10 minutes to write out their memory. This allowed them time to write about an experience without being rushed, but did not provide extended time for deliberation or to edit or revise their writing, thus increasing the spontaneity of their responses. In these ways, we attempted to control for unnecessary method variance. It was also decided that participants would write their memories by hand on paper that could be transcribed into type-written text. In doing so, we were able to test a large group of participants at a time. Participants' personal identification information was removed from their materials in an effort to reduce their social desirability concerns. However, it was likely that, given free choice to write about any important experience, participants would choose one that they would enjoy recalling (rather than a painful or negative experience that was also significant), which was consistent with our motivation hypothesis, that people are more likely to recall experiences related to their motives.

- **Step 4: Determine the unit of analysis to be coded.** Because each memory pertained to a specific event, the whole memory was used as the unit of analysis for the content scoring procedure. To score the memories for differences and similarities, it seemed most reasonable to identify meaningful phrases within the memories that had these respective qualities.

- **Step 5: Select or develop a content coding system.** In a review of the literature on motivation and personal narratives, we found McAdams's (1985) life experiences scoring system, in which each personal narrative could be sorted into agency, communion, both, or neither, based on broad thematic scoring criteria, to be suitable to measure of task (agency) versus social (communion) themes.

In a review of the literature on social perception and cognitive complexity, we found that the Conceptual/Integrative Complexity System (Baker-Brown et al., 1992) was the best system to measure differences and similarities in the text of the memories because it had operational definitions for both differentiation and integration. However, because this system is based on ratings, it did not provide a way to measure differentiation and integration independently—which was essential for our hypotheses. Therefore, we decided to develop a scoring system for the two components of complexity, differentiation and integration (Woike, 1994, 1997; Woike & Aronoff, 1992), to measure the frequencies of two forms of complexity found in open-ended responses. This system allowed for the separate assessment of the components of differentiation and integration.

- **Step 6: Obtain pilot data to test and to refine the coding system.** In developing the system for cognitive complexity, we consulted different systems in the literature (see Woike & Aronoff, 1992, for a complete list). We then wrote out descriptions of the characteristics that pertained to differentiation and integration for a group of naive coders to find in pilot data of autobiographical memories. From these initial scorings, the system was further developed; in some cases categories were eliminated because they could not be found reliably in the pilot data. In other cases, categories were too broad and were divided into subcategories that could be identified reliably. The system went through several revisions; each revision in the scoring definitions was tested with pilot data until each subcategory would be identified reliably, based on the definition in the scoring manual. In the end, we had four subcategories for differentiation and four for integration.

- **Step 7: Train coders and ensure that intercoder agreement is satisfactory.** A scoring manual was developed that included clear definitions and examples for each of the subcategories. We used the pilot data to create sets of practice exercises to allow the coders to learn the system and to assess their reliability. Reliability was determined using the formula for establishing reliabilities with frequency data.
Content Coding of Open-Ended Responses

Discussing earlier, each coder needed to achieve 90% agreement with the expert scoring key before he or she could begin scoring the data.

- **Step 8:** Obtain final material to be analyzed. Participants were pretested for agentic and communal motives as described above. Those who scored high on one motive and low on the other were called back to participate in a follow-up study by research assistants unaware of the hypotheses. In a neutral classroom setting, all participants received the same written and verbal instructions for completion of the memory writing task. Their names did not appear on the materials.

- **Step 9:** Code the material with identifying characteristics removed, and determine inter-rate; or perform computer-assisted content analysis. Two coders categorized the memories for agency and communal themes, and another group of coders scored the frequencies of the categories of complexity in the memory data. Their reliabilities were determined, using Cohen's kappa for the category data, and the formula for determining the reliability of frequency data for the complexity data, as discussed above.

- **Step 10:** Analyze the data; carry out cross-validation if appropriate. To test the hypotheses, we first conducted a chi-square comparing the number of agentic versus communal individuals who had memories categorized as agentic versus communal. As predicted, we found that more agentic participants wrote about a significant memory that was categorized as agentic, whereas the more communal individuals wrote about memories categorized as communal. Next, we looked at the complexity of the memories within a 2(motive) × 2(theme) multiple analysis of variance (MANOVA) with differentiation and integration as the dependent variables. An interaction was found between motive, theme, and type of complexity, in which agency-motivated individuals whose memories had agentic themes organized their memories using more differentiation and communal-motivated individuals whose memories had communal themes organized their memories using more integration.

- **Step 11:** Interpret results. Compare your findings with norms, if available. From these findings it appeared that personality motivation did indeed influence what and how autobiographical memories were recalled. Multiple tests of the content hypothesis (Woike, 1994, 1995; Woike et al., 1999, 2003; Woike & Polo, 2001) have replicated the same thematic patterns under different conditions. The influence of these motives on the complexity with which personal experiences are recalled in written recollections has also been replicated (Woike, 1994; Woike et al., 1999; Woike & Polo, 2001). Further studies have shown evidence that the organizational patterns found in written narratives reflect the influence of motivation on the patterns of encoding, processing, and retrieval of the memory of personal experiences. Computer-based encoding and retrieval studies have found that these motives serve a selective and organizing function for incoming information, and then influence the retrieval of this information as well (Woike, Lavezzary, & Barsky, 2001; Woike, Mcleod, & Salzberg, 2007). This work has led to the development of a model of the influence of personality motivation on memory processes beyond written recollections (Woike, 2007). Without the initial open-ended response data from which the scoring system for differentiation and integration was derived, it would not have been possible to make this discovery.

---

**Complexities: Determining Reliability and Validity with Open-Ended Response Data**

This final section deals with some considerations for determining the psychometric properties of open-ended data. Open-ended data differ from questionnaire data in several ways. Therefore, it is important not to subject data derived from coding open-ended responses to statistical techniques that have been designed to test reliability and validity of questionnaire data.

Questionnaires have different properties. In using questionnaires, it is assumed that each item is discrete and can be presented in any order to the participant and that this will not affect the participant's score. Data from open-ended responses are typically a coherent set of verbal responses that form a specific rather than a random pattern. Open-ended responses typically have a beginning, middle, and end. If the sentence were to be scrambled (as is possible with questionnaire items), the response would lose its meaning. Therefore, a technique like factor analysis, which is based on having a fixed number of responses to a number of discrete questions, would not be an appropriate technique to use to examine the internal prop-


certics of open-ended responses. Other techniques, such as split half reliability, are also not appropriate for the same reasons.

Test–retest reliability may also be difficult to achieve with open-ended data using standard procedures. Even if the participants are asked the same questions, at different points in time, there are an infinite number of ways they can respond to an open-ended question, whereas questionnaires have a limited number of choices. Test–retest reliability may also be lower with open-ended data if questions involve creativity or storytelling, as the participant may make efforts to not repeat the same response. As open-ended responses may have more problems with reliability than questionnaires, the validity of open-ended response instruments may arguably be easier to establish. By asking people directly what they think and feel, we know about their cognitive representations of their thoughts and feelings and what sorts of information are important to them, rather than those of the researcher.

Recommended Readings


References


Pennebaker, J. W. (1997). Writing about emotional ex-
en coding of open-ended responses.

