A HANDBOOK OF

Clinical Scoring Systems for Thematic Apperceptive Techniques

EDITED BY
SHARON RAE JENKINS
A HANDBOOK OF
Clinical Scoring Systems for
Thematic Apperceptive
Techniques
Series in Personality and Clinical Psychology
Irving B. Weiner, Editor

Bornstein/Masling (Eds.) • Scoring the Rorschach: Seven Validated Systems
Calhoun/Tedeschi • Facilitating Posttraumatic Growth: A Clinician’s Guide
Dana (Ed.) • Handbook of Cross-Cultural and Multicultural Personality Assessment
Eisen/Quas/Goodman (Eds.) • Memory and Suggestibility in the Forensic Interview
Exner (Ed.) • Issues and Methods in Rorschach Research
Frederick/McNeal • Inner Strengths: Contemporary Psychotherapy and Hypnosis for Ego-Strengthening
Gacono (Ed.) • The Clinical and Forensic Assessment of Psychopathy: A Practitioner’s Guide
Gacono/Meloy • The Rorschach Assessment of Aggressive and Psychopathic Personalities
Ganellen • Integrating the Rorschach and the MMPI-2 in Personality Assessment
Goodman/Rudorfer/Maser (Eds.) • Obsessive-Compulsive Disorder: Contemporary Issues in Treatment
Hall/Bernieri (Eds.) • Interpersonal Sensitivity: Theory and Measurement
Handler/Hilsenroth (Eds.) • Teaching and Learning Personality Assessment
Huprich (Ed.) • Rorschach Assessment of the Personality Disorders
Hy/Loevinger • Measuring Ego Development, Second Edition
Jenkins (Ed.) • A Handbook of Clinical Scoring Systems for Thematic Apperceptive Techniques
Kelly • The Assessment of Object Relations Phenomena in Adolescents: TAT and Rorschach Measures
Kohnstamm/Halverson/Mervielde/Havill (Eds.) • Parental Descriptions of Child Personality: Developmental Antecedents of the Big Five?
Loevinger (Ed.) • Technical Foundations for Measuring Ego Development: The Washington University Sentence Completion Test
McCallum/Piper (Eds.) • Psychological Mindedness: A Contemporary Understanding
Meloy/Acklin/Gacono/Murray/Peterson (Eds.) • Contemporary Rorschach Interpretation
Needleman • Cognitive Case Conceptualization: A Guidebook for Practitioners
Nielsen/Johnson/Albert Ellis Institute • Counseling and Psychotherapy With Religious Persons: A Rational Emotive Behavior Therapy Approach
Nolen-Hoeksema/Larson • Coping With Loss
Rosowsky/Abrams/Zwieg (Eds.) • Personality Disorders in Elderly: Emerging Issues in Diagnosis and Treatment
Sarason/Pierce/Sarason (Eds.) • Cognitive Interface: Theories, Methods, and Findings
Silverstein • Self Psychology and Diagnostic Assessment: Identifying Selfobject Functions Through Psychological Testing
Taylor (Ed.) • Anxiety Sensitivity: Theory, Research, and Treatment of the Fear of Anxiety
Tedeschi/Park/Calhoun (Eds.) • Posttraumatic Growth: Positive Changes in the Aftermath of Crisis
Van Hasselt/Hersen (Eds.) • Handbook of Psychological Treatment Protocols for Children and Adolescents
Weiner • Principles of Rorschach Interpretation
Wong/Fry (Eds.) • The Human Quest for Meaning: A Handbook of Psychological Research and Clinical Applications
Zillmer/Harrower/ Ritzler/Archer • The Quest for the Nazi Personality: A Psychological Investigation of Nazi War Criminals
CONTENTS

Preface  xi
Sharon Rae Jenkins

SECTION I   GENERAL PRINCIPLES

1  Introduction
Why “Score” TATs, Anyway?
Sharon Rae Jenkins  3

2  Teaching How to Learn Reliable Scoring
Sharon Rae Jenkins  39

3  Future Directions for Thematic Apperceptive Techniques
Sharon Rae Jenkins  67

SECTION II   PERCEPTUAL-COGNITIVE SYSTEMS

4  Edith Weisskopf’s Transcendence Index
Benjamin P. Chapman  89

5  Scoring Manual for Edith Weisskopf’s Transcendence Index
Edith Weisskopf, Max Prola, Luis E. Perez, and Sharon Rae Jenkins  103

6  Richard Dana’s Objective TAT Scoring System
Cara L. Santa Maria  115
SECTION III PSYCHODYNAMIC SYSTEMS

14 Reuben Fine’s Scoring Scheme for the TAT and Other Verbal Projective Techniques 261
   Sidney Ornduff

15 Scoring Manual for Reuben Fine’s Scoring Scheme for the TAT and Other Verbal Projective Techniques 271
   Sidney Ornduff
17 Primary Process Scoring Manual 293
Robert R. Holt

18 Pathogenesis Index 347
Bertram P. Karon

19 Scoring Manual for the Pathogenesis Index 365
Bertram P. Karon

20 TAT Oral Dependency Scale 385
Steven K. Huprich

21 Scoring Manual for Oral Dependency 399
Steven K. Huprich

22 Symbiosis 411
Frank Summers

23 Scoring Manual for Symbiosis 419
Frank Summers

24 The TAT Ambitious-Narcissistic Style Scale 429
David W. Harder and Deborah F. Greenwald

25 Scoring Manual for the TAT Ambitious-Narcissistic Style Scale 437
David W. Harder and Deborah F. Greenwald

SECTION IV SOCIAL-EMOTIONAL SYSTEMS

26 Singer and Wynne’s Communication Deviance Scoring System 449
Benjamin P. Chapman
SECTION V NEEDS AND CONCERNS FOCUSED SYSTEMS

35 The Need- Threat Analysis
A Scoring System for the Children’s Apperception Test
Louis A. Chandler

36 Scoring Manual for the Need-Threat Analysis
Louis A. Chandler
Sharon Rae Jenkins

The story—from *Rumplestiltskin* to *War and Peace*—is one of the basic tools invented by the human mind, for the purpose of gaining understanding. There have been great societies that did not use the wheel, but there have been no societies that did not tell stories.

*Ursula K. LeGuin, 1979*

When I am having difficulty understanding a therapy client’s problem, I usually want to ask for examples of the problem. Sometimes this works, but more often, the client’s response is hesitant and fragmented. A small change in wording is invariably helpful: “Tell me a story about a time when that happened.” What follows usually flows into a clear illustration of the problem and its context, antecedents, and sequela. Omissions can be important, too. Taping the session would allow for post hoc analysis of the response using relevant scoring systems from this book.

As a field, psychology has been struggling for over half a century to integrate science and practice. As a culture, one challenge for American society from the beginning has been to integrate into a functioning social system peoples having very different nationalities, ethnic habits, and personal histories. Our technology-driven lives only increase in complexity, with resulting tensions toward social fragmentation straining all our integrative functions.

Storytelling is a quintessentially integrative function, as shown by studies of narrative memory. Stories bring people, ideas, and feelings together around campfires and research groups. Folktales build cultures; bedtime stories raise children. Stories show what people and their societies value and wish to communicate. Understanding stories helps us understand these things, which makes them useful for clinical work with clients who have trouble understanding themselves. This use of stories has a 70-year history.

The tensions in that history are in part the tensions between science and practice. The skilled clinician accrues a repertoire of analytic approaches over a lifetime of experience—but where is the objective evidence of reliability and validity? The researcher demonstrates associations between features of stories and people’s career choices and long-term health outcomes—but how is this relevant for an individual client? How can we make generalizations about individuality?

A fundamental hypothesis behind this book is that structured scoring systems for thematic apperceptive techniques (TATs) are one way to bridge the scientist–practitioner gap by enabling scientifically sound, efficient, and clinically informative examination of
clients’ stories to answer focused clinical questions relevant to diagnosis, estimation of prognosis, and effective treatment planning. Testing this hypothesis requires a body of research using specific systems for clinically appropriate purposes. Such research has been difficult because the scoring manuals and practice stories necessary for both clinical and research use have not been available widely, if at all. This collection of scoring manuals makes available a selection of promising ones, and the summary chapters for each explain their best uses, evidence for validity and reliability, and priorities for future research.

This book should be in the library of every faculty member and clinical supervisor who is responsible for teaching courses in psychological assessment or supervising assessment students in clinical, counseling, school, or forensic psychology, whether in academic or practice settings, practicum sites, or internships. It may also be useful as an assessment course text. In chapter 2, procedures for learning how to score will be especially useful for that purpose, as its pedagogical principles generalize to data from other less structured assessment techniques. Practicing assessment psychologists will find the book a handy reference resource, given its coverage of common clinical problems. Researchers in these areas who are interested in storytelling techniques, especially thesis and dissertation students, will find it useful for identifying interesting research problems and choosing variables, as well as for training scorers. They may also turn to chapter 3 for broader suggestions for future research.

Any movement from general to specific, or vice versa, involves a leap of inference. The traditional clinical approach to story analysis has involved large leaps of inference—albeit highly educated—using abstract theoretical constructs and frameworks to structure the assessor’s understanding of the stories and thus of the particular client. This approach allows for a highly detailed and individualized understanding of the person, but might make it difficult to see the client’s commonalities with others. Yet if no generalizations can be made, how can we know whether what helped the last client might help this one?

Structured scoring systems minimize leaps of inference by providing the assessor with detailed, specific descriptions of what to look for in each story: words, phrases, images, structural features, degrees of intensity. These scoring category criteria indicate what to do if the material described is found: assign a point or a scale rating, or make some other decision. One requirement of these descriptions is that they be sufficiently clear and detailed that most appropriately trained people, given the same rules for observation, will usually see the same thing in the same place. This requirement of interscorer reliability meets the scientific criterion for objectivity, that is, agreement on whether or not the phenomenon to be identified has in fact been observed.

Academic psychology has had a 55-year history of such scoring systems for personality variables, most notably the human motivation scoring systems pioneered by David C. McClelland, John W. Atkinson, David G. Winter, and their colleagues, as discussed in more detail in chapter 1. The fact that there is a parallel 55-year history in clinical psychology is less well known, in part because the lengthy and detailed materials for using these systems have rarely been published. Phebe Cramer’s 1991 book, *The development of defense mechanisms* (Springer-Verlag), is a notable exception. For my Faculty Development Leave from the University of North Texas, I set out to locate interesting systems for which there was both some evidence of interscorer reliability and some evidence for validity of a clinically useful nature, as described in more detail in chapter 1. Unfortunately, from the start my search was limited by my lack of adequate competence in languages other than English, which forced the omission of Vica Shentoub’s highly recommended 1990 *Manuel d’utilisation du TAT* (Dunod).
Not only were there many more systems with publication records than I had guessed, but also there were so many that I had to set priorities for choosing among similar systems! Of the 37 clinically interesting systems that were initially identified and ultimately not included, 14 were dropped for lack of at least one solid published clinical validational study or evidence of one in progress, and seven did not report evidence of interscorer reliability. In some cases, more recent systems known to me had incorporated older ones, making pursuit of the latter unnecessary; seven were redundant with and superseded by a system that met more criteria better.

The next challenge was to find the originator and to explore the possibility of chapter authorship. In that, too, I was surprisingly successful, given the ages of some of the earlier authors. Melvin Feffer was pleased and flattered to hear from me in his retirement, he said, and he gave permission for me to revise and elaborate on the manual that he had thoughtfully archived at the Library of Congress American Documentation Institute. Sadly, he died before seeing the final product. Sidney Ornduff was in a somewhat similar situation with Reuben Fine, whose manual she obtained before his death and developed for the studies described in her contribution herein. The authors who present their work here are to be commended for rooting through files that were not always ready at hand, and that in some cases might have required Stygian labors for their retrieval from basement boxes of dissertations long past.

Of the nine systems that were pursued but are not included here, two authors could not be located or did not respond to communications. Four authors declined to have their systems included. Only three systems were lost due to the tidy habits of retired authors who had discarded their materials.

The systems included here are organized into four sections, presented in roughly chronological order of their first appearance in publication, as are the systems within each:

**Perceptual-Cognitive:** Edith Weisskopf, Richard Dana, Melvin Feffer, George Ronan and Margaret Gibbs, Barry Ritzler

**Psychodynamic:** Reuben Fine, Robert Holt, Bertram Karon, Steven Huprich, Frank Summers, David Harder and Deborah Greenwald

**Social-Emotional:** Margaret Singer and Lyman Wynne; Antoinette Thomas; Giuseppe Costantino and Robert Malgady; Hedwig Teglasi, Constance Locraft, and Kelly Felgenhauer; David Joubert; Michelle Hoy-Watkins and Valata Jenkins-Monroe

**Needs and Concerns Focused:** Louis Chandler, David Ephraim

Each system is discussed in more detail toward the end of chapter 1.

Although Weisskopf died some years ago, Max Prola had preserved a version of her scoring manual as an appendix to his dissertation. With the much appreciated help of Clifford Swenson and Chris Smith at Purdue University, I subsequently located her original manual and practice stories in the collection of her papers at the State University of West Georgia. I thank Myron House and Laura Henry at the University of West Georgia’s Special Collections for their assistance during a pleasant but busy afternoon in their archives.

For my personal history with TATs, I thank Richard Teevan, who chaired my undergraduate honors thesis at Bucknell University on achievement motivation and fear
of failure in groups. When I was ready to find a graduate school, Jacqueline Fleming wel-
comed my interest in Matina Horner’s research on fear of success, taught me the scoring
system, and invited me to their research group. There I met Abigail Stewart, who coinci-
dently came to Boston University to teach in the same week that I arrived there as a stu-
dent in the new personality program. She taught me her scoring systems and supervised
my dissertation using human motivation variables and her own Self-Definition/Social
Definition system. Much of chapter 2 grew directly from my experiences of learning to
score through individual study with Fleming and in Stewart’s scoring groups.

When Carol Huffine recruited me to teach research methods courses at the
California School of Professional Psychology, Berkeley/Alameda (CSPP-B/A), I was quite
surprised to learn how differently my clinical psychology doctoral students approached
TAT stories. It was disorienting not to have the focus of a structured scoring system’s
organizing principles for looking at a story. And they saw things that they said were obvi-
ous, but I could not see them! Sometimes I was not alone in that, either—so much for
interscorer reliability, which did not seem to bother them as much as it did me. That was
my introduction to the clinical uses of TATs, which was sufficiently provocative to inter-
est me in learning more by taking a postdoctoral clinical certificate at CSPP-B/A, where
I subsequently studied assessment with Gerald Michaels, Jacqueline Singer, David Stein,
and Susan Fair.

Most of my clinical training with TATs has come from books by Magda Arnold, Leo
Bellak, Phebe Cramer, William Henry, and Silvan Tomkins, with my clinical faculty and
supervisors helping with the integrative functions of weaving the resulting observations
into the rest of the data from the classical full battery assessment. I came to appreciate
those authors’ depth of insight, while also wondering, on internship working with Kevin
Riley, where they could possibly find the time to take this approach with all of their cli-
ents. Clearly, the haste of modern life would not make this easy. Arguably, one advantage
of structured scoring systems is that they enable well-focused information gathering with
relatively less time and cogitation.

Several senior clinician scientists have been generous with their time and encour-
Apperception Test* (Guilford), was among my first sources of systems to pursue. Her willing-
ness to spend time with me to offer suggestions about avenues to search and people to
contact was essential to my rapid progress. Richard Dana became enthusiastic quickly,
offering advice whenever I asked and sometimes before I was aware of needing it. His
ambitions for the book inspired me. I thank Hedwig Teglasi for her 1993 book, *Clinical use
of story telling* (Allyn & Bacon), another major reference source, and for her professional
contributions to making TATs accessible to clinicians working with children, for our con-
versations, and for her enthusiasm for the work.

I am much indebted to Charles P. Smith for the careful structuring of his 1992 vol-
ume, *Motivation and personality: A handbook of thematic content analysis* (Cambridge University
Press), which was the first inspiration for this project. Examining it closely shaped my
thinking about how to organize this book and what features likely would help readers.
His suggestions at an early stage proved strategically useful.

Virginia Demos and Chris Fowler lent their sensitive ears and inquiring minds to
the early stages of my ideas for the project and its later development. Bonnie Strickland
alerted me to articles of likely interest, and she and Marjorie Nott shared their home for
workspace during my leave time. Bob Holt and Bert Karon were always ready to engage
with the progress of the work. Irving Weiner and Drew Westen furnished stimulating
ideas in writing, presentations, and conversations.
Important boosts in the search process came from Amanda Phillips, Gladys Croom, Stephanie Dudek, Clifford Swenson, Sandra Russ, Jack Gerber’s Rorschach Discussion List, the Projectives Discussion List, Robert Garlan, Norm Abeles, Lee Zimmerman, and David B. Baker, director of the Archives of the History of American Psychology at the University of Akron. The assistance and hospitality of the library staff at the University of Massachusetts at Amherst’s W.E.B. DuBois Library was invaluable to this guest user. The availability on their shelves of their entire collection facilitated the necessary hand search of tables of contents of the three major journals that published structured scoring systems in my pursuit of the many systems that evaded electronic indexing. My literature searches there formed the backbone of this volume.

The staff of the University of North Texas Interlibrary Loan merits special mention for handling the incoming stream of dissertations that might contain usable scoring systems. Among our Psychology Department staff, Lee Ward provided crucial support by decoding my edits, anticipating what I should have said, and formatting references with her expertise using American Psychological Association style; Phyllis Dever and Stacy Suits solved problems that, thanks to their timely intervention, never came to my attention.

Among my students, I thank Lauren Dobbs, Melissa Leeper, Luis E. Perez, Rachel White, and Derrick Carter for trusting me to guide their honors theses into potentially Herculean labors. Numerous scorers contributed suggestions and examples for the development of Ephraim’s Psychocultural System, Feffer’s Interpersonal Decentering, Huprich’s Oral Dependency, Thomas’s Affective Scale, and Weisskopf’s Transcendence Index, as those authors have acknowledged. I appreciated their comments, additions, and elaborations as they pilot tested the appendix for chapter 2 on learning scoring. Diana Brown and Lizzie Woodruff were helpful in so many ways that I am sure I have not recalled them all, because they recognize how helpful unobtrusiveness can be.

At Routledge, I thank Susan Milmoe, my original editor, for her encouragement and support; Steve Rutter, her successor, for his patience with my creative approach; Nicole Buchmann, for her rapid responses at crunch time, also a time of transition for the organization; George Zimmar, Mimi Williams, Robert Sims, and the production staff at Taylor and Francis for seeing it through the final process; and Larry Erlbaum for his congenial hosting of the social hours at the Society for Personality Assessment meetings that helped me network to locate systems.

Finally, I thank Kwame Azalius Ross for the concept of accountability to the work.

Sharon Rae Jenkins
General Principles
Introduction
Why “Score” TATs, Anyway?

Sharon Rae Jenkins

Introduction

The Thematic Apperception Test—“the TAT” (Murray, 1943)—has a long, honorable, and well-argued 65-year history in psychological research and clinical assessment. In a recent surge of activity, several books have taken stock of that history (Douglas, 1993; Gieser & Stein, 1999; Robinson, 1992) or have reviewed important research and clinical applications (e.g., Aronow, Weiss, & Reznikoff, 2001; Costantino, Malgady, & Rogler, 2002; Cramer, 1991, 1996, 2006; Dana, 2005; Costantino, Dana, & Malgady, 2007; Kelly, 1996, 1997; Smith, 1992a; Teglasi, 1993, 2001). In response to the historical and cultural specificity of the original pictures, new picture sets have been developed, each with its own scoring system, four of which are described in this book. As there is only one “the TAT”—Henry Murray’s (1943) set of published picture stimuli—all of these and the associated scoring systems are here referenced collectively as thematic apperceptive techniques (TATs).

Clinical and counseling psychologists have long known the usefulness of TAT story protocols for those insights into how a client sees the world that can be gained through systematic idiographic interpretation (Aron, 1949; Arnold, 1962; Bellak, 1954; Henry, 1956; Shneidman, 1951; Stein, 1981; Tomkins, 1947). Academic researchers have made heavy use of similar stories with objective, structured scoring systems to measure human motivation and other personality characteristics (e.g., Atkinson, 1981; McAdams, 1980, 1985; McClelland, 1965, 1975, 1985a, 1985b, 1989; Smith, 1992a; Stewart, 1992; Winter, 1973). Unfortunately, clinicians and academic researchers seldom discuss these common interests, leaving a gap between the wide clinical application of TATs and the scientific evidence for their validity. Despite early attempts at integrating these views (e.g., Dana,
1955a, 1959; Holt, 1978; Murstein, 1972; Zubin, Eron, & Schurei, 1965), further developments in clinical research historically have been slow. In academic psychology, Smith (1992a) collected empirically supported scoring systems for 14 of the most widely used narrative-based research measures of motivation and personality. Each is represented by a research summary, a detailed scoring manual, and practice stories with expert scoring that are designed to make a system self-teaching. There are also several integrative chapters on historical, conceptual, and methodological issues. That volume was the original model and inspiration for this one.

This book brings together the best available clinical scoring systems for TATs, presented in research summaries along with detailed scoring manuals and practice stories. Chapter 2 concerns how to learn and teach these systems. It is designed to enable clinicians, researchers, faculty, and supervisors to apply these systems efficiently and accurately. Chapter 3, on future directions, discusses needed clinical applications and research programs to encourage further research contributions. Like Smith (1992a), this one is essentially a user’s manual designed for professional self-teaching, training research teams, and classroom instruction.

Why Structured Scoring Systems?

The strengths of structured scoring systems lie in their flexibility, efficiency, generalizability, consistency, and objectivity. The strength of flexibility is that, like the Rorschach Ink Blots and unlike structured self-report scales, TATs are essentially data-gathering techniques that entail a two-step process. First, storytellers are shown a series of pictures and are asked to tell a story about each one, under standardized conditions (or in specified experimental situations). Second, and separately, useful information is gleaned from the resulting protocols. Structured self-report measures depend on conceptualization of the construct for the proper wording of items, which might not generalize across conceptualizations, theories, or cultures. In contrast, any theoretical perspective or data-analysis approach can be used with TATs once the stories have been gathered; the data are transparent to purpose. It is this second step that is evaluated for psychometric validity and reliability, as those terms are meaningless for narrative data. Anastasi (1988, p. 23) defined a psychological test as “an objective and standardized measure of a sample of behavior.” For structured self-reports, the measure is the sample, but for TATs, the objective and standardized measure is quite separate from the sample.

Unlike the Rorschach, which has a standard set of 10 card stimuli, the first step also allows flexibility for TATs, as a set of stimuli must be chosen. Although these are typically pictures, sentence cues have been used effectively in research (Jenkins, 1987, 1994; Smith, Feld, & Franz, 1992) and hold promise for clinical use with visually impaired persons. For some standard picture sets, the choice of stimuli and their order are fixed (see chapters 12, 29, 32, and 33 in this volume). For others such as the Murray TAT, some clinicians prefer to choose the pictures that seem most relevant to each client’s situation, in order to sample the client’s responses to similar nontest circumstances that hopefully will generalize to that specific life situation. This line of thinking—generalizing to specific circumstances—is unfamiliar, even alien, to most researchers.

TAT data gathering permits flexible applications, but this strength must be balanced by efficiency. In fact, arguably it is the efficiency of structured scoring systems that makes the flexibility of TATs a strength. Like the Rorschach, TAT story protocols
Introduction

provide a literal embarrassment of richness that is daunting for students and can be exhausting for the expert. Hopefully the present volume will help to make the process more efficient than it was when Fosberg (1949) proposed a minimum one-year academic course in projective testing alone (to be supplemented by continuing education workshops), with three years of course work ideally recommended as adequate preparation for clinical assessment. Manualized structured scoring approaches permit efficient analysis of these useful data sources by giving systematic decision rules for identifying features of stories that are linked to specific theoretical constructs or characteristics of the storyteller. Even a novice clinical assessment student may be able to use TAT data to answer specific referral questions if an appropriate scoring system is available and the student has attained the interscorer reliability standard (facilitated perhaps by chapter 2 in this volume). This efficiency allows for better focused psychological services to more clients at lower cost.

This more structured and efficient approach to TATs makes possible the systematic study of clinically important processes, as valid story data can be gathered either by traditional individual clinical assessment or via group administration, and by nonclinicians (Smith et al., 1992). Pretenure faculty need not be discouraged from undertaking research programs using TATs, because student assistants can be trained to score large numbers of stories reliably (see chapter 2), thus enabling more rapid data analysis and publication. Furthermore, once the data are in hand, several scoring systems can be applied for different publications, provided only that the theoretically relevant correlate variable data have been gathered. This feature permits graduate students to join forces for economical cooperative data gathering, with each one scoring the TAT stories using a different system.

This efficiency is not costly to the data, either. Unlike manualized psychotherapy approaches, scoring manuals for assessment data are not intrusive. They do not change the data that have been gathered, they only assist focused and strategic data analysis to answer specific referral questions. New questions can be answered long after the data have been gathered by simply applying another scoring system. The only limit to the number of scores that can be extracted is the number of scoring systems currently available and the time available for using them. The only cost comes from the investment in learning what to look for in stories; that process is transparent to the storyteller.

Generalizability to storytellers’ lives is an important strength of TAT data, regardless of the analytic approach. Emerging basic psychological and cross-cultural research on narrative processes suggests that storytelling may be one crucial link between the individual’s intrapsychic functioning and the construction of a life in a society (e.g., De Vos & De Vos, 2004; LeVine, Strangman, & Unterberger, 1966; McAdams, 1997; McClelland, 1961; Veroff & Feld, 1970). How a person tells a story reflects a host of mediating processes that connect perceptions, thoughts, and feelings to actions, knitting together the psychological context and behavior in vivo.

Just how this works is not precisely clear as yet, but there is evidence (in this book and others cited previously) that these processes and the stories that illustrate them provide useful data that are quite different from the data yielded by most other forms of clinical assessment. If we consider stories to represent samples of the storyteller’s imaginative thought and inner experiences about situations similar to those in the pictures (McClelland, 1980), then it is only a small step of modest inference to assume that what the client says in the story resembles what the client might think or feel in similar real-life situations.

It is crucial to note—and is historically overlooked—that these experience samples might generalize well to thoughts in real life, but this does not mean that they will necessarily predict well
to actions in real life! There are a host of reasons why people do not do everything that they are capable of thinking or saying. In fact, one noted weakness of structured self-reports is that people do not always do what they say they do, or will do (Mischel, 1968). Structured scoring systems improve generalizability by enabling the scorer to identify and classify specific features of stories that can be connected to parallel inner experiences—and perhaps overt actions—in vivo. This parallelism can be tested directly using collaborative assessment models (Fischer, 1994) or therapeutic assessment interventions (Finn, 2007; Finn & Tonsager, 1992, 1997).

The consistency of analysis gained from structured TAT scoring, like its efficiency, is a unique benefit (over and above the richness of the unscored stories) that further amplifies the usefulness of the data. The consistent interpretation of story data begins with stable and precise application of the system’s scoring rules, as evaluated by calculating interscorer agreement or intrascorer agreement over time. Once the system has been learned, it can be applied to the same stories—or to different ones—by the same or different scorers, the same way now and next year. This consistency is attained by having a standard frame of reference for understanding stories so as to generalize the same observations across different storytellers, which makes systematic research, as well as systematic clinical use, possible.

The consistency of structured scoring systems enables two particularly helpful applications. First, admittedly at this writing some way in the future, the use of several scoring systems in combination can provide a consistent frame of reference for idiographic interpretation, much as the Rorschach Comprehensive System is now used. Second, this frame of reference might serve as a platform for comparison of clients’ scores with the clinician’s database of local norms and with existing research literature using samples from similar populations. Third, it enables evaluation of change over time in the individual client on the constructs of interest.

Consistency does not enforce uniformity, though. In-depth idiographic content analysis is not ruled out by structured scoring. It can be used along with specific scoring systems and can be enriched thereby. For example, suppose that a client has told three stories about achievement concerns (identified by the Psychocultural System) that involve characters in parent–child roles and negative affect (scored with the Affective Scale). The clinician might examine the specific nature of the concern and its apparent function in the relationship that produces negative affect, and in whom, for more precise discussion with the client.

Finally, although flexibility, efficiency, generalizability, and consistency are helpful in both clinical and research work, structured scoring systems allow for the objectivity that is the sine qua non of science in our field. This objectivity is exemplified first by the attainment of interscorer reliability between different observers applying the same scoring rules. Only individual scoring systems, not “the TAT,” can be evaluated for validity and reliability, making scientific evaluation of TAT data possible. The phrase validity of the TAT is meaningless, because validity is specific not to the pictures, but to the set of scores derived from the population, purpose, and circumstances involved in any given data collection, as discussed in this chapter. Furthermore, for TATs validity is specific to the method of obtaining scores from stories. (Curiously, it is conventional to call objective the structured self-report items commonly used in social and personality psychology to elicit respondents’ subjective perceptions and attitudes toward themselves and the world. The content of these items is certainly not objective in any meaningful way, as the experiences reported cannot be independently verified apart from the respondents’ introspection.)
Why This Book, This Way?

Numerous clinical scoring systems for TATs have been published in the last half century, but few have been available for use beyond the laboratories of the originators and their students. For reasons of length, journal articles typically include no more than a brief description of the scoring categories, which is not adequate for either clinicians or researchers to attain interscorer reliability. Many clinically useful systems are sufficiently complex and subtle to require for training a detailed scoring manual with several sets of stories to be scored for practice, along with expert scores and commentary on the more difficult ones. Working with these materials allows others—who may differ in theoretical orientation, clinical specialty, level of experience, institutional context, client population, and personality style—to replicate the scoring decision process accurately. Because such materials are rarely circulated widely, few of these systems have accumulated the amount of published empirical support that we would ideally like to see for clinical practice.

Furthermore, the more subtle the theoretical constructs, the more complex the system, and the higher its information yield, the more training material (and for some, such as the Adult Attachment Projective [AAP], supervised practice) is needed. A full range of examples and practice scoring materials helps the interested user learn the scoring procedures and contingencies of scoring decisions. A detailed, specific scoring manual provides the learner with the rule system for decision making that constitutes the statutory law that minimizes the leaps of inference that the scorer makes in the process of scoring decisions. The process of working through the practice scoring materials corresponds to the study of the case law, by which the scorer learns how to reason deductively from the rule system to its application across a range of specific instances. These instructive examples show the application of statutory principles and help the learner to make the occasional unavoidable inferences by internalizing a reasoning process similar to that used by the expert scorer. The use of a scoring group for establishing and maintaining interscorer reliability parallels other learning processes relying on the case-study method (Stenzel & Feeney, 1970), and is discussed in chapter 2.

The goals of this book are (1) to raise awareness about the availability and usefulness of TAT scoring systems for research, training, and clinical practice; (2) to provide the materials needed for learning and using the most useful available clinical systems; and (3) to facilitate their use by making independent learning and systematic research easier. The systems gathered here have the most extensive supportive research evidence among those for which the original authors or interested successors could be located, had preserved the needed materials, and were willing to contribute same to this volume. These systems and their training materials are presented for clinical use as well as to facilitate the development of the needed research literature. If this book does its job, it will enable such a wealth of research as to make itself outdated rapidly.

Each system is represented by a chapter that describes the system within its conceptual framework, summarizes the available research evidence, and suggests an agenda for future research that is designed to inspire graduate students and other researchers. Scoring manuals for most of the systems—those that are not proprietary to publishers and can be self-taught feasibly—are also included. Some have numerous practice stories with expert scoring, others only a few, depending on their availability.

The following sections set this work in its historical, cultural, and disciplinary context, first recounting a brief (and necessarily too general) history of structured scoring
systems; second, reviewing the challenges that developed with the scientist–practitioner
gap; third, considering from a clinician’s viewpoint the scientific status of the methods
historically called projective; and, finally, reviewing the plan of the book and introducing
the systems that follow.

What Led Up to It, and What Will Happen Next?

“Projective” Assessment and TAT Interpretations: A Brief History

TATs have been called an ambiguous stimulus on which practitioners and scientists of all
persuasions have projected their own assumptions, views, and preferences for how science
or practice should be done. In the early years of TATs, psychological science and practice
were better integrated than they are at present, and a variety of theoretical and methodolog-
ical approaches were used for story interpretation. Holt’s (1950) publication of Charles E.
Thompson’s 377-item bibliography on the TAT contains 80 that Holt identified as “principal
references”—many including or reviewing research—and 82 research and methodological
studies, along with 28 published case studies.

The “Projective Hypothesis” and Assumptions About Consciousness

Historically, the clinical use of TATs has been supported by Frank’s (1939) “projective
hypothesis”. Consistent with the psychoanalytic conceptualization of psychological
functioning, this hypothesis presented stories as reflecting the internalization (i.e.,
“taking in”, or introjection) of early experiences that are then externalized (i.e.,
“projected”) onto situations that evoke the original learning experience. As usually
presented in undergraduate psychology textbooks, this projection process, and often
the original learning situation itself, is presumptively unconscious, theoretically due
to defensive repression of traumatic memories that arouse extreme, intolerable anxiety
and other uncomfortable affects.

One unfortunate corollary of the theoretically unacceptable impulses and uncon-
scious processes involved in the projective hypothesis is that then by definition, in assess-
ment and psychotherapy the interpreter–analyst is presumably better able to understand
the client than is the client. This might best be illustrated by Murray’s (1943, p. 1) unfor-
tunate choice of words in the TAT manual:

Special value resides in [the TAT’s] power to expose the underlying inhibited tendencies
which the subject, or patient, is not willing to admit, or can not admit because he is uncon-
scious of them…. As a rule the subject leaves the test happily unaware that he has presented
the psychologist with what amounts to an X-Ray picture of his inner self.

Revolt by female clients against this one-down, almost Machiavellian attitude among mental
health professionals and the increased popularity in the 1970s and 1980s of feminist ther-
pists, who prioritized empowering the client, may have contributed to the subsequently
decreased popularity of assessment techniques that appeared to depend on the projective
hypothesis, as also happened with psychoanalysis, in part for similar reasons.
Rigor and Scoring Systems

To Murray’s credit, however, although he resisted systematic scoring of the sort presented here (Gieser & Stein, 1999), he also emphasized the importance of rigorous training and discipline for the interpreter:

Even an old hand at the game must rely on the same process—empathic intuition first and last, disentangled as far as possible from personal elements. No true scientist will scorn the use of a function which when properly disciplined is capable of yielding precise and pertinent information. Of course intuition alone is highly unreliable; what is required is a rigorously trained critical intuition… TAT stories offer boundless opportunities for the projection of one’s own complexes or one’s pet theories, and the amateur psychoanalyst who is disrespectful of solid facts is only too apt to make a fool of himself if, in interpreting the TAT, he gives free rein to his imagination. The future of the TAT hangs on the possibility of perfecting the interpreter … more than it does on perfecting the material. (Murray, 1943, p. 6. italics in original)

According to Piotrowski (1950), interpreters of “projective” tests quickly broadened that concept in their reading of data to include conscious, acceptable impulses and realistic, healthy adaptations as well (Bellak, 1954, p. 2). Indeed, Frank’s (1939) original description emphasized the meaning that storytellers give to the test stimuli, as well as their subsequent actions and emotions, as expressions of personality paralleling their typical responses in social situations. Contrary to the usual undergraduate textbook depiction of idiosyncratic TAT interpretation (Mihura & Alperin, 2000), a half century ago Piotrowski was moving to formalize Murray’s trained intuition, calling for “a dictionary and rules which would enable us to translate the test data into habitual modes of behavior in interpersonal relationships” (p. 102).

Within a decade, papers presenting systematic objective (Dana, 1959) scoring systems were being published. These included both research-focused (Aron, 1949; Weisskopf, 1950) and clinically oriented systems (e.g., Cox & Sargent, 1950; Eron, 1950; Holt, 1956; Libo, 1956, 1957; Shorr, 1948) and the first empirically derived content analysis research measures of human motives by McClelland and Atkinson (e.g., Atkinson, 1958; McClelland, Atkinson, Clark, & Lowell, 1953; McClelland, Clark, & Atkinson, 1949). Although individuals may have specialized in either academic or clinical work, familiarity with both uses of TATs, and mutual respect among these communities, seemed more the rule in those early years than at present.

What made these systems objective was not only the existence of documented decision rules for quantifying text material that minimized the level of inference exercised by the scorer, but also their simplicity, clarity, and concreteness under the logical positivist insistence on establishing the objectivity of observations by confirmation of hypotheses (e.g., Carnap, 1936, 1937). Confirmability required that anyone who applied correctly the rules required to make an observation should see the same phenomenon. Clear and concrete decision rules facilitated consistent scoring decisions by making the relevant material easier to identify in the stories despite differences in training, personality, and preferred theory among scorers.

Theory-based interpretation of TATs and other less structured measures demanded higher level inferences to deduce the meaning of the data from the principles of theory. Theory was needed to specify the conceptual structure to be imposed on the data to evaluate each theoretical construct. Unfortunately, not everyone made the same deductions
the same way from the same rules, even when they used the same theory, resulting in rather more variable interpretation that relied on the interpreter's understanding and application of theory. The master interpreters previously cited built their careers on their ability to connect intricate theoretical filigree to their patients' labyrinthine problems using the stories as a conduit for images and intuitions. Learning to do this well was generally a lifelong enterprise.

In the fertile period for scoring system development from 1950 to about 1980, authors defined their goals diversely. Some attempted comprehensive catalogs of all relevant features of story narratives that were exhaustive of the data, and quite likely the interpreter (Aron, 1949; Bellak, 1954; Fine, 1955; Friedman, 1957; Haworth, 1963, 1966; Shatin, 1953, 1955; Weisskopf, 1950). Others identified important aspects of personality that they expected would be captured in most stories (Dana, 1959; Eron, 1950). Several selected quite specific theoretical constructs that should be evident in both TATs and Rorschach protocols or other less structured measures, and developed scoring systems for both kinds of data (Harder, 1979; Holt, 1956; Masling, Rabie, & Blondheim, 1967; Singer & Wynne, 1966). Weisskopf's (1950) and Murstein's (1972) goal was to identify the specific characteristics of picture stimuli that encouraged richer, more informative stories for clinical use. Arnold (1962) developed a method for extracting the essence, or import, of each story for interpretation and then studying the sequence of imports across the series of stories. These different approaches rested on a variety of assumptions about the nature of story data as a representation of something about the person and as a tool for identifying problems and predicting behavior.

**The Scientist–Practitioner Gap**

Although there had always been partisan arguments, between 1970 and the present the early integration and mutual respect among clinical scientists and practitioners in psychology as a field deteriorated steadily. Fewer clinicians using TATs concerned themselves with research evidence; some of the master interpreters were actively hostile to the emerging structured scoring systems. Researchers and textbook writers began dismissing the less structured measures out of hand rather than contributing to building the necessary technical scaffolding. Among academic TAT researchers, psychometric questions were too often handled dismissively in favor of getting on to the interesting substantive questions, to the frustration of one writer in that tradition (Fleming, 1982).

Nevertheless, studies using human motivation scoring systems flourished following the publication of the first collection of research articles and scoring manuals for the big three motives—need for Achievement (n Ach), need for Affiliation (n Aff), and need for Power (n Pow; Atkinson, 1958), the precursor of the one by Smith (1992a). McClelland and Winter's (1969) major cross-cultural longitudinal study of business executives in India followed, as did studies using survey research methods (Veroff & Feld, 1970) and Winter's (1973) revision of the need for Power system (McClelland, 1975). Although these systems were originally empirically derived and atheoretical, accumulating findings led to the development of theory (Atkinson, 1981; Atkinson & Birch, 1970; McClelland, 1961, 1981, 1985a, 1989; Winter, 1973; Winter, John, Stewart, Klohnen, & Duncan, 1998), a psychometric literature (e.g., Lundy, 1985, 1988; McClelland, 1980; Smith, 1992b; Winter & Stewart, 1977), applications in business (Litwin & Stringer, 1968) and clinical psychology (Cutter, Boyatzis, & Clancy, 1977; McClelland, 1977; McClelland, Davis, Kalin, & Wanner, 1972), and computer models of theoretically expected patterns of storytelling (Atkinson, Bongort, & Price, 1977; Reuman, 1982).

Surprisingly, this literature is often unknown to present-day clinicians and to researchers in other traditions, one of many puzzles in the history of this interesting
method. This track record has typically gone unreferenced in clinical assessment texts and review articles. The present isolation of the clinical practitioner from the academic scientist communities regarding TATs has discouraged the growth of the kind of coherent research literature that clinical researchers have produced to support the Rorschach Comprehensive System over the past 30 years. Acknowledgement of the human motivation paradigm's relevance for clinical approaches to TATs is relatively recent but growing, as evidenced by the coverage of this literature in recent books (e.g., Cramer, 1996; Ganellen, 1996; Meyer, 2004; Teglasi, 1993) and journal articles reaching an audience of clinical psychologists (e.g., Blankenship et al., 2006; Pang & Schultheiss, 2005; Schultheiss & Brunstein, 2001).

The increasing cultural alienation between psychological scientists and practitioners just described has been a major social-structural barrier to the understanding and effective use of TATs. Research-oriented and clinically oriented psychologists appear to sort themselves out as early as graduate school (Weiskopf-Joelson, 1955), and communication between them thereafter is sparse (Zachar and Leong, 2000), so that their frames of reference diverge as they gain experience with contrasting tasks and data. One result is a lack of consensus on the best way to understand and use TATs that satisfies the requirements of both research and clinical work. The cultures of academic science and clinical practice are sustained by differences between both the functions of culture members in their daily work lives and the institutional structures that shape those functions. These structural and functional differences in turn support quite different value systems for what makes a measure a good one.

**Challenges From Classical Test Theory**

Differences of purpose, incompatible frames of reference, and resulting failures of communication have contributed to a narrowing of perspectives and hardening of positions on both sides of the gap, further restricting the knowledge bases in both communities about the best uses of TATs and similar measures. In their struggle for hegemony in the literature, proponents of each perspective have increasingly in the past two decades adopted a dismissive stance toward each other's goals and the tools used to attain them. As clinical researchers began to import into personality assessment the psychometric procedures of classical test theory designed for structured cognitive ability tests, misapplication of these tools to less structured measures generated hostile reviews based on classical test theory (e.g., Entwisle, 1972; Klinger, 1966) that centered on TAT scores' generally lower internal consistency reliability (across stories) and test–retest stability.

The historical evolution of the concept of reliability is a curiosity of our field beyond the scope of this book. However, understanding it is important for TATs because it has so often been a locus of criticism and contention. A too-brief review of this history reveals that the concept of reliability evolved in an effort to identify the sources of inconsistency in scores on structured ability tests. This inconsistency was attributed to random error, that is, fluctuation in scores (error of measurement) when the same person took the same test on repeated occasions, or when some items on the test did not correlate highly with the rest. In the case of TATs, which are less pervasive in American culture than are standardized cognitive ability tests, consistency across repeated administrations is of less practical interest and importance as a source of measurement error than is consistency across scorers, that is, interscorer reliability.

Some of the assumptions underlying the concept of reliability are peculiar to structured ability tests. The first is that the abstract entity called an *ability* is an unchanging
aspect of the person, constant over time and situations. The second is that whenever the person's score differs from this "true score", this mistake in the test's estimation of that ability reflects flaws in the test and not real changes in the person. This rules out a considerable amount of individuality in how people approach tests, not only habitually but also due to day-to-day fluctuations in motivation and priority setting.

Structured ability tests use a response format that structures examinees' tasks by imposing strict constraints on the examinee's behavior. One reason for this is to reduce random error, that is, to bring the test scores into conformity with these theoretical assumptions by ruling out non-test influences on scores. Considering that a test samples behavior in order to generalize to theoretically similar non-test behavior (Anastasi, 1988), scores on these and similar highly structured tests generalize well to situations that impose similar constraints (McClelland, 1980).

The concepts of test–retest and internal consistency reliability might generalize well from ability tests to highly structured personality tests that involve stable traits and impose similar constraints on the test taker. However, the assumptions underlying these concepts simply do not fit either the kinds of behavior sampled by less structured methods such as TATs, or the range of non-test behavior to which they are designed to generalize. Indeed, one problem of more structured measures is that they do not predict so well to the less constrained behaviors of most interest to most people, such as mental health, career choices, and success in life (McClelland, 1980; Mischel, 1968). But why would this be a surprise, given that statisticians know that restricting the range of a variable makes it mathematically less likely to correlate with other variables, all else equal? If restricted-range variables relate less well to other variables, why would restricted-range samples of behavior relate well to non-test, less restricted behavior?

This is especially not surprising, given that some of the restrictions imposed to sample behavior "reliably" create problems of their own. Different test takers respond to those constraints differently, but predictably—that is, they are consistent in giving patterns of answers that may have nothing to do with the construct being measured and everything to do with the shape of the constraint imposed by the test. These patterns include response sets that are specific to the design of item response formats (Block, 1965; Schwarz, 1999), personality variables such as social desirability and need for approval that relate to the content of items (Crowne & Marlowe, 1960), and response styles of other sorts (Meyer, 1999). This kind of measurement error is systematic—constant over time and situations, like an ability or personality trait—and is indistinguishable from the construct being measured without complex data analysis techniques such as structural equation modeling, which clinicians rarely use with their clients.

The structure and assumptions underlying most clinical TAT scoring systems do not fit the requirements for evaluating internal consistency reliability that are described by Streiner (2003). Neither do the theoretical properties of human motive scoring systems, in particular (Atkinson, 1981; McClelland, 1980; Smith, 1992b). However, appropriate scientific evaluation of our tools remains important for both good practice and good science. How might this best be done?

**Strengths of Integrating Science and Practice**

First, it is important to recognize that a true integration of science and practice requires adaptation on both sides, including learning each others’ languages, and change is not
Introduction

13

easier for clinicians and researchers than for clients. The main difficulty of these historical
oppositions is that they have obscured the potential value of a more balanced approach that
treats each data source within a framework that specifies its contribution to a comprehen-
sive view of the client. The beginnings of such a framework can be glimpsed in recent
work that clarifies the individual and combined contributions of the Rorschach Com-
prehensive System and the Minnesota Multiphasic Personality Inventory-2 (MMPI-2)
in psychological assessment (Gane len, 1996; Meyer, 1999; Meyer, Riethmiller, Brooks,
Benoit, & Handler, 2000). One important dimension of the framework involves appraising
the sources of systematic measurement error variance, such as format-specific response
sets and related biases (Block, 1965; Schwarz, 1999), and irrelevant personality variance
(e.g., social desirability, defensiveness, and other self-presentation biases).

How these combine for the individual client is important in interpretation when
the client’s Rorschach and MMPI-2 findings suggest contrasting conclusions (Gane len,
1996; Meyer, 1999; Meyer et al., 2000). Meyer showed that convergent validity is higher
when examinees have similar response styles on both kinds of measures (i.e., either
open on both or guarded on both) than when they respond differently to the different
methods. The same is likely true for TATs. Hopefully the multidimensional framework
that is beyond the scope of this book not only can provide for the incorporation of other
measures into both the larger conceptual picture and a comprehensive treatment plan-
ning strategy, but also can inspire the development of new measures in areas that are
currently underpopulated.

TATs, like the Rorschach, may be especially suitable for an integrated scientist–practi-
tioner approach because of the combined richness of their data and flexibility for various
rigorous interpretive strategies. Commitment to such an approach eventually compels
innovative and appropriate answers to psychometric questions like those described
already. By first gathering a sample of client data, then subjecting it to careful objective
scoring consistent with theory and validation studies, users of these techniques may
attain both the rich detail that clinicians find most useful and the rigor that scientists
require. Unfortunately, scant institutional support has been available for creating scoring
systems, and there are continuing widespread failures of textbooks and journal review
processes to distinguish between psychometric criteria appropriate for more structured
measurement approaches and those suitable for less structured measures. As a result,
there remain risks of both clinical misapplication and scientific misevaluation.

An integrated scientist–practitioner perspective necessarily must bring together the
quite different strengths of story protocol material in both its research and clinical uses.
Clinicians of various theoretical persuasions appreciate the richness and depth that a
story protocol contributes to our understanding of a client’s social world (Schorr, 1993).
Compared with self-report measures like personality inventories and symptom scales,
it is more difficult for the storyteller to manipulate the data because there is literally no
right answer or wrong story. A fortunate advantage of narrative protocols for research
purposes is that data gathered under one research plan can be scored long after the
fact from a range of quite different perspectives, an activity that I hope this book will
encourage.

But when it comes to the process of developing and validating a measure, TAT scor-
ing systems, like Rorschach Comprehensive System scores, are still confronted with a
monolithic psychometric frame of reference that defines the established normal science
paradigm (Kuhn, 1962) designed for more structured cognitive ability measures. The
rules derived from this vantage point historically have been misapplied to less structured
measures, which are best used for constructs having different theoretical properties,
sources of error variance, and fields of generalization than do the more structured ability measures. In the wake of recent attacks on the Rorschach Comprehensive System, it is clear that the scientific issues involved are broader than TATs. Thus, these issues require discussion elsewhere both as general principles for evaluating less structured measurement methods and as technical specifications unique to story protocols. A more appropriate frame of reference is needed that integrates clinical richness with scientific precision. The following section builds a case for what that frame of reference should look like.

**Toward Scientific Practice with TATs**

As is clear from Murray’s (1943) emphasis on the “true scientist” as previously cited and from the volumes of research since then, scientific practice has been a major rationale behind the use of TATs from the beginning. The need for scientific evidence is more urgent now. Parity of mental health services in relation to medical services is by no means secure in either public or private arenas, and both have become political footballs in recent elections. As decreased funding for mental health produces pressure for the deskilling of psychological services to masters-level practice, the more specialized and potentially cost-effective assessment tools of doctoral-level psychologists become less and less available to the general public, lowering the visibility of our practice to society and to third-party payers in a self-fulfilling prophecy. This is a greater hazard for less structured measures than for others because they are relatively time-intensive and individualized. The efficiency in training, research, and practice gained by structured scoring systems is essential to the effective use of TATs in managed-care environments, and improved empirical support for the use and interpretation of all assessment techniques is increasingly seen as important for ethical practice.

**Stories and Science**

One major cost to our functioning as a field of the gap between science and clinical practice is that scientific evaluation of our measures rarely resembles the ways in which those measures are best used in clinical practice. Psychological research proceeds normatively, but clinicians make decisions ipsatively. The prediction of the average behavior of groups on single variables is of little use to clinicians compared with the precise description of multiple interacting influences in dynamic tension within the individual. Clients’ actions in the midst of their lives are influenced more by the external and internal contexts of those actions than they are by comparisons with normative samples. Pictures and the internal narratives that they stimulate may be key to the development of the kind of person-situation interactional psychology proposed in the 1970s and 1980s (e.g., Endler and Magnusson, 1976; Magnusson, 1988), whether one conceptualizes the assessment process as the activation of internal object representations or as an analog behavioral observation.

Against the conventions of tests-and-measurement textbooks, the outline given to authors for this volume asked them to present the evidence for their system’s validity first, then to discuss its generalizability across populations, and finally its reliability. The priority of position given to reliability over validity in the chapter order of most textbooks is bewildering, given that so many of the classical psychometricians said clearly
that validity was the far more important characteristic for evaluating a measure (e.g., Anastasi, 1988). The view that follows integrates the perspective of academic scientists who study TATs with that of clinicians who use them to understand and help their clients. Hopefully, what emerges will enrich the tactics available to those in both roles.

Validity Considerations for TATs

The validity of a measure is nearly a misnomer, as validity is a property of the interpretation of a set of scores obtained from a specific sample of people gathered for a particular purpose under certain circumstances (Anastasi, 1988; Cureton, 1951; Shneidman, 1959). Thus, all clinicians share responsibility for determining the kind and degree of validity that each measure has for their clients. Content validity, however, is indeed a property of TAT scoring systems, not of the set of scores. A distinct advantage of content validity for TATs is that what appears in the stories is a literal sample of imaginative thought (McClelland, 1980), not a sign or a symbol—though these can be inferred from stories as from other human products. These samples can be understood by the clinician as representing special instances of the abstraction described in the scoring categories, and yet understood by the client as concrete examples of a personal recurring and problematic action pattern. Stories have face validity, once told and scored, even though the storytelling task imposes none when the client first confronts it.

Construct validity indicates in addition that the properties of the set of scores are consistent with the theoretical properties and nomological network of the construct being measured (Cronbach & Meehl, 1955). That is, if the construct is theoretically stable, such as an intellectual ability or personality trait, the scores should have high test–retest reliability, even over quite long intervals. Scores for measures of unstable constructs such as mood or state anxiety should not be so stable; high test–retest reliability for those scores (i.e., consistent scores for inconsistent constructs) is evidence that the measure is invalid. Scores for situation-specific constructs such as test anxiety or state-dependent memory should only be consistent when the person is in the specified situations and should not generalize to other situations, by definition (Anastasi, 1988).

McClelland (1980, 1981) and McClelland, Koestner, and Weinberger (1989) extended this reasoning about congruence between the measurement method, the construct, and the criterion behavior to be predicted to demonstrate construct validity. His studies emphasized the structural similarity between the measurement situation and the person's population of life situations to which generalization from the behavior sample—in this case, imaginative thought—is desired. Specifically, these studies showed that structured self-report measures tend to predict scores on other structured self-reports, whereas scores from less structured non-self-report measures (TAT stories) predict better to less structured situations, such as people's daily life activities and long-term outcomes.

In part, McClelland argued, this was due to the theoretical properties of the constructs being measured—stable cognitive schemas, attitudes, and abilities versus situationally activated motives—but an additional issue was the sources of error variance typical of each measurement method. That is, self-report measures are notoriously subject to social desirability and other systematic biases due to self-presentation and social comparison processes. They are also vulnerable to format-specific response biases coming from the response options that examinees are given (Block, 1965; Schwarz, 1999). These forms of systematic measurement error tend to be stable and to cause items and measures
to intercorrelate cross-sectionally and over time even when the constructs that they measure are unrelated (Campbell & Fiske, 1959). Those tendencies increase reliability, but at the expense of lowering validity.

In contrast, TAT scores for motives, theoretically and empirically being more situationally influenced, tend to differ across situations (yielding lower internal consistency reliability among stories) and over time (showing low test–retest reliability to the extent that major events and changes in people’s lives might influence their actual motives, as shown by Jenkins, 1987, 1994). Furthermore, storytellers tend to see their task as a demonstration of creativity, so they tend to avoid telling similar stories if they can (Lundy, 1985; Tomkins, 1947; Winter & Stewart, 1977). Nevertheless, these scores correlate well with the kinds of situations people choose to be in and what they do when there, even in longitudinal studies over more than a decade (Jenkins, 1987, 1994; McClelland, 1965, 1980, 1985b; Stewart, 1980), showing high predictive validity despite low reliability.

Because the lesser structure of TATs allows more variability in behavior generally, when compared with structured self-reports, one common source of systematic measurement method error variance that might dilute the validity of scores is simple response productivity: story length, which must be evaluated and perhaps controlled statistically (McClelland, 1980). Another might be story richness, as represented by Edith Weisskopf’s Transcendence Index (TI) herein, which is conceptually comparable to Lambda in the Rorschach Comprehensive System. For some scoring systems, telling longer, richer stories is likely to gain higher scores; storytellers who give shorter, more literally descriptive responses are likely to score lower. But in contrast to the measurement error common in structured self-reports, story length and richness might vary across stories and over time, to the extent that storytellers find some pictures more engaging than others, are more energized or fatigued at different times, or are otherwise affected by changes in their internal state. To the extent that this happens, unsystematically, what we would otherwise think of as systematic method error becomes random error, changing independently of the construct of interest. As such, it is likely to lower both internal consistency reliability and test–retest reliability, even when the construct being measured is relatively stable. However, this random error is cancelled out when per-story scores are aggregated into a summary score.

Local norms are often the best for valid clinical interpretation (Anastasi, 1988), as these will capture both ethnic and regional variations in idiomatic meanings, which are especially evident in orally given narrative data. For TATs, these norms should include not only scores, but also annotations to scoring manuals that indicate how common idioms were classified and other difficult scoring decisions were made, in order to ensure consistency of those decisions across clients and over time (see chapter 2). The ubiquity of computer use for practice management is a considerable asset for such record keeping, as it allows clinicians to maintain easily complete databases of their assessment scores from their clientele as local clinical scientists (Stricker, 2006).

Situation Specificity of TATs

A major strength of TATs that has been obscured by the often inappropriate psychometric emphasis on internal consistency and test–retest reliability is their usefulness for appraising the situation specificity of storytellers’ social perceptions and responses. Presenting a series of pictured situations to a storyteller is a way of sampling objectively and systematically, in a context standardized by administration instructions, from among the storyteller’s imaginative repertoire of possible social scenarios. To the extent that the pictures differ,
the stories and scores should differ. After all, storytellers understand their task as one of telling a different story about each picture. To the extent that they do so, the assessment yields a broader sample of behavior that should generalize to more of the storyteller’s various life contexts.

If the assessment clinician or researcher seeks more information about certain kinds of situations to answer a specific referral question or to test a specific hypothesis, then more pictures that sample that class of situations can be used. For example, pictures that might plausibly be a romantic couple sample the storyteller’s perceptions, ideation, and affect relevant to romantic situations; pictures of cross-generational dyads sample implicit experiences of parent–child relationships (see chapter 27 in this volume). Survey researchers often use this principle, called a stratified sampling procedure, to oversample particular population groups for certain purposes. Just as survey researchers might be interested in comparing storytellers’ responses to different populations of situations (as done explicitly by Thomas’s Affective Scale in chapter 27 in this volume), as well as in their more trait-like average responses across all situations. This information can be obtained easily by aggregating (i.e., adding or averaging) scores across all pictures to generalize to average (trait-like) behavior (e.g., “dangerousness”) or by aggregating only the scores from one kind of picture to generalize to only that class of situations (e.g., probability of thoughts related to violent childhood abuse, or to street fights).

Given that the set of pictures used for an assessment can be seen as a tool for systematic sampling of the storyteller’s range of life circumstances, how should that process be structured? What dimensions are most important? For general-purpose, comprehensive clinical assessment, three dimensions seem salient and generalizable across cultures as a sociological framework for organizing the human life course: social roles, as defined by age-gender combinations (Neugarten & Gutmann, 1958); activity context, as defined by characters’ focusing on people or on impersonal tasks; and emotional tone, as defined by the facial expressions and postures of the characters. Classification by these or other dimensions allows for comparing the storyteller’s responses to different kinds of situations, as well as comparisons across individuals and studies.

### Population Generalizability

One important aspect of any psychological frame of reference is its cross-population generalizability. Recognizing the importance of validity generalization across populations, this issue is discussed in each scoring system’s summary chapter in this volume. Individual psychosocial functioning, if adaptive, is shaped by its sociocultural context, and one culture’s healthy adaptation might be another’s insanity. In a rapidly shrinking world, generalizable measurement techniques are increasingly desirable, but the challenges of linguistic and cultural translation and cross-cultural norming are just beginning to be appreciated fully (Dana, 1993, 1998, 2005; Helms, 1992).

Clinical TATs in the United States have a 50-year history of accumulated cross-cultural knowledge via such authors as Richard Dana, George De Vos, David Ephraim, and William Henry. The need for racially and culturally specific stimuli was recognized early on, as described by Michelle Hoy-Watkins and Valata Jenkins-Monroe in this volume’s chapter 33. Although debates continue about the culture-specific details of picture cues and identification with one main character in the story, and about the influence of cultural
differences in values, beliefs, and symbols, it remains that nearly all cultures use storytelling to socialize their children (Howard, 1991), whereas few cultures use Likert scales or multiple-choice questions in that capacity. Thus, what TATs require participants to do is likely to be familiar, even enjoyable, to them, in contrast to most of our multiscale personality inventories. Although cross-cultural generalizability of results is a property of the specific picture cues and interpretive system used, few individuals in most cultures will have difficulty with the basic data-gathering task of telling a story about people in pictures.

Reliability

Many of the classical psychometricians observed that less structured personality measures are clinically valuable, but that classical test theory is not appropriate for appraising their validity or reliability (e.g., Anastasi, 1988; Guilford, 1954; Harrison & Rotter, 1945). They framed clearly two basic assumptions underlying all psychological measures that, if taken seriously for TATs, serve as an adequate response to the misapplications noted in the literature.

1. The ways in which the validity and reliability of measures are evaluated must be consistent with the theoretical properties of the constructs measured by the scoring system (e.g., measures of transient states such as mood should not show stability across situations and over time; wide-band constructs with fewer redundant items will have lower internal consistency reliability).

2. In contrast to the priority implied by the order of chapters in most measurement textbooks—including Anastasi (1988)—validity should take precedence over reliability (Anastasi, 1988; Guilford, 1946; Gulliksen, 1950; Karon, 1966, 1981; Thorndike, 1985).

McClelland (1980) added a third: The ways in which the validity and reliability of measures are evaluated must consider the measure’s sources of error variance, both stable and unstable, and whether the criterion correlate variables share the same sources of systematic measurement method error variance. Psychometric work on TATs not only requires bridging the scientist–practitioner gap; on the academic side it exists in the gap, as some academic personality researchers have never fully accepted even the human motivation paradigm. One challenge on this side is that the human motivation tradition’s use of story data makes it appear “too clinical”, and therefore scientifically suspect to those who equate science with structure. The use of formal scoring systems has not been well understood, in part due to its omission from or misrepresentation in measurement texts. For example, although Anastasi (1988) speaks well of both Exner’s (1978, 1986) then-emerging Rorschach Comprehensive System and McClelland (1985a) and Atkinson’s (1958, 1981) work on need for Achievement as promising scientific developments, this brief acknowledgement is a belated exception to its context in the book, a general damnation of projective techniques. Thus, her positive, albeit cautious, view of these systems is apparently often overlooked.

There have been occasional attempts at rapprochement, however, with efforts on both sides of the growing scientist–practitioner gap to hear and address the assumptions of the other’s frame of reference and to develop new constructs and language reflecting
that conceptual adaptation. On the more clinical side—but writing for research audiences as well—Karon (1966, 1981) distinguished the kinds of validity and reliability important for projective measures, as did McClelland (1980) for the research use of human motivation scoring systems.

To make the methodological differences and their implications more salient, McClelland has repeatedly offered concise labels for these paradigms that draw attention to the methodological reasons for the low correlations between story-based measures of motives and structured self-report measures of similar-sounding constructs. He first called these approaches operant and respondent, respectively, in part to recast in behavioral terms the theoretical assumptions attending the projective hypothesis (McClelland, 1980). Then he presented findings contrasting his measure of the achievement motive with measures of achievement values such as the Edwards Personal Preference Schedule (McClelland, 1985a) and did the same for affiliation (McClelland, 1989). Finally, McClelland et al.’s (1989) empirical illustration of the different properties of implicit (i.e., measured in stories) and explicit or self-attributed (i.e., structured self-report) motives seems to have recruited the attention of the academic science community (e.g., Baumann, Kaschel, & Kuhl, 2005; Brunstein & Maier, 2005; Schultheiss & Brunstein, 1999, 2001; Schultheiss et al., 2005; Spangler, 1992; Woike, 1995; Woike, Lavezzary, & Barsky, 2001) and a smaller but well-respected group of clinical researchers (e.g., Bornstein, 1998).

Returning to the goals of the original psychometricians, the purpose of appraising reliability is to identify the amount of error variance in a measure’s scores. But classical test theory considers only random error, and does not evaluate systematic measurement method error variance (Campbell, 1996), so it gives only a partial picture, even for the purpose for which it was originally developed. Because they impose less restriction on the range of the storyteller’s responses, TATs and other less structured measures may have less built-in stable measurement error than do more structured measures of traits and abilities. Thus, they may be more sensitive to changes in the person across time and situations, and consequently can be especially useful for measuring less stable constructs such as mood, as well as other kinds of situation-specific responses such as motives and interpersonal scenarios. As already described, TATs in particular can be used flexibly to sample a broad or narrow range of situations represented in the stimuli used to generate stories, so that the results might better generalize to a broad or narrow range of situations in the storyteller’s life.

This does not mean that reliability is unimportant in evaluating TATs. The most crucial source of random error threatening all less structured measures is inconsistency in scoring. Attaining and maintaining interscorer reliability is thus a central issue in evaluating TAT scoring systems, important enough to merit its own chapter in this book (see chapter 2).

This does not mean that other forms of reliability are unattainable for TATs, given that the construct measured is narrow-band, theoretically stable, and trait-like (i.e., consistent across situations), or that the sample of stories is large enough, due to the influence of number of items on the calculation of Cronbach’s alpha (Streiner, 2003). Dana’s (1959) three scales presented in chapter 7 measure narrow-band, trait-like perceptual constructs, and these show high internal consistency reliability with only five stories. By careful matching of theoretically appropriate constructs and methods with precisely chosen psychometric challenges, Hibbard, Mitchell, and Porcerelli (2001) produced a landmark paper on TAT internal consistency reliability, showing that samples as large as 10 stories show high internal consistency reliability for the much more complex, wider-band Social Cognition and Object Relations Scales (SCORS; Westen, Lohr, Silk, Gold, & Kerber, 1990), whereas smaller samples do not.
A different approach to increasing the number of items is to obtain more “items” from the same number of stories. Using a computerized administration technique for which students typed their own stories, Blankenship and Zoota (1998) increased the internal consistency reliability for need for Power, and Blankenship et al. (2006) for need for Achievement, by separating the traditional four aspects of a story requested in the administration instructions (tersely: what is happening, what led up to it, what are people thinking and feeling, how will it come out) into separate questions, each with its own paragraph. Thus, five “stories” yielded 20 “items”. However, this approach is generally inappropriate for clinical use because it tends to disrupt the narrative coherence needed for clinical purposes and for some other scoring systems. The better response to the classical test theory challenge of internal consistency reliability for clinical scoring systems is to gather a large enough sample of stories, 10 apparently being a reasonable minimum for a broad-based assessment. As discussed previously, though, the minimum number of stories needed may depend on the kinds of situations in the client’s life to which generalization is wanted, and the number of stimuli needed to give an adequate sampling of those situations for the purpose at hand.

More recent responses to the reliability challenge have bypassed classical test theory in favor of item response theory (IRT; Tuerlinckx, DeBoeck, & Lens, 2002) and the Rasch model (Blankenship et al., 2006). These studies have applied these forms of item analysis toward the goal of identifying and selecting pictures that evoke, or pull for, more need for Achievement imagery than other pictures as a way of reducing variability of scores between pictures. They are valuable for having demonstrated solutions-in-principle to the internal consistency reliability problem for TATs by increasing the structure of the stimulus situations in the pictures. It is unclear whether these solutions will apply to other scoring systems, as IRT and the Rasch model both were designed for ability-like constructs that are presumed stable over time and situations rather than being transient states or situation-specific responses. Furthermore, for clinical purposes it is usually desirable to choose pictures of different kinds of situations and people, rather than a more structurally homogeneous set, in order to generalize the findings to a wider variety of the client’s life circumstances.

Meeting the Challenges: Goals of the Book

This book was written for both clinical practice and academic research communities. It is designed to make the most useful available clinical scoring systems accessible to potential users, providing scientifically and clinically promising tools for the improved use of TATs.

Plan of the Book

Chapter 2 presents a strategy for learning to score structured scoring systems that is suitable both for individuals who are using a manual for self-teaching and for supervised research teams. This chapter grew from my work with undergraduate scorers that in our lab has enabled them to attain and maintain high interscorer reliability standards even with quite complex systems. The Appendix to that chapter is designed as a handout for scorer trainees; the bulk of the chapter is directed to the supervisor. It covers a detailed description of the learning process and its challenges, along with suggestions for using
the group's dynamics to improve the accuracy and efficiency of the learner's developing competence with scoring. Chapter 3 extends the contents of the present volume toward the future, supplementing the research recommendations offered by authors for their specific systems and noting existing systems not included here that merit further study. It proposes more general avenues for conceptual and methodological developments, discussing the kinds of research and clinical problems for which TATs are most useful. Students and early-career researchers hopefully will find there some helpful inspiration, ideas, and strategies for theses and dissertations using these methods.

The remaining chapters present the research summaries and scoring manuals. An overview chapter describes each system, covering its conceptual framework; theoretical, clinical, or empirical basis and assumptions; scoring categories; research evidence; an optional illustrative case; and suggestions for future research that are meant to inspire theses and dissertations. Most are followed by a detailed scoring manual. Many of these include stories that readers can score for practice, along with expert scoring so that readers can check their scoring accuracy.

**Systems Represented Here**

Few of the systems represented here were located by electronic search, as the indexing keywords used were either too general or too specific, and searchable abstracts appeared only in the mid 1970s. I began with those references cited by Cramer (1991, 1996), Teglasi (1993), and Zubin, Eron, and Schumer (1965) that had promising titles. Early consultation with Phebe Cramer and with Richard Dana was also helpful in locating both specific systems and promising search strategies. When it became clear that the most likely journal sources were the *Journal of Personality Assessment* and its ancestors, the *Journal of Clinical Psychology*, and the *Journal of Consulting and Clinical Psychology*, I spent much of my faculty development leave during 2000–2001 in the stacks of the W.E.B. Dubois Library at the University of Massachusetts at Amherst hand searching the tables of contents for these journals—and occasionally others—back to 1945, looking for suggestive titles and review articles. Stricker and Healy's (1990) and Vane's (1981) reviews were notably helpful. A few other useful leads came from colleagues, conference presentations, and Internet discussion lists.

My criteria for pursuing a system further included (1) evidence of the existence of an adequate scoring manual, typically a citation to same in the article or presentation; (2) evidence of interscorer reliability; (3) empirical evidence for validity or clinical utility in samples with probable generalizability; (4) importance, promise, and clarity of the theoretical constructs measured; (5) feasibility of learning to score largely through self-teaching with the manual (i.e., not dependent on extensive knowledge of a particular theory); and (6) recommendation by a colleague. Not all systems met all criteria. Unfortunately, it was necessary to omit systems not in English, which eliminated Shentoub's (1990) highly recommended system.

The systems are organized in the table of contents loosely according to theoretical orientation, with these in a rough historical sequence, the earliest publication first. Similarly, they are ordered chronologically within groups. In the section that follows, they are described in historical order of publication, without subgrouping. Table 1.1 summarizes in crude categories each manual's theoretical foundation, how the scores are derived (which is useful for conceptualizing the distributions of scores and their statistical properties, and
<table>
<thead>
<tr>
<th>System</th>
<th>Theoretical Foundation</th>
<th>Nature of Scores</th>
<th>Main Validity Type</th>
<th>Best Referral Question(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Diagnosis</td>
<td>Prognosis</td>
</tr>
<tr>
<td><strong>Perceptual-Cognitive</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcendence Index</td>
<td>Eclectic</td>
<td>Content Analysis</td>
<td>Content</td>
<td></td>
</tr>
<tr>
<td>Objective TAT Scoring</td>
<td>Eclectic</td>
<td>Content Analysis</td>
<td>Criterion</td>
<td>✓</td>
</tr>
<tr>
<td>Interpersonal Decentering</td>
<td>Piaget</td>
<td>Two-step</td>
<td>Content Criterion</td>
<td></td>
</tr>
<tr>
<td>Personal Problem Solving</td>
<td>Cognitive behavioral</td>
<td>Rating</td>
<td>Content Criterion</td>
<td></td>
</tr>
<tr>
<td>Picture Projection Test 1</td>
<td>Eclectic</td>
<td>Mixed</td>
<td>Criterion</td>
<td></td>
</tr>
<tr>
<td><strong>Psychodynamic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fine's Scoring Scheme</td>
<td>Psycho-dynamic Mixed</td>
<td>Content Criterion</td>
<td>Content</td>
<td>✓</td>
</tr>
<tr>
<td>Primary Process</td>
<td>Psycho-dynamic Mixed</td>
<td>Content Construct2</td>
<td>Content Construct</td>
<td></td>
</tr>
<tr>
<td>Pathogenesis Oral</td>
<td>Psycho-dynamic Two-step</td>
<td>Content Construct 2</td>
<td>Content Construct</td>
<td></td>
</tr>
<tr>
<td>Dependency</td>
<td>Psycho-dynamic Analysis</td>
<td>Content Construct 2</td>
<td>Content Construct</td>
<td></td>
</tr>
<tr>
<td>Symbiosis</td>
<td>Psycho-dynamic Two-step</td>
<td></td>
<td>Content Criterion</td>
<td></td>
</tr>
<tr>
<td>Ambitious-Narcissistic Style</td>
<td>Psychodynamic</td>
<td>Content Construct²</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>--------------</td>
<td>--------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social-Emotional</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affective Scale</td>
<td>Two-step</td>
<td>Content Criterion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpersonal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empathy</td>
<td>Two-step</td>
<td>Rating Content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contemporized Themes</td>
<td></td>
<td>Rating Content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concerning Blacks</td>
<td></td>
<td>Rating Content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Needs-Concerns Focused</td>
<td></td>
<td>Rating Content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychocultural Scoring</td>
<td></td>
<td>Rating Content</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anthropology, Motivation</td>
<td></td>
<td>Rating Content</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

³ Has its own picture set.

² Rorschach scoring manual also published.

Note: Nature of scores: Content analysis (identified content scored present-absent), Rating (global ratings of entire story), Two-step (first identify content if present, then score or rate it), or Mixed (some combination of these for different scores). Main validity type: Content (congruence of category descriptions with construct and/or behavior represented), Criterion (scores differed between theoretically relevant groups or correlated with other theoretically relevant measures), or Construct (Criterion, plus scores converged with measures of the same construct using other measurement methods). Diagnosis = Utility for identifying one or more Axis I or II diagnostic criteria; none are comprehensive. Prognosis = Utility for estimating prognosis under different situational conditions. Treatment planning = Utility for suggesting likely response to various treatment alternatives and interventions. General Social, Close Relationships = Utility for identifying client’s strengths and weaknesses in general social interaction and close relationships, respectively.
the types of validity evidence that are available thus far. The likely utility of each system for addressing different aspects of referral questions is also indicated crudely. This classification is impressionistic and not necessarily well supported by current validity evidence, except in the case of diagnosis; a check in that column indicates that at least one validational study differentiated clinical groups significantly or showed convergence with other clinically validated scales as expected. For the other four referral question categories, classification is based on the content or criterion validity of the measure and the likely utility of that information for estimating prognosis under different situational conditions, suggesting likely response to various treatment alternatives and interventions, or identifying client’s strengths and weaknesses in general social interaction and in close relationships, respectively.

The first reasonably detailed published scoring manual found to include interscorer reliability data is Weisskopf’s Transcendence Index. It was not designed for scoring stories, but rather for identifying ways in which people who are asked simply to describe a picture “transcend the stimulus” spontaneously, for example, by attributing thoughts, feelings, abilities, and physiological characteristics to the characters; by going beyond the picture in time or space; by specifying social relations such as kinship, occupational roles, or possession; or by making editorial comments such as imperatives, evaluation, or noting the atmosphere or symbolism of the picture. Weisskopf’s primary use of the system was to compare the usefulness of different stimuli and administration conditions for eliciting more transcendent descriptions, and thus presumably richer, more informative stories for clinical application.

One long-standing question about the validity of TAT data has been, “What constitutes an adequate story for interpretation?” Although even very brief stories can sometimes be clinically useful, more often they are not, and it is scientifically silly to say only that short stories are invalid unless something valid is obvious. Regarding the storytelling process as a process of thought sampling highlights this question, parallel to population sampling in survey research, because the validity of data rests in part on the validity of the sampling process. Stories with little transcendence do not provide a trustworthy sample of thoughts for interpretation beyond the observation of apparent constriction in responding.

I think that the TI is a good candidate for a TAT validity index comparable to Lambda in the Rorschach Comprehensive System. A person who does not give much more than a literal description when asked to tell a story about a picture is simply not giving enough information—regardless of the reason—for the stories to be useful for more than individualized appraisal. If so, then eliminating uninformative stories, or correcting for TI scores along with story length, in research may well improve the quality of our results. Although the various scoring categories have not been studied individually, their face validity for the most common content domains of transcendence makes them worthy of study for their potential generalizability to clients’ repetitive patterns of social behavior—and deficits thereof.

Fine’s Scoring Scheme measures the manifest content of the story characters’ feelings and outcomes, scored as either present or absent in the story, and of interpersonal dyads, scored as to whether the latter are moving toward, away from, or against one another. These classifications are quite detailed, with 36 subcategories for feelings and outcomes, and with 16 dyad types that include distinctions between initiating members of asymmetric dyads (e.g., separating mother’s movement with respect to child from child’s movement with respect to mother). The comprehensiveness of this system should make it useful for a variety of purposes, and the concrete nature of the categories should make it easy to score.

The Primary Process System scores content that meets criteria for this classical Freudian construct by reflecting attention to libidinal or aggressive themes, and also
scores formal aspects of stories that suggest defenses against or control of these contents. Manifest content is identified and scored in the story, and in some cases is then given an additional score for severity. Although high scores might indicate psychopathology—especially if uncontrolled—adaptive regression in the service of the ego (ARISE) might also yield notable scores. This is one of several systems developed during an era in which it was common to create in the same article parallel scoring systems for the same construct using two or more measurement methods so as to have validational associations with a parallel system scored from other less structured data—in this case Rorschach Ink Blot protocols. Like the Oral Dependency Scale, the Rorschach version of Primary Process has more extensive research support, but there is also some evidence of TAT validity.

Dana’s Objective TAT Scoring System was designed to capture aspects of“The psychological relatedness of person and environment,” specifically the storyteller’s approach to the assessment situation (i.e., following directions to include specific information in stories), the similarity of responses to those of nonpatient samples, and the rarity of responses that are more common among patients. In contrast to some of the more intricate, exhaustive, and multifaceted systems of the time, such as Reuben Fine’s and Robert Holt’s, Dana’s was designed to be of broad theoretical and practical importance, but to minimize inference so as to be simple enough to be scored by a clerk. These three scales discriminated between contemporaneously defined “normal”, neurotic, and psychotic persons. Like Weisskopf’s TI, Dana’s system has been used in numerous unpublished dissertations that were necessarily beyond the scope of the present review, but that might be informative if pursued.

Singer and Wynne’s Communication Deviance (CD) System is based on clinical observations of families in which CD in parents’ stories has been related to the later development of schizophrenia spectrum disorders in their children. Rather than identifying content as do most other systems, CD captures the structure of communication and problems of integration in language. These include closure problems, disruptive behavior, peculiar perceptions and verbalizations, and other formal characteristics. This system has produced a relatively large number of publications among the systems herein. These are primarily from large-scale family studies, mostly from the University of California–Los Angeles (UCLA) Family High Risk Project. Benjamin Chapman’s very thorough review of this literature in the present chapter 26 traces the system’s history from the original 1966 publication, based on National Institute of Mental Health (NIMH) family studies that included both TAT and Rorschach manuals, to its use in the UCLA studies. The use of Singer and Wynne’s evolving theory about the role of parental communication in the development of schizophrenia, plus the availability of clinical and research observations of the NIMH families, makes this system both theoretically and empirically derived. Supportive findings have come from Spanish-speaking and Norwegian samples as well as from studies primarily of Anglo Americans.

The authors of the current highly developed scoring manual declined permission to publish it here. Although no reason was given, one plausible rationale might be the possible sensitivity of its implications for parental functioning in the event of child custody litigation. The original manual is presented in Singer and Wynne’s 1966 publication along with its Rorschach counterpart, but it lacks the specificity of the one that has accrued the most empirical support.

Pathogenesis is based on clinical observations of the unconscious defenses that may be activated within different relationships. The content-valid whole-story score involves first locating instances in which a dominant and a dependent character are interacting in a situation in which their needs might conflict. Each such instance is then scored as to whether or not the dominant character takes the other’s needs into account, the latter being the pathogenic
response. Despite the theoretical emphasis on the unconscious, the content of the thought sample clearly depicts behavior that, if acted out, could well have the consequences that have been supported by research. Although it originated with studies of families having a child with schizophrenia, research has shown its validity for parents who abused their children, parents having a child who has become delinquent, and for distinguishing therapists who are or are not helpful. Like CD, its closest relative here (but very different in what is scored), and like Symbiosis, the Affective Scale, Interpersonal Decentering, and parts of Fine's Scoring Scheme, Pathogenesis is highly interpersonal and arguably has its best use in the study of dyads.

The TAT Oral Dependency Scale identifies content associated with this Freudian construct and is quite simple to score. Some categories involve only recognition of relevant words or phrases (e.g., parental figures; help; helplessness, loneliness, or depression; food- and mouth-related references; belief in luck or magic), and others need only minimal inferences (e.g., passive dependency, overly optimistic endings). The Rorschach version (Bornstein, 1996) has received considerable research support, but the TAT system has had much less published research attention. Like hostility and aggression, dependency is not socially desirable in mainstream American culture, and more literal representation in stories might have complex associations with overt behavior. Future studies using this system should examine the relative usefulness of the scoring categories separately until it is clear which yield the best data for which populations. The success of the Rorschach system, and the usefulness of manifest story content as a representation that likely generalizes to clients' real lives in individualized interpretation, make this system worth further study and development.

Melvin Feffer's Interpersonal Decentering is simple in structure but requires making careful distinctions. Scoring this two-step system involves first isolating instances of interaction between story characters, and then rating the cognitive complexity of each as undifferentiated, sequentially coordinated (i.e., action–reaction statements without internalized states, at three levels of complexity), or simultaneously coordinated (i.e., reflecting a character's internal awareness of another, at five levels of complexity). In the original study, it was validated against Feffer's better-known Role-Taking Task, as described in chapter 8. It provides a content-valid sample of how the storyteller conceptualizes relationships; the most typical score calculation method represents the balance between action–reaction and internalized interactions.

The Psychocultural System identifies the substantive content with which cultures and individuals are most likely to be preoccupied. Although some of its categories imply the pursuit of goals (e.g., Achievement, Affiliation, Esteem, Mastery, Power), others imply ongoing engagement in social processes (e.g., Competition, Harmony, Nurturance, Responsibility, Solidarity). It is perhaps the most thorough and exhaustive extant representation of storytellers' concerns and preoccupations among all those that have been designed to measure one or more of Murray's (1938) needs and presses. However, this system is more cosmopolitan in that it was created by psychological anthropologist George De Vos for cross-cultural studies, and thus includes more categories that would concern members of collectivist cultures. Some of the categories appear to parallel the human motivation systems for needs for Achievement (McClelland et al., 1953; McClelland & Koestner, 1992), Affiliation (Heyns, Veroff, & Atkinson, 1958; Koestner & McClelland, 1992), and Power (Winter, 1973, 1992). However, the Psychocultural System scores only the presence of imagery, and not the subcategories reflecting an imagined sequence of goal pursuit that characterize the human motive systems. One beneficial connection to theory is that David Ephraim organizes these concerns along the two major axes...
of agency and communion, also called dominance–submission and love/affiliation–hate/separation, that have become prominent higher-order factors for interpersonal relationships in recent decades (e.g., Alden, Wiggins, & Pincus, 1990; Benjamin, 1993; Leary, 1957).

The Symbiosis System appraises the kind and degree of self-other differentiation problems between story characters, because a deficient awareness of people as distinct individuals is important in several character pathologies. Symbiotic attachment is scored by first locating representations of dyads, as for Fine’s scoring scheme, and then scoring the presence or absence of each form of symbiosis according to the direction of attachment. Thus, problems of psychological merger, intrusiveness, negative reactions to separation, jealous reactions, dependency, and attempts to control the other can be seen as diffused among relationship types or localized in just one. The validational family studies have shown that symbiosis scores differentiate among patients, and also in parallel, among parents of patients.

The Ambitious–Narcissistic Style Scale identifies indicators of this character style using five theory-based content-valid categories: Intrusiveness/Thrusting, Exhibitionistic/Voyeuristic, Urethral Excitation, Mastery/Competence/Power, and Self-Potency. Each, if present in the story, is then rated at one of three intensity levels. Oedipal and autonomy concerns are also rated, primarily to assist in differentiating these from the other categories. Most of the extant theory and research is on men, though there is a little evidence for women. The single published study is especially strong, showing both convergent validity with parallel scores from Rorschach and early memories protocols and criterion discrimination between men judged by clinical raters as having or not having this style.

The Affective Scale requires first identifying the role relationships in a story and then rating each on a four-point scale of positive versus negative affect. It developed in a classical fashion from clinical observations that led to qualitative data gathering, which in turn was formalized into a content-analysis scoring system. Its simplicity comes from its reliance on formally defined role relationships among story characters, the central assumption being that perceptions of relationships among story characters are drawn from—and thus generalize to—the role relationships in the storyteller’s life. Thus, married partners’ depictions of married story characters represent their own implicit definitions of marital roles, including the assumptions, attitudes, and actions that they expect from each other and from others toward their spouse and their couplehood. The validational study showed that couples whose spouse representations were more positive than their representations of parents had better dyadic adjustment than did couples whose parent representations were rated more positively than their spousal dyads. However, Thomas wisely provided for the generalization of her approach to a variety of relationships that might be of concern to different populations. There is no obvious reason why it should not be applicable to members of cultures that operate under differently structured kinship systems.

The Tell-Me-A-Story (TEMAS), designed for children and adolescents, is a particularly important innovation for its picture set, which not only reflects people with racial minority features in culturally familiar urban settings but also includes a nonminority version. The design of both the pictures and the scoring system is well grounded in the research history of other TATs regarding use of color and moderate to low levels of ambiguity compared with the Murray TAT. The setting, postures, and facial expressions of the characters suggest problems to be solved, sometimes with alternative courses of action suggested. The theory-grounded scoring system was designed to capture personality functions drawn from ego psychology, along with a host of descriptive indicators for clinical use. An up-to-date review of TEMAS literature appears in Costantino, Dana, and
Handbook of Clinical Scoring Systems for Thematic Apperceptive Techniques

Malgady (2007). The published picture set and the manual are available from Western Psychological Services.

The Need–Threat Analysis for the Children’s Apperception Test (CAT) identifies the major sources of stress perceived by children, which typically revolve around getting needs met or the threat posed by deprivation. Five content-valid need–threat binaries are scored present or absent in each story: Independence–Domination; Affiliation–Rejection; Security–Insecurity; Achievement–Failure; and Aggression–Punishment. A sixth category for “other” may be added. A computerized content analysis dictionary has been developed to assist interscorer reliability, but the story context must be considered to validate each score. Because of the emphasis on content validity to inform treatment interventions, the major research presents frequency norms for each of the 10 CAT cards. There is scant evidence of gender or age differences in scores. This system seems likely to be valid for adults and for application to stories elicited by other picture stimuli as well.

The Personal Problem-Solving Scale-Revised uses global rating scales for structural features of stories to provide a comprehensive picture of each story as presenting a problem to be solved (implicit in the plot) and one or more solutions. The subscales of the Story Design category indicate the structural aspects of problem appraisal, those of Story Orientation describe approaches to problem solving, the number of solutions is simply counted, and the components of Story Resolution identify different aspects of success. This cognitive behavioral framework is explicitly a content-valid translation to story data of several self-report problem-solving scales that have yielded few significant findings, exemplifying the gaps among what people say, what they think, and what they actually do. The system is very efficient; the several supportive studies thus far have used only three stories.

The Adult Attachment Projective (AAP) is a new and promising method that, like the TEMAS, the Contemporized Themes Concerning Blacks, and the Picture Projection Test, uses new stimuli developed to sample construct-relevant experience, in this case involving interpersonal attachment. The eight AAP stimuli are simple black line drawings, given in a fixed sequence, representing increasingly stressful situations that evoke attachment concerns for most people. The resulting scores include classification of the storyteller as Secure-Autonomous, Dismissing, Preoccupied, or Unresolved. Supportive findings have included native speakers of French and German, and several studies based in other cultures are in progress. The scoring system is well grounded in attachment theory, and requires considerable in-person training and certification for attainment of interrater reliability; thus, no manual is included in this volume. Interested readers should contact Dr. Carol George at Mills College (george@mills.edu) for information about scheduled scoring workshops.

The Empathy System has only three rating scales that are applied globally to all stories, but like those of Westen et al.’s (1990) SCORS, the distinctions between levels must be made precisely and are facilitated by sophistication with theory. The four levels of Perceptual Integration identify the accuracy of the storyteller’s perception of emotional tensions shown in the picture and the congruence between the latter and the story. The four levels of Affect Source distinguish increasing sophistication of coordination between internal and external sources of affect for story characters. The five levels of Mutuality of Autonomy capture the differentiation of individual characters’ internal worlds and the coordination among different characters, consistent with their internalized values and concern for others. It is well grounded in developmental theory and in the authors’ theoretical analysis and integration of the research literature on prosocial behavior more broadly. The authors have built on and adapted features of several previous scoring systems, including one for Rorschach Ink Blots protocols (Urist’s 1977 Mutuality of Autonomy),
strengthening the carefully explained scoring categories. The few studies to date are nevertheless very strong, with the major validational study relating Empathy scores to teachers' ratings of children's empathic behavior.

The Contemporized Themes Concerning Blacks Test (C-TCB) is a revision of Robert Williams's (1972) Themes Concerning Blacks Test. It consists of a set of 27 chromatic cards representing contemporary African American daily life and a scoring system grounded in Africentric theory. The content-valid categories consist of eight personality traits rated globally on five-point Likert scales and nine affective states rated globally on four-point Likert scales for all stories. The situational themes include parenting and family experiences, racism, media portrayal of African Americans, substance abuse, and HIV/AIDS. The personality traits include interpersonal, family, and group concerns; goal setting and motivation; self-concept and identity, autonomy, and Black pride. A careful analysis by a panel of raters evaluated the content validity of the card pulls for their intended themes and the personality traits and affective states attributed to each. Frequency norms for 100 participants are given herein.

The Picture Projection Test (PPT) is a revision of a scoring system that has been used with different sets of picture stimuli, culminating in a new set of chromatic paintings that can be computer-administered, potentially over the Internet. The scoring categories were developed to parallel conceptually those of the Rorschach Comprehensive System (Exner, 2003). The findings reported here are especially exciting in that they show surprising consistency between TAT/PPT and Rorschach measures of form, affect modulation, and various forms of experienced distress. That the effect sizes of this consistency are modest reflects the considerably different characteristics of the stimuli and the task involved, namely perception with minimal ideation for the Rorschach compared with construction of a narrative for the TAT. Interestingly, the pattern of correlations with Rorschach scores for PPT scores obtained from stories told to TAT pictures is remarkably similar to that from stories to his new PPT pictures. This suggests that the underlying constructs are remarkably robust when aggregated across 8 to 10 stories, even when different pictures were used by different examiners—as was the case for the archival TAT stories.

What's Not Here

The two most obvious absences from the array of the best available clinical scoring systems for TATs are the two that have received the most research attention since 1990: Cramer's Defense Mechanism Manual (DMM) and Westen et al.'s (1990) SCORS. Cramer's system describes three major defense mechanisms—Denial, Projection, and Identification—as exemplars of successively more mature defensive processes. She published her scoring manual, along with an extensive literature review and theoretical integration, in Cramer (1991). Cramer (1996) updated her accumulated research findings again in the context of narrative interpretation more broadly conceived, including chapters on other scoring systems. Cramer (2006) included important extensions of theory and application as well. Her key empirical publications (Cramer, 2005; Cramer & Blatt, 1990, 1993; Cramer, Blatt, & Ford, 1988) merit careful attention for their integration of theory and data analysis. She has now made the DMM available on the Web at http://www.williams.edu/psychology/faculty/cramer/cramer.html (select "Defense Mechanism Manual").

Unlike most of the systems collected here, Westen et al.'s (1990) SCORS was developed empirically—though guided by psychodynamic theory—by comparing patients in
three diagnostic groups. The SCORS identifies differences in the stories told by patients diagnosed with borderline personality disorder and those told by patients having depression or schizophrenia-spectrum diagnoses. The SCORS too has generated a quantity of significant results. Many of these papers, along with the SCORS manual, are available on Westen's Web site at http://www.psychsystems.net/lab/.

Finally, given the importance of violence, aggression, and hostility among patients and forensic populations, the absence of a system covering this area is notable. However, though often researched, few studies of aggression represented in TAT stories have shown consistent findings from any scoring system beyond an initial study (as discussed by Ritzler in chapter 12 in this volume); thus, none is included in the present volume. Rorschach research on aggression has found that more overtly aggressive responses that include human movement (AG) are less related to socially unacceptable aggression than is more subtle and symbolic representation of aggression via simple aggressive content (e.g., guns, wolf teeth; Meloy & Gacono, 1993). It may be that overtly aggressive content in TAT stories is so transparently undesirable in so many social circumstances that its presence is a better clue to an oppositional response style or a failure of self-censoring control (e.g., Holt's Primary Process in chapter 16 in this volume) than to likelihood of aggressive behavior. In any case, Hafner and Kaplan's (1960) scale has been included in Ritzler's PPT, permitting study of its vicissitudes in the context of other personality and situational variables.

Looking Ahead: An Integrated Approach

TATs begin with perceptually sensible stimuli (pictures of people in situations) and ask the storyteller to make sense of human social interaction. Structured scoring systems define some dimensions that might characterize such interaction. These might reflect what the client's dilemma is, especially if the client sees the task as to make up a story about the presenting problem. But arguably the most useful information for the clinician concerns what it is that makes this a dilemma for the client.

The systems presented here lay out the psychological context of the client's problem: what the client cares about (Need–Threat, Psychocultural), who is involved and whether they are supportive (AAP, Affective Scale, Fine's Scoring Scheme, Pathogenesis), the client's empathy for (Empathy) and depth of insight into those others and their relationship (Interpersonal Decentering) as compared with constriction (TI), the clarity of the client's communication (CD), the accuracy (or at least conventionality) of the client's social perceptions and responses (Dana's Perceptual Range and Perceptual Personalization, PPT, Primary Process), the client's responsiveness to the social environment's reasonable demands (Dana's Perceptual Organization, AAP, PPT) and ability to problem solve about those (Personal Problem-Solving System–Revised, Contemporized Themes Concerning Blacks Test, TEMAS), the client's dependency versus independent functioning (AAP, Fine's Scoring Scheme, Oral Dependency, Symbiosis), the client's emotional experiences and ability to modulate these (Contemporized Themes Concerning Blacks Test, Fine's Scoring Scheme, PPT), and the client's characterological strengths and problems (Ambitious-Narcissism, Contemporized Themes Concerning Blacks Test, Need–Threat, Symbiosis, TEMAS). If the client cannot see the big picture, the well-informed assessor might be able to piece it together using the mosaic thus created.
An array of these more systematic approaches should not replace individualized idiographic interpretation using narrative analysis any more than the medical student’s review of systems will replace the seasoned physician’s immediate recognition of a specific constellation of symptoms, or than either should replace magnetic resonance imaging (MRI). Each view is an essential backdrop to the next more resource-intensive and expensive step, and each provides its own unique information yield without which the diagnosis may be missed and the prescribed regimen inappropriate for the patient’s needs. None can be replaced with either guesswork or the more wasteful practice of providing cheap and possibly misdirected treatment in lieu of a clear and solid picture of the client’s problem. The information yield of structured scoring systems such as those presented here may make it easier for each generation of novices to acquire rapidly and more easily an introductory level of competence in focused topic areas. All clients, regardless of payment method, deserve treatment that is guided by appropriate assessment to be efficient, problem focused, and effective.

References


Teaching How to Learn Reliable Scoring

Sharon Rae Jenkins

Introduction

Accurate, efficient, even enjoyable scoring is key to both scientific clinical use of thematic apperceptive techniques (TATs) and the rapid production of research using structured scoring systems. Successful teaching and learning of low-inference structured scoring systems provides more systematic structure for the learning process than does teaching idiosyncratically organized clinical TAT interpretation; these differ dramatically. Perhaps the best argument for structured systems is that they make learning good idiographic interpretation possible in less than 10 years. There will never be a substitute for 10 years’ experience, but structured systems give the clinician some skills to use productively meanwhile.

Academic faculty approach teaching quite differently from clinicians. The former provide more systematic structure for information and processes, whereas the latter tend to see structure as an imposition on the client’s preexisting structure (which for novice clinicians it sometimes is). For these systems, a less structured teaching strategy is inefficient at best and frustrating to the scorer at worst. Good teaching, like good parenting, arguably requires reliable structure and seriousness of purpose at the outset, with a gradual lessening of external controls as structure and process become internalized.

The manuals in this book provide a structure and materials for self-teaching, and the Appendix to this chapter gives directions to the scorer for using them. The teaching process can best be described as teaching how to learn reliable scoring, more a matter of supervision and guidance in the concept formation process than of pedagogy. This chapter presents some guidelines for supervision of those working with these materials so as to develop scoring expertise rapidly and accurately.

This chapter is designed to help both the individual learning alone and the supervisor of trainees (the latter may be experienced professionals or undergraduate students—or non-students). The former may turn at the end of the next section directly to the Appendix,
and then review the remainder of this chapter as needed for clarification. The Appendix is directed to the trainee to guide self-teaching, and is designed to stand alone or to supplement a supervisor’s guidance. Most of the rest of this chapter is directed to the supervisor.

Orientation to the Scoring Process

Objective observation, meaning that two individuals following the same observation rules under the same conditions will note the same phenomenon, is a central rule of empirical psychological science (e.g., Carnap, 1936). The scoring manuals in this volume provide a variety of such rule systems that have shown adequate objectivity in previous research. However, the application of these rules by research teams in other laboratories might reveal unarticulated theoretical assumptions and unexpected effects of cultural differences. Training scorers to apply manuals accurately and to follow them consistently thus involves not only learning the manual as given, but also attending to any needed local cultural adaptations (Hwang, 2006; Jenkins, 2000). This labor-intensive process has historically been an impediment to research with both TATs and the Rorschach systems.

This chapter is designed to reduce that impediment by describing the group learning process used in my lab with several of the systems in this book (Dobbs, Leeper, & Jenkins, 2004; Jenkins, Dobbs, & Leeper, 2004, 2005; White, McFarland, Londenberg, Chlipala, & Jenkins, 2004). Just as these structured scoring systems make the interpretation of stories more focused and efficient for specific purposes, so also the group learning process herein described can expedite scoring while also improving accuracy and reducing scorer attrition.

Considering scorer training as a developmental process with three phases allows for the differing needs of scorers at each phase. A more structured first phase facilitates rapid learning of the system’s structure and content and establishes group norms for the scoring process. In the second phase, shifting the responsibility for structure to the group builds scorer competence and confidence while also supporting convergence on a shared cognitive strategy. In the third phase, supporting increased autonomy with continued objectivity as the internalization process is completed minimizes scorer drift and makes the scoring process more efficient and faster. Allowing a minimal level of error by omission ensures that scoring will be conservative, reduces false positives, and improves specificity. It also reduces scorer boredom, burnout, and attrition from the project.

Learning a scoring system is best done with an intensive period of work in which the trainee first studies the manual closely, and then reads and scores a number of stories systematically, looking for specific features that are identified by the manual. These features are described by the manual’s scoring categories in terms of the words, phrases, images, or other patterns in the story that are to be given a score. The score might be a simple notation of presence or absence of the feature, or might involve rating the feature—or the whole story—on an ordinal scale representing the degree or intensity of the feature. Quite complex and subtle characteristics of stories can be scored in this way, provided that the description in the manual is sufficiently clear and specific.

For individuals learning on their own, a partner is extremely helpful at some point in the process, first as a sounding board for alternative understandings of a category or scoring dilemma. Second, in order to report interscorer reliability on the new data, a second scorer (and perhaps a third) will be needed. Third, having two learners who can confer to develop any scoring conventions that may be necessary is more likely to result in
reliable scoring, that is, scoring with which scorers trained in other situations would agree. For most of the systems in this volume, scoring partners need not be psychologists; the Appendix was developed for undergraduates.

The following sections present my approach to supervising the learning process, which is heavily indebted to my graduate education in TAT scoring with Jacqueline Fleming and Abigail Stewart, and which comes from the methodological paradigm described in more detail by Smith, Feld, and Franz (1992). For readability, I have omitted numerous well-deserved personal communication citations to Fleming and Stewart, and others to the Smith, Feld, and Franz chapter, which my training predated.

The Appendix summarizes what I have said to my student scorer groups for the past several years, and consists of edited transcripts of my instructions to a recent scoring group that I conducted. Lately, I have been able to hand students the Appendix and a manual with only a polite preamble, and they have been able to proceed rapidly into our scoring groups with little further active direction from me; however, as described in the phase 3 section of this chapter, ongoing monitoring of interscorer reliability is important. The Appendix is designed to be copied and given to scorer trainees, along with a copy of the scoring manual to be studied for the next meeting.

Components of effective supervised learning include (1) proper selection and preparation of trainees; (2) clear initial orientation that begins with more structure, teaches problem solving, increases the scorer's self-awareness and self-observation, and prepares the scorer for eventual authority over the results; (3) a priori stress inoculation that prepares the trainee for the challenging aspects of the learning process; and (4) for group supervision, good use of the scorer group's process dynamics during training. The next section focuses on undergraduate scorers but applies equally—though with fewer caveats—to graduate students.

Selection and Preparation of Undergraduate Scorer Trainees

In my experience, bright, ambitious undergraduates can become reliable scorers of surprisingly complex systems if taught properly. I prefer to supervise scorers in groups of four to eight—a minimum of two two-person teams, perhaps with an alternate in case of attrition. This is large enough to lighten the load of scoring a large data set and to encourage supportive friendships among the kinds of students who are typically attracted to and selected for this work. It is also large enough to dilute somewhat the possible personality conflicts that may arise with the intensity of collaborative work, and to include a variety of individual perspectives that can cancel each other out in the scoring process and thus reduce scorer drift.

In regular group meetings, individuals often benefit from hearing the reasoning behind others’ scoring decisions. A larger sample of individual idiosyncrasies makes it more likely that at least one person will share or understand the idiosyncrasies of the system's author. However, to avoid compounding confusion, it is important that individuals learn not to be wedded to their own unique views if articulating these is distracting the group from the goal of developing a common understanding of the system.

I find that the best undergraduate scorers are bright psychology majors recruited from methods and measurement courses who enjoy learning new things and are verbally astute, motivated to learn about research, bound for graduate school, and needing letters of reference from faculty. The latter extrinsic reasons may be their initial motives for volunteering; the first two are my major selection criteria.
All other things remotely equal, I choose a student who is curious and intrinsically motivated by enjoyment of learning over a student who articulates graduate school ambitions but who cares more about grades or about pleasing me than about exploring ideas for their own sake. My preferred students can use the group process readily to improve their scoring skills and to sustain their reliability. However, they may be more vulnerable than the externally motivated to discouragement and disappointment in themselves during the stressful parts of learning described below. It may be harder for them than for others to manage the initial failure experiences and the later resurgences of same that are involved in the trial-and-error part of the learning process.

In my experience, these undergraduates have some advantages over graduate students as TAT scorers for systems not requiring advanced clinical skills or knowledge of theory; however, the best of them can handle even quite sophisticated systems (see Feffer’s Interpersonal Decentering in chapters 8 and 9 of this volume). Those who are essentially pregraduate students are just as bright and curious, but are less overloaded, anxious, and preoccupied than current graduate students—indeed less than they themselves will be in that role eventually. Furthermore, they are probably intellectually underemployed by their undergraduate coursework. Learning to score offers a cure for their boredom by presenting an engaging and challenging task at which they can self-observe their own growing verbal competence. At this earlier stage of their education, they may be a little more concrete, and thus less prone to make abstract inferences that decrease interscorer agreement, compared with graduate students.

Bright undergraduates, and some graduate students, who otherwise make good scorers have corresponding weaknesses that must be managed in teaching them. First, some of the brightest and most creative have largely evaded the education process by virtue of their preexisting verbal skills, and may even have developed a certain contempt for mass higher education, perhaps an artifact of its increasing dependence on large classes and depersonalized multiple-choice testing. Although capable, some are unaccustomed to focusing their considerable resources in a sustained fashion. Proper scorer supervision can give them the necessary discipline and break them of undesirable habits, preparing them for graduate school in attitude as well as in verbal sophistication.

Second, once they recognize a challenge, in rising to it such students may mobilize resources of creativity that decrease interscorer agreement if not controlled. Third, some—most often women—lack insight into their strengths, and their low self-confidence reduces their conviction in mounting a persuasive argument in favor of their scoring decisions, even when their reasoning is sound. They may also tend to obsess.

Supervising scorers in groups helps to control these specific problems, builds character, teaches discipline, and develops academic skills that will serve them well in graduate school. These latter include skill and persistence in mounting a persuasive evidence-based argument and explicating a self-critical insight. This process also increases and maintains their motivation to learn and execute competent scoring. As discussed in a later section, these potential problems are addressed directly in the Appendix and in the brief lecture portions of my first and second group sessions, and indirectly throughout the process by jokes and teasing as well as by the way I structure the process.

Depending on the needs of a particular group, I might also say to reassure anxious strivers and to defuse potential competitiveness, “You are working with the best and brightest of the psychology majors at this university, who are in turn often the brighter undergraduates. Our situation here may be unfamiliar to you; very bright people may be used to being the smartest in any group, and may assume that this is expected of them. Now you are with others, some of whom may be like you in this way, and you may
understand that about each other. One reason I chose you all was so that you could learn
from each other about your differences and similarities. I will be interested to see what
you learn from this about how to work with others who are at your level in some ways.
Watch, notice, and learn from what they do. This will be good preparation for graduate
school and your future career as a psychologist.” One serendipitous benefit of learning to
score for graduate students who hope for academic careers is that it enhances one’s capac-
ity to focus on picayune detail, arguably a beneficial skill for faculty life.

Use of this group process has important secondary benefits for faculty whose
students will be scoring. My scorers are invited to participate in weekly undergraduate
research team meetings as one benefit of their labors. These meetings serve a variety
of informal training purposes that incorporate both their own proffered agendas (e.g.,
information about graduate study, how to choose a graduate school, how to write an
honors thesis) and mine (e.g., training in literature review writing, statistical analysis,
and preparation of presentations and publications). Experienced scorers are thus well
prepared to use the scored data for honors theses and unusually advanced professional
offerings, to the benefit of all concerned and the field as a whole.

The First Phase of Learning

There are three general phases of learning to score: (1) initial work with the manual and
practice stories, which teaches the basics of the system, and in group supervision estab-
lishes processes for arriving at a common understanding within the group; (2) the move
beyond the expert-scored learning materials, which for groups teaches reliance on the
collective agreement process for individual scoring decisions; and (3) the move to stories
told to different stimuli or by different participant populations, which teaches gener-
alization of this process from familiar material and scoring decisions to new material.
Although I give trainees the entire Appendix at the beginning, I pace them through their
reading at the points where their experience prepares them to understand what is said.
That is, I suggest that they read only the “General Orientation” section at phase 1, then
“Tips for Sessions 3–4” before those sessions, phase 2 when the manual’s practice stories
are exhausted, and phase 3 when the first collaborative story set is completed.

A good scoring group is a better decision maker than any of its members because
their individualities will average out with discussion, which likely makes for better intescorer reliability and more generalizable scores. At the initial phase, scorers learn to rely
on the manual and expert scoring as absolute authority. As they develop a consensual
decision-making process that reinforces close adherence to the scoring manual and expert
scoring, group members can internalize that reasoning and can apply it to new mate-
rial with minimal inference. At the second and third phases, once they have achieved
documented reliability, they will move beyond these materials to generalize these prin-
ciples, recognizing and making decisions from the authority of their own expertise.

Initial Orientation and Instructions

The first phase of learning has three parts: (1) studying the manual; (2) scoring the prac-
tice stories; and (3) beginning to develop self-awareness about personal idiosyncrasies
that might affect their scoring. These phases and skills probably apply to nearly all content analysis systems—broadly conceived—and can be adapted readily for scoring of similar narrative material such as the Rorschach Comprehensive System (Exner, 2003).

My initial and subsequent structured presentations to trainees about the scoring process define the scoring process and its challenges so as to convey my confidence in their competence, my expectations for their performance, and support for the stresses inherent in the learning process. The first step is to orient them as a group to the process of scoring and its challenges. The essence of this lecture, provided in the first “General Orientation” section of the Appendix, is (1) a general orientation to the scoring process; (2) how to study and learn the manual; (3) how to use the practice scoring materials and expert scoring; and (4) a brief overview of expected stressors. The second (and sometimes third) session is usually spent addressing their questions about the process. In the third or fourth session, we cover “Tips for Sessions 3–4” in the Appendix, which includes (5) how to recognize and manage implicit inferences and (6) how to develop personal scoring notes to complement the manual.

Although the overt purpose of the first lecture is to tell them what to do and how to do it, several less obvious functions are served: (1) to teach them to depend on and internalize the scoring manual and expert scoring of the practice stories as the final authority in scoring decisions; (2) to encourage and support self-critique of their own scoring errors as an important part of the self-teaching process; (3) to prepare them to develop as a group a positive, cooperative working climate that encourages active learning and mutual respect; (4) to defuse some of the potential competitiveness that may develop in students who are accustomed to being the best in any class that they take, and for whom anything less is experienced as a demoralizing failure; and (5) to prepare them for constructive handling of the expected disagreements on scoring decisions that will arise in the learning process.

First Session: General Orientation

When I meet with scorers for the first time, I begin by describing the scoring process as a process of concept formation, the ability to recognize familiar patterns in new material. The learning process is the acquisition of that familiarity to the point where recognition requires little thought, like being able to tell an apple from an orange. I give them each a copy of the Appendix, asking them to read the first sections through “Learning the Manual” and to do what it says for our next meeting, that is, to study the manual thoroughly, making notes as described in the Appendix, and then to score the first 5 to 20 stories, as appropriate to the complexity of the system.

The simpler systems have fewer categories and require primarily recognition of concrete semantic category memberships signaled by a few specific words, such as portions of Steven Huprich’s Oral Dependency, or Antoinette Thomas’s Affective Scale. A distinction of language here: The Thomas system is a two-step system requiring a first reading to locate scorable material—role relationships—and a second reading to assign a value—affective valence. The phrase scoring category does not quite apply to such ordinal scales, as they are scored as alternatives rather than as presence–absence as are most of the others. Some of Barry Ritzler’s Picture Projection Test categories are also scored as ordinal scales. More complex systems such as David Ephraim’s Psychocultural System, Reuben Fine’s Scoring Scheme, Michelle Hoy’s Contemporized
Themes Concerning Blacks Test, or Edith Weisskopf’s Transcendence Index (all herein) have many categories requiring subtle distinctions. For these, scorers might benefit from learning a handful of categories well before moving on to others.

For most novice scorers, it is easier to grasp the scoring process if on the first day they start by scoring just one or two categories for the first five stories to be scored, review the expert scores for those categories, then score the next one or two categories for the same stories, and so forth. For the simpler systems, they might be able to manage all the categories for 10–20 stories, given 2 or 3 days to do it along with studying the manual. For the most complex systems, I might ask them to score just four or five categories for the first meeting, perhaps for the first week, for only five stories the first session, with graded increases thereafter paced to their comfort level. Those who have scored other systems or other kinds of content analysis will probably find learning these systems to be less challenging and might start with more categories simultaneously.

When they have a sense of the process and the categories, they will be ready to score all categories in each story. However, the most complex systems can be very demanding on the scorer’s attention and concentration. For the sake of both efficiency and accuracy with these systems, I would consider first determining the number of variables that scorers seem able to hold in attentional focus simultaneously without strain. Then I might have them score the entire story set—even the entire data set—for just those categories. The remaining categories could be scored by the same scorers on a second pass through the data, or by a second team.

As I structure the group’s work process and define the number of stories to be scored for each meeting, I keep in mind the research on human motivation and performance. Need for Achievement ($n_{Ach}$), which likely improves scoring accuracy, is best aroused and maintained by working independently with control over outcomes on tasks that are moderately difficult (challenging), neither too easy (boring) nor too difficult (anxiety arousing or frustrating) and by providing rapid feedback as to task success (McClelland & Koestner, 1992). TAT scoring is thus an excellent activity for arousing and maintaining $n_{Ach}$ As the scoring process becomes familiar and easier and threatens to become boring for the experienced scorer, I find that $n_{Ach}$ can be maintained by engaging scorers in increasingly difficult tasks so as to pace the level of challenge along with their improved abilities. Such tasks related to scoring include helping them to make more precise verbal distinctions by benchmarking, encouraging them toward increased self-insight and self-awareness, and eventually suggesting that they balance more obsessive distinction making with a more rapid scoring pace, all of which are discussed in following sections.

**Second Session: Scorer Group’s Problem Solving**

This section describes instruction that is best delivered in the second session after the first few stories have been scored. Now you have a chance to shape the group discussion process by example. Your routine questions and prompts in response to the scorers’ questions are important examples of the questions that scorers should learn to ask themselves when confronting difficult decisions. Asking this sequence of questions in response to questions about particular scoring decisions models the decision-making process of scoring. It also conveys implicitly your expectation that scorers will have made good-faith efforts to problem solve about their own errors. These questions are as follows:
1. “Let’s look at the manual. How far did it take you with this story, and where did you get stuck? What fits, and what doesn’t? What words or phrases make it fit, and what makes you uncertain?”

2. “What examples in the manual come closest to the part that gives you trouble? How closely did they fit, and how did they not fit?”

3. “Who [in the group] got this one right? How did you see it? What did you notice that told you it should [not] score?”

4. [To the one who got it wrong:] “Can you make sense of it that way [as in 3]? What note can you add to your manual to help you get it right next time?”

Although this process might seem overly concrete to those accustomed to more idiographic and theory-based interpretation, it allows for the precise rendering of even subtleties such as “emotional tone”. The point is to achieve a sufficient tally of specific material, not a thorough, complete, and broad-based interpretation. This approach is also quite effective for working with undergraduates who have been educated recently in states that have emphasized mass standardized educational testing that requires memorized facts as a criterion for high school graduation.

Scorers’ responses to these questions make their reasoning processes available for group discussion, which should consist largely of comparisons among scorers’ rationales rather than a leader’s reiteration of the manual. At this phase, the best rationale is the one most consistent with the manual and in agreement with the expert scoring. As scorers learn to recognize and articulate the features that distinguish this best rationale, they are preparing to move to new material having no expert score.

At the end of the second session, I first ask trainees for feedback about how this process has been working for them. Then I ask how many stories they feel they can score by our next meeting, and we set as a goal the lowest common denominator. Then I say, “What we’ve accomplished today is to look at the reasoning process that goes between the manual and the expert scoring. Now, after you’ve finished scoring the first five of the stories for our next meeting, look at the expert scoring and then correct your own scoring. Then score the next five and correct those, and so on. Try to finish [the goal number]. When you’re done, look over your errors and see if you can see patterns in them. Then try to go through today’s reasoning process yourself and see how far you can get with it. Try to figure out your errors and where you went wrong. What was the expert probably noticing that you missed, or what might have been missing in a response that you scored and the expert didn’t? Then in our meeting we’ll focus on just those errors where you can’t figure out what was wrong.” This is the group’s routine process until all of the practice stories have been scored, checked, and discussed.

**Increasing Self-Observation and Self-Awareness**

The first step of increased insight is for scorers to learn to observe their own internal processes as they score, in order to recount and correct their scoring rationales as they learn the system. Such self-observation can be encouraged by questions such as #4 in the previous section that ask for explicit self-correction, and by an emphasis on elaborating on
their manual notes to document their decisions. Such documentation should be part of the permanent file for the project. In session 3 or 4, most scorers will be ready to read the Appendix section on “Tips for Sessions 3–4,” which is designed to increase self-awareness.

Especially early in the learning process, it is helpful to emphasize improvement and insight into one’s individuality as the criteria for progress in learning rather than being right the first time. Accurate self-critique (e.g., “I know I tend to overscore this category because certain stories to this picture remind me of how my family expressed closeness”) should be rewarded as an indicator of increasingly sophisticated self–other individuation in forming scoring judgments.

The Scorer Group’s Process Dynamics

Once the manual has been learned, application to new data is a social process. When scorers are working as a group, developing a systematic work process will not only improve interscorer reliability, but also keep scorers motivated for productivity. The group’s emotional climate can be created and mobilized to facilitate learning and to minimize interpersonal conflicts. An enjoyable group process acts as a secondary incentive for participation by students who rarely can be paid for their work, reducing scorer attrition.

But more important for effective scoring is the development of a collective decision-making process in which individual scorers’ personal idiosyncrasies can cancel out as they learn about how their own individuality is reflected in their scoring decisions. It is helpful to have scorers who are aware of the local cultures of the storytellers, because the general principles of the manual might translate accurately into somewhat different words, depending on language and local cultural meanings. For example, in scoring an Ach, what counts as a traditional achievement goal may differ cross-culturally, and is to be scored consistent with the storyteller’s culture.

In the process of splitting fine verbal hairs, my teams sometimes discuss the meanings of particular words in scorers’ own families and regional cultures. Some scoring differences arise from personal meanings within scorers’ verbal associative networks that come from their own or family members’ histories of involvement with particular occupational or leisure-time activities, regional or ethnic cultures, religious traditions, atypical family values, and other sources of individuality that make life interesting and scoring agreements complex (see “Making Your Personal Manual” in the Appendix).

A properly developed group process can help students develop insights into the sources of their individuality that will make them better scorers as they learn to distinguish their personal from more general cultural meanings. (This should also make them better candidates for careers in clinical and counseling psychology.) A properly developed process is defined by the group serving this purpose effectively while arousing minimal defensiveness and other undesirable reactions in its members.

At this point, I teach them to use as their frame of reference what the group would likely agree on rather than their own individuality as the basis for difficult decisions. To accomplish this, it is necessary to establish from the beginning a group process that reinforces their recognition of each other’s valid arguments based on the manual and expert scoring, and that reduces their competitive investment in “being right”—and possible shame at “being wrong”—in their initial decision.
The Second Phase: New Data

The goal of phase 2 is for each scorer to establish and maintain interscorer reliability with the other scorers on previously unscored data, scoring independently without discussion. This requires using a minimal-inference reasoning process to generalize from examples in the manual and practice stories to cover instances not included in either. When the group has converged on sound reasoning processes using the manual and has finished scoring the available practice stories, the group is ready to move beyond the manual's expert scoring to previously unscored data. At this point, I move the whole group to a set of stories told to the same picture, preferably a picture that is well represented in the manual and expert scoring.

(In my lab, stories are organized, typed, and scored by picture to facilitate consistent and efficient scoring. This practice minimizes possible bias among stories told by the same storyteller. For consistent reliability, clinicians should score each of the client’s responses independently of all others rather than making some scoring judgments based on scores of other responses. Some clinicians are in the habit of using a client’s scores on other stories to shape questionable or unclear scoring decisions in favor of similarity to scores on other stories, violating the rule that stories should be scored independently.)

For some manuals having few practice stories, this move will be precipitated earlier by the exhaustion of those stories. In such cases, the “Tips for Sessions 3–4” section of the Appendix can be assigned at Session 2 and the “Moving to Phase 2” section along with the new story set. More time and discussion might then be needed at phase 2 to establish the necessary consistent distinctions and to elaborate the manual appropriately for the less familiar stories. Good documentation of these elaborations will be especially important in these cases to guide future researchers, and its availability should be noted in publications, as discussed in the section on “Finishing Up.”

Phase 2 prepares the scorers to make decisions as a group and to see the group as the authority, in the absence of an external expert. When this learning process has consolidated in phase 3, pairs of students may each work on batches of stories told to different pictures. A batch is defined by the group or pair as a reasonable number of stories to finish scoring by the next meeting, the size of which will vary with story length and system difficulty. Because the first batch of stories for any new picture or population will typically generate more disagreements, and thus more discussion, while scorers orient to the new material, I typically start with half the usual number of stories that I would expect to finish were they familiar.

Orientation to Phase 2

The general orientation to phase 2 embeds additional self-esteem stress inoculation in my explanation of the changes in group process. Helpful process changes in the move to new stories are (1) establishing turn taking within the group when reviewing and recording their scoring decisions, and when summarizing the day’s work, to equalize their engagement with the process, decision-making power, and the workload, which often results in (2) increasing their reliance on the group’s common frame of reference for rationales to support their inferences; and (3) teaching them to formalize the rationale for
recurrent picture-specific or population-specific inferences as conventions (e.g., benchmarks) for scoring. You will probably want to decide in advance how you will centralize the recording of individual and consensus scores for the calculation of interscorer reliability and for later data analysis. Each scorer should record the consensus scores for reference in later scoring decisions.

Turn Taking and Reliance on the Group

Whereas in phase 1 discussion was organized by scorers’ questions about understanding the manual and its examples, in phase 2 it proceeds systematically through each of the stories that has been scored. The first step of each meeting is to review the scoring for each story in order, by each scorer in turn presenting her or his scores and the others identifying their agreement or disagreement. This turn-taking process sets important process precedents in that it positions each scorer to take responsibility for having an opinion worth considering seriously, and minimizes a tendency to rely on one as a putative expert, which might result in incomplete processing of story information and scoring criteria. Scorer Erin Cobb (personal communication, July 15, 2005) noted that “consistent acquiescence to [an expert partner’s] rationales leaves one feeling like a puny pushover. It’s pretty easy for me to just nod my head when [given an explanation]…. For first-timers, it should be priority to vocalize why you thought something was the way it was. If not, you end up just nodding your head.”

A variation on the sequence of questions in phase 1 provides the structure for discussing and reconciling discrepancies in scoring as each occurs. Because scoring should be conservative, the burden of proof is on the ones who scored the category present—or highest, for ordinal scales, unless stated otherwise in the manual—so this scorer addresses the following sets of questions:

1. “Why did you score it [that way/so high]? What did you see in the story that makes it scorabale that way? Where in the manual does it describe that as something that would score? How does this meet those criteria? Where does it say that in the story?”

2. “What examples or practice stories are most like that one? How were they scored?”
   (When the scorer has thus made the case, those who didn’t score it [so highly] respond to the third set of questions.)

3. “How did you see it? Did you consider scoring it [that highly]? What told you not to score it?”

Addressing these questions thoroughly at first, and then in more abbreviated form as scorers learn the routine, sets a second important process precedent, establishing the structure of a thorough survey of relevant information when reconciling discrepancies. This serves further to reinforce the importance of considering all opinions, even if one scorer is clearly more accurate than the others and even in those cases for which the non-scorers readily admit having overlooked an obvious piece of evidence. You might then reassure them that anyone can have a moment of oversight and that this is a routine part of learning how to focus their attention in this new way. This gives scorers the opportunity to share their tactics for attention management.
As the group engages in these discussions, they reinforce each other’s learning of the manual, identify and learn to manage their individual differences in language use, and build functional working relationships. In the process, they develop a consensual reasoning process and a common frame of reference for applying the manual to storytellers from the cultures represented in the data. It is especially helpful, of course—and in some cases essential—for the major cultural groups represented in the data to be represented likewise in the scorer group. If the scorers have opportunities to discuss their own cultural backgrounds among themselves, it might be easier for them to notice when a particular scoring discrepancy might be rooted in culture-specific experience or use of language.

**Using Conventions to Increase Interscorer Reliability**

Later in phase 2, when they have become accustomed to the process of reconciling discrepancies, I introduce scorers to the concepts of *interscorer reliability* and *scoring conventions* (see the Appendix section on “Some Common Conventions and Strategies”). There is little on the former in this chapter or the Appendix because the appropriate way of calculating reliability is specific to the mathematical properties and underlying assumptions of each system. Also neglected here are the specific means of computerizing the data, as this process will probably differ among labs; indeed, in my lab, it differs among scoring teams based on individual scorers’ computer skills and interests. I do give scorers an oral explanation of the principles and practice of interscorer reliability determination that is appropriate to their system, and, when feasible, I show them a scatterplot of two scorers’ scores along with the reliability coefficient for the plot.

New picture stimuli may require finer distinctions on one or more dimensions that are different from the stimuli used for the practice materials, and storytellers from differing populations may respond quite differently to some aspects of the pictures. Thus, each new picture or population might require scorers to agree on certain specific scoring conventions for that situation to produce consistent scoring decisions.

Scorers who have attained consistent reliability with the original practice materials have typically internalized the decision-making process well enough to develop picture-specific scoring conventions that produce consistent local applications of the general rules of the system. If the practice stories are too few to establish reliability, the development of conventions may not be so confident, but remains necessary, as is careful documentation of those decisions and benchmarks. These conventions might, for example, identify specific frequently occurring words, phrases, or images that might be loci of disagreements between the scorers in the first set of previously unscored stories and at the beginning of each story set told to an unfamiliar picture.

**Benchmarking**

In such cases, after the scorers have conferred about their disagreements using the three sets of questions listed in the previous section and have identified several similarly problematic responses, the group of responses should be reviewed together along with the relevant scoring rules. To make a general decision about these cases, it is
often helpful to try to put them in some order or in clusters according to one or more relevant similarities and differences. Then it should be possible to decide on one or more cut-points or decision rules that seem most faithful to the manual. Where to draw the line may be somewhat arbitrary. The important thing is to be consistent about where the line is. This process often generates a list of words and phrases that are to be scored and another list that will not score, as described in the Appendix under “Benchmarking.” Working with these lists makes scoring both faster and more accurate.

An example of such a specific convention taken from Rorschach Comprehensive System scoring (Exner, 2003) is the use of talking as a benchmark for distinguishing active from passive human movement. Arbitrarily, talking is scored passive, as is anything more passive than that (e.g., whispering); verbs that are more active than talking (e.g., shouting or even talking loudly) are scored as active. Similarly, TAT scorers might review the specific words and phrases often used for a picture in a data set and agree on such benchmarks.

Calculating and Monitoring Interscorer Reliability

In scoring for research purposes, it is necessary to determine and report interscorer reliability for the stories told to each picture separately. This is because for each system and storyteller population, certain pictures are likely to elicit stories that call for difficult decisions regarding one or more categories because of the imagery in the picture. For example, in some populations, Card 1 (boy with violin) elicits many stories about the boy preferring to play with friends rather than practice the violin. Whether or not those stories are scorable for need for Affiliation (n Aff; Heyns, Veroff, & Atkinson, 1992) hinges in part on whether playing with friends is depicted as motivated by a wish to approach companionship or friendship, or instead to avoid the violin. In such cases, initial interscorer reliability might be low until the problematic cases have been identified and the scoring dilemma articulated and resolved—the purpose of developing conventions.

I typically begin monitoring interscorer reliability with the last 30–50 stories of the story set scored at phase 2, provided that at least 20–30 stories have already been scored and conventions developed. Discussion of the statistical evaluation of interscorer reliability is beyond the scope of this chapter because it is specific to the number of scorers being compared, scalar properties of the scores (nominal/categorical, such as presence–absence content analysis; ordinal, such as ranks and rating scales; interval, such as frequency counts; or ratio, which are generally rare for these scores as a score of 0 usually does not represent a complete absence of the underlying characteristic, only its low representation in the story sample) and the assumptions about the underlying construct (e.g., a simple behavior count; a continuum of personality dispositions divided for convenience into a crude integer scale). The reader should consult appropriate statistical resources, which might include the editorial preferences of journals that are being considered as publication outlets. In our lab, we usually use category agreement (Smith et al., 1992) to monitor agreement on presence of presence–absence categories in each story, and either Spearman’s rho or intraclass correlation for summary scores.

Most of the scoring manuals presented here include appropriate interscorer reliability calculations for their systems. Some of the older ones are exceptions in that they use percentage agreement, and it might not be clear whether this is calculated only on presence or on both presence and absence of a category. The latter approach might be
inflated by agreement on absence, which results in overestimation of scoring accuracy for infrequent categories (Smith et al., 1992).

If all scorers’ interscorer reliability with their consensus scores is at or above the minimum needed for publication, the group is ready to move to phase 3. If not, the group scores another set of stories together in the same way, with interscorer reliability calculated similarly at the end. For a difficult system, challenging stories, or a small sample study, all stories might be scored this way and average reliability with consensus scores reported.

Additional Stress Innoculation

In conjunction with beginning to monitor interscorer reliability—which might induce anxiety or competitive feelings—I continue to talk about the stresses of the learning process. Now I emphasize the challenges presented by the process of articulating and defending one’s position for individuals who care about getting things right and not making mistakes (see the Appendix, “Moving to Phase 2”). I emphasize that mistakes are inherent in the learning process and that the point is to learn from them by returning to the manual and studying it carefully to understand why the wrong decision was made and to correct one’s thinking. I provide stress innoculation by warning them that these necessary mistakes can be discouraging, especially when they are repeated, and that most people will repeat some.

The Third Phase: Working in Pairs

This phase begins when the group members have attained consistent reliability with each other and their consensus scores—or else have been eased out of the group into an alternative aspect of the research project. The goal in phase 3 is to increase efficiency and speed of scoring without loss of accuracy. For the sake of moving through a large data set, the group may be divided into pairs that score different sets of stories and that confer only with each other to finalize scores. The now familiar aspects of group process that will generalize to work in pairs include turn taking in reviewing scores, and self-critique in consultation now with the pair partner to resolve discrepancies.

At this phase, scorers are typically well practiced in the discrepancy resolution process and can carry out this work in the absence of a supervisor, needing assistance only in the determination of a very few decisions. However, supervision is helpful for the first several pair meetings as they become accustomed to the new data and consolidate their work process, and again as they begin each new story set. The new supervision task at this phase is to begin calculating interscorer reliability between the pairs and with their consensus scores on a regular basis as each story set is finished. As there is no majority when pairs disagree, discrepancy resolution might be a little more complex and depends for its success on the skills and self-confidence developed in the group.

The central data quality issue at this phase is to minimize scoring drift, which is the tendency of individual scorers to become less reliable over time due to regression toward their different individual means. It is helpful for this issue to take care in forming pairs. Whenever possible (i.e., not precluded by scheduling conflicts), I pair scorers who are about equally reliable but not necessarily with each other. That is, I pair scorers
who agree less within the larger group, preferably who tend toward opposite biases when they disagree, rather than those who score most similarly. This reduces pair-wise scoring drift. It is helpful if they seem to get along well, but more important for data quality that they be equally assertive. Having compatible schedules and other logistical matters should be lesser considerations if these are not absolutely prohibitive.

**Orientation to Phase 3**

When I orient scorers to phase 3, I typically explain the rationale behind my pairings as given in the previous section, to reinforce their understanding of the importance of thorough discussion and resolution of their differences. I remind them to continue using the group’s decision-making process, not only each other, as their frame of reference for inter-scorer reliability, and to identify and document additional picture- or population-specific conventions that may be needed for stories obtained from new stimuli. When they have completed the first batch or two of stories and have settled their initial conventions, we discuss increasing the pace of scoring and how to reduce infinite regress into interminable analyses. When they will be entering their own data, we discuss any necessary logistics.

**Moving Right Along**

As their expertise and their capacity for discriminating among fine shades of meaning increase, some scorers may tend to bog down in perseveration over details that are smaller than the storyteller could detect. This is more likely to exhaust the scorer than to improve the accuracy of scoring. It may also generate interminable arguments between scorers about trivial distinctions. These likely will not increase the accuracy of scores but will waste time and risk making the scoring process so tedious and aversive as to reduce accuracy. I prefer to rely on the presence of another scorer to catch possible oversights, and I maintain n Ach arousal by suggesting increased speed as a positive value for high performance. I so indicate to the scorers as described in the Appendix. At this stage, as each pair becomes able to sustain acceptable levels of interscorer reliability, I find it helpful to ask them to monitor their scoring pace and try to increase it.

**Picking Up the Pace**

Suggesting attention to speed can be used to draw further attention to scorers’ self-monitoring processes. I invite them to test their growing observational skills by scoring several stories faster than usual, and then going back and reviewing them more slowly to see whether they would change any of their scores. This tactic can be quite effective for systems and categories that rely heavily on visual scanning for literal recognition of keywords and synonyms that might not occur in every story or are even more rare. Accuracy of scoring for these kinds of categories often suffers as much from scorer boredom and inattention as from lack of diligence. However, it should be used with caution for categories that require identification of more complex patterns and those for which
each story receives at least one score. For this reason, there is no section in the Appendix on this topic; the supervisor should appraise the capabilities of the scorers in relation to the demands of the system and should address this topic orally as needed.

Avoiding Hypotheticals

Although it may be helpful at earlier stages to discuss scoring decisions in terms of hypotheticals (i.e., what words, if added, changed, or removed would make a response scorable or not), as the distinctions involved become finer this can become tedious and interminable. At that point hypotheticals should be discouraged in favor of responses that have actually occurred. Although there might be some small gain in refinement at that point, the cost to smooth and efficient group functioning outweighs this benefit. At this point, I might say, “Let’s let that one go. There will always be one or two responses where even excellent scorers disagree because the story is just not clear. If it hasn’t come up before this point, it’s really not likely to come up again often enough to make discussing it worthwhile. Let’s save our energy and move on.”

After a Break

I typically caution scorers not to score when they are too tired or stressed to concentrate effectively, such as during exam week. When a scorer has not scored for a couple of weeks, especially over vacation breaks, a refresher review of the manual is needed, along with rescoring the practice stories and checking against the expert scores. As described in the final Appendix section, it is also helpful to rescore and calculate intrascorer reliability for a previously scored story set before beginning a new one. This will be most effective if the former set is somewhat similar to the new one.

Finishing Up

Monitoring and Reporting Accurate Scoring

For TAT scoring systems, good interrater reliability is the primary indicator of score adequacy for research purposes—necessary but not sufficient. Interscorer reliability should be computed and reported separately for stories to each separate stimulus in the data set and for each individual category, perhaps as a range. That for scorers should be tracked to identify any retraining needed, but need not be reported unless requested by a journal. I typically calculate reliability on the last 30–50 stories of each set, after the initial development of conventions, more for systems having rarer presence–absence categories, because a low base rate will lower interscorer reliability, all else equal. Adequate interrater reliability reported for other studies indicates only that it is possible to score the system reliably. But as the scores are a product of the stimuli, the stories, the manual, the scorers, and the latter’s state of mind as well as expertise, there is no “interrater reliability of the system” beyond its application in a particular study, any more than there is a “the TAT” above and beyond the Murray (1943) picture set.

The interrater reliability calculation indicates only that these scorers, working independently, were able to apply the manual to these stories consistently so as to arrive at the
same conclusions for most of the decisions involved. Inadequate interrater reliability for a study typically indicates slippage in any one or more of these four areas, or a mismatch between them. This could be due to a lack of clarity and specificity in the manual; gaps in the preparation or diligence of the scorers; cultural differences among the manual developer, the storytellers, and the scorers; or some combination of these.

Documentation of Scoring Decisions and Archiving of Revised Manuals

Because of this central role of scoring, and because of the wide variability of stories told by clinical populations—in contrast to the college students so often studied in the human motivation research tradition—questions and dilemmas that arise in the scoring process should be documented and archived as part of a revised manual that is considered part of the data set. This should be made available to other researchers on request for the benefit of consistent application across laboratories, and this availability should be noted in reports. In the event that outside examination of the data becomes desirable, for example by researchers wishing to conduct additional analyses as provided by the American Psychological Association’s Ethical Principles and Guidelines, the scoring documentation should be part of the information given.

As a scoring system matures and the scoring rules become more explicit and less ambiguous, less inference is needed in applying the rules to most responses, and the expert scoring in the revised manual appears more consistent to the novice. These refinements are the product of teaching the system to additional scorers and of elaborating the scoring rule descriptions to anticipate and address frequently asked questions. As such, the revised manual becomes a valuable scientific contribution in its own right, quite apart from any substantive findings from its scores.

The essential material to be archived is what would allow other well-prepared scorers to attain reliability with the original scorers for the stories in question. Typically this would include the notes on scoring conventions and benchmarks made by scorers for each picture that were used as scoring guides for the data set. Scorers’ personal manual notes on their own unique scoring challenges are not so relevant and need not be included. However, if more than one scorer finds a clarifying note useful for the same data set, it may be that this note belongs in the more general consensual scoring notes.

The chief decision for manual archiving is documentation accessibility. One option for such materials is the archiving service of the Library of Congress; another is the Education Resources Information Center (ERIC). For manuals having historical value, the Library of the History of American Psychology at the University of Akron might appreciate a copy. For dissertations and theses, manuals should be added as appendices. Most of these are now available to university libraries through interlibrary loan.

References


### Notes

Although these systems are for the most part too complex and context dependent to be fully scorable by computer, some categories that have a limited number of keywords might have their scoring facilitated by Nudist or a similar thesaurus-type software package. This could reduce errors of omission by allowing scorers to locate the relevant keywords efficiently and then to make a separate context-based judgment for each. This has not been tested in my lab, however.

I thank the scorers whose initial efforts to learn the systems in this book taught me what they needed to know: Meryl Bazaman, Diana Brown, Diane Brüger, Summer Daé Burkman, Kirstine R. Carter, Linda Chlipala, Joni DeGrado, Lauren Dobbs, Michele Elefante, Graciela Fraga, Andrea Haag, Tamara Hale, Melissa Leeper, Lindsay Londenberg, Mike McFarland, Ashley Nicklas, Cara Santa Maria, Luis Perez, Enedelia Saucedo, Rachel White, and Lizzie Woodruff.

### Appendix: To the Scorer

You have been invited to learn this scoring skill because you are bright, competent, and probably underused by your education to this point. You deserve a challenge that will increase your abilities. In this process, you will learn to argue your case, but also to be flexible, and to respect the different strengths that your colleagues may bring to the table. You might be used to being “right”—or feeling like you should be—nearly all of the time;
now you will learn to collaborate with colleagues who have different opinions and to put your arguments together to come up with the best answer.

There are three general phases of learning to score. The first phase involves learning the manual and getting used to the scoring process. For that phase, read the “General Orientation” section, and have a couple of scoring sessions as that section describes. Then when you have had a little scoring experience, read “Tips for Sessions 3–4” before those sessions. Read “Moving to Phase 2” when the manual’s practice stories are exhausted and “Phase 3” when you have attained interscorer reliability (if you are working with a group) or when you feel comfortable with phase 2 (if you are working individually).

General Orientation

Scoring TAT stories involves reading each story and deciding whether or not it meets the criteria described in the scoring manual for each of the scoring categories of the system. That is, does the story include the content described by the category or the rating scale, or is the story told in the way described by the category? During this first phase of the learning process, you will study and memorize your scoring manual. Then each day you will score some stories, compare your scores with the correct expert scoring, and try to figure out where you went wrong when you disagreed with the expert scoring.

Please plan to score for half an hour to an hour nearly every day, especially while you are learning, but not more. Distributed practice is more efficient and less exhausting for this than massed practice. Pick a time each day when your head is usually clear and you are not too stressed or tired to focus, and set that aside as your scoring time. You might have to try some different times until you find one that works in terms of your energy level and mental attitude.

This Appendix and your manual are designed for self-teaching, but it is easier to learn and stay reliable if you work with a group. It also helps to start out with a more intense period of immersion while you are learning the manual and getting used to the process. In your group meetings, you will review your scoring together each day or two for the next week or two, looking at how close your scoring is to the expert’s, and whether there are any patterns in where you are going wrong. You will also discuss any aspects of this process that might be challenging. This may include parts of the manual or expert scores that you find hard to understand, or particular categories that may be hard for you to get right. This discussion is likely to be helpful in cases where several of you are making similar errors; one of you may have an insight about how you went wrong in your scoring decisions and can explain it to the others.

The process of learning to score TATs is partly cognitive and partly emotional. The cognitive part involves your learning to discipline your attention so that you do not miss things that you should score, and to split verbal hairs accurately. That is, you will develop a sophisticated eye for fine distinctions of meaning between words that score and words that do not, which you may now see as almost synonyms. Gradually, as you learn the system, what begins as a laborious decision-making process will become a recognition process. You will become so familiar with the scoring criteria that you might say, like one scorer I taught, “I know I’m getting it because most of the time I can just see it; I don’t have to think about it. It just gets my attention.” At this point, you may find yourself automatically scoring incidental print such as billboards, cereal boxes, and your textbooks.
It is helpful to know about the emotional part of scoring to prepare you for the stresses of the process. You are bright and competent, or you would not have been invited to score. But the process can be stressful too, and some stress inoculation is useful. The process of reconciling disagreements can be hard on the self-esteem, but we will come back to that later.

The cognitive part of learning to score involves reading and understanding the scoring manual that you are working with. When you go home tonight, study your manual using the strategy described in the next section. Then score a few practice stories; you will discuss those in your next meeting.

Now, turn to the section of your manual that describes the scoring categories of the system. This section tells you what it is that you are looking for in the story and what to do when you think you see it. There are also examples illustrating the rules and the decision process. Applying these rules is a process of making generalizations to new stories from the abstract rules and the particular examples given in the manual, and deciding whether each new story fits that rule or resembles an example closely enough to be scored. As you are learning and applying each rule, you will be thinking of all the different ways that scorable material might appear.

These decision-making rules can be thought of as statutory law, like the laws passed by Congress and state legislatures. When you read stories, you are sifting through the text looking for evidence of something to score—that is, words and phrases that fit the rules given. This is the same process that lawyers and judges use when applying statutory laws to particular cases. The historical record of the decisions they make is called case law; each new case is argued based on both statutory law and the accumulated precedents of case law. Studying the statutory law alone is not enough, because storytellers have many different styles and thus different ways of expressing the characteristics that the scoring system is designed to identify.

Each picture has common things people say that will often be scorable. As you learn the scoring system, you learn to recognize these things, and they become familiar to you. Then you consult the manual for your specific rules and run the word or phrase that looked familiar through the rule system to see if it meets the criteria for scoring. As you get more experienced at that process and with the stories people often tell about a picture, you will find it easier to recognize what you are looking for. The patterns will become repetitive, and scoring will often feel like a reflex. Although it will be slow to start with, you will find that you score more quickly as you go along.

And at times you will be wrong a lot, which brings up the emotional side. Everybody who has done this has gone through periods of being wrong a lot. It builds character. Not only that, but if you have been getting through school writing papers the night before they are due and still getting A's because you are so smart, you will benefit from some humility.

When you finish scoring the practice stories in the manual, you will be starting on a new set of stories from a different sample of storytellers than the ones who told the stories in the manual. There will not be expert scoring for those, so your group will have to become the experts. Then you will determine reliability by comparing your scores within the group. In doing that, you will be explaining your scoring decisions to each other, referring to the statutory law in the manual and also to the case law of your notes about the scoring examples and previously scored stories that resemble the one you are discussing. So when you find that a particular decision is difficult for you, make a note about the evidence for and against scoring it each way. That will help you learn the manual and will also be important as a record of your group’s decision processes for future scorers to use.
Learning the Manual

For your next meeting, take home the scoring manual and study it, and then score the first batch of stories. This kind of studying is a more complex and thorough process than most people are used to. Memorizing the manual is only the beginning. You must embed the category descriptions into your own cognitive structure so that you can articulate them for yourself in your own terms and can execute consistently and accurately the decision process they describe.

Begin by outlining the manual—memorizing as you go—so that you can have the list of categories and brief scoring criteria in front of you as you score. Leave extra spaces so you can make notes on each section and bring a copy to each scoring meeting. As you discuss each story, you can add notes on how it is scored, if you missed something or had questions.

Scoring the Practice Stories

When you have finished outlining the manual, locate the practice story section and bookmark it, and then locate the expert scoring for the first story, bookmark it, and hide the expert scores for the second story. Then read through the first story with the first scoring categories or decision tasks at hand, and look for evidence that the story might merit a score. At first you may need to go through the story sentence by sentence while you try to keep the scoring category details in mind. If you see something that looks plausibly scorable; then go through the scoring criteria systematically to make sure it fits in all ways.

When you have decided whether you see it or not and how to score it, compare your score with the expert scoring. Then turn back to the second story and do the same thing, hiding the expert scores for the third story, and so on for the first five stories. If your manual has so many categories or complex decisions that it is hard to keep them all in your head together, you might want to do a smaller number for the first five stories until you have some idea how the process works and you feel clear about them. Then go back and do the same for the rest. If you are feeling confident after five stories, you might wish to continue scoring with a few more in the same reading.

Make notes on your copy of the stories as to what fits and why—for example, by underlining the word or phrase that seems to capture what you’re looking for—or make a note in the margin. You will be finished scoring the story when you have decided for each category whether you see or do not see enough, or any, evidence that it meets the criteria for scoring in a particular category. Then compare your scoring with the expert scoring, and see whether you agreed with the expert or not. In scoring your stories, always work by yourself, and do not discuss with others until you have made your scoring decisions.

At each meeting, the group will go over the stories that you scored and will discuss how to understand what the manual says to do so that you make the same decisions as the expert. There may indeed be parts of the manual that are unclear or not very specific about how to score a particular story, but we figure the expert is always right, so the question is how to understand and apply the rules consistently to each story. There will sometimes be borderline cases in which particular word choices make a decision difficult. However, the expert scoring is always right, and your job is to reason from the scoring manual to the expert’s decision, using the latter to understand how the manual should be applied: what fits the rules, which kinds of expressions are strong enough and which are too weak, and where to draw the line.
Thus, learning a system is partly a memorization task (i.e., learning the category definitions and their boundaries) and partly a generalization task (i.e., learning how to apply the decision-making process consistently to different kinds of stories). You will be scoring each story by applying the statutory law of the manual, plus the case law reflected in the examples and in your previously scored practice stories (when you have some), to each scoring decision. You should add to your notes in your manual the reasoning behind each difficult decision so that you have a record of your own case law and your own precedents.

Attitude

Even for established professionals, self-teaching presents its own unique challenges involving the right mental attitude. Those who wish to learn to use these systems will benefit from approaching the manual with a receptive, noncritical attitude during the learning process. You are learning how someone else defined these concepts, which might differ slightly from the meaning that they have for you. Accepting the manual’s rules for scoring as the Voice of Authority is essential until internalization has occurred. Some basic principles should be kept in mind or on a large wall poster above the scorer’s work area:

1. The manual is always right but sometimes needs clarification, so you must provide it for yourself by diligent study of the literal words and phrases given, and of the expert scoring, so as to make appropriate generalizations to new material.

2. The unfortunate reality: Sometimes the expert is inconsistent or obscure. This makes (1) more difficult but no less true.

3. The moral of the story: Diligent persistence pays off.

Tips for Sessions 3–4

As you begin to understand the scoring process, you will likely find that it is useful to be able to observe your own thoughts, feelings, and decision-making processes and to reflect on them. This is a particularly helpful skill to have if you want to go into clinical or counseling fields. You might begin by watching your own developing attentional process. For some systems and categories, noticing key words is important, but do you more often miss those that appear at the ends of lines? How can you solve that visual scanning problem? For some systems and scorers, visualization of the story might be good, but for others it might be a trap.

Do Not Score Your Head Movie

These are low inference scoring systems, which means that you have to score concretely rather than abstractly. That is, you should think in terms of being able to underline specific words and phrases that meet the scoring criteria and serve as evidence that the story should be scored. You must be sure that the scoring criteria are met by what the
storyteller actually says in the story, not by what you imagine while you are reading. We say, "Score the words on the paper; do not score the movie you get in your head when you read the story." If you are inclined to visualize as you are reading, you may tend to score what you see in your mind’s eye so that you are no longer scoring the storyteller’s words. Accuracy of scoring depends on scoring only what the person actually says as opposed to what you think they mean.

One of the challenges of learning to score is learning not to use your imagination in this otherwise helpful way that has helped you to get this far in your life. Folks who are bright and verbal can often find creative reasons to score things that are not so obvious to others. One challenge, then, becomes knowing when to quit. The two golden rules of scoring are, “When unsure, score like everyone else (especially the expert) would,” and “When in doubt, don’t (score).” Consider that a story is written as a brief snapshot, or maybe a video, out of the characters’ lives. Score only what occurs within the framework that the author gives.

Making Your Personal Manual

The fine line to walk is between making too many inferences (i.e., making leaps from the story on paper to your visualization of the action and your assumptions about what you would do if you were the character) and taking the manual too literally, making no inferences at all. The problem with the first is that we are all different people, and we may visualize the story action and infer the character’s motivations differently because of the personal meaning of the story for each of us; this is bad for interscorer agreement. But if we make no inferences at all, it becomes impossible to make generalizations to new stories told about new pictures with different wording.

One purpose of the scoring group is for scorers to develop skills as a group in learning to walk that fine line and in helping each other control your creativity and individuality so that you all apply the rules in the same way. Because we mostly grow up in different families, each of us has particular ways of understanding certain words, phrases, or social conventions that may come to light in the process of discussing the assumptions behind our scoring decisions. For the purpose of improving your scoring reliability, each of you may have specific points on which you need to make personalized notes in your manuals about the meaning of certain things for you as they are to be applied in scoring. We can think of this as a personal scoring manual that elaborates on the author’s manual. These personal manual notes may be a little different for everyone, because they mediate between our own individual cognitive structures, the author’s meaning, and the general wisdom of the group, for the sake of improving interscorer reliability.

Moving to Phase 2

Now you are ready to become experts yourselves, as you move to new stories that do not have expert scoring. At first you will all be scoring the same stories; later you will be working in pairs. For your next meeting, score the first batch. When you meet, you will compare scores to identify agreements, and then you will discuss and reconcile discrepancies in scoring.
Reconciling Discrepant Scores

Because you will not have expert scoring to compare with anymore, now you will be developing the expert scoring by consensus among the group. When scorers differ, you will explain why you scored the story as you did, and the other scorers will explain why they scored as they did. Your leader will referee the discussion and will help you to put together the two sets of reasoning about why the response should be scored in a particular way to come to the best decision. Keep track of your original score—do not “correct” it—and record the consensus score with a note.

Sometimes the process of reconciling discrepancies will be a simple matter of having overlooked something that is obvious once someone else has pointed it out. This is how you learn to be more careful. Other times you will have considered a score but decided against it for a reason that you can explain. Then it is up to the one who scored it to persuade you that the scoring criteria are met. At still other times, you will not have considered that score and it is not at all obvious, in which case the other scorer will have to make a very persuasive argument before you agree. If you are the one who scored it, it is up to you to make a persuasive case, putting together the manual rules, the expert scoring, and your notes on previous decisions. If you did not score it, it is your job to resist until the other person’s argument and evidence from the story persuade you. Remember, this is not an exercise in either creativity or compliance; follow the rules, do not work too hard, and try to score only what would be obvious to most people in the group.

If you are a person who is not used to making mistakes, this part might be painful. As the scoring process involves working in groups and then in pairs, you will be comparing your scores with those of others. Now that you are beyond the stories that the expert has scored and you are working together to become the experts, a certain amount of contentiousness might arise, along with a certain investment in being right. Be warned in advance that probably you will not win more than half of those battles, but doing well at scoring is not about that. Doing well at scoring is being able to explain why you made the judgment that you made, which should improve your ability to stand up for your ideas.

Be aware that most people do not improve continuously at a steady rate. Often the first wave of confidence after a much improved batch of scoring is followed by a discouragingly worse one. Error patterns also change, with a period of overscoring followed by one of underscoring, or vice versa. Try to be patient.

There will be three kinds of outcomes of these discussions. First, when you have been able to make a clear decision, the rationale for decision making will be clearer to everyone, and you will add that rationale to your scoring manual. That rationale might be specific to that picture but also might generalize to other pictures. The addition to your manual will help you make that decision the same way again when it comes up next.

The second helpful outcome will be that both scorers will be more accurate with that decision next time, having an explicit reason for scoring, which may involve refining your scoring rationale a little bit. That is a normal part of the learning process. The third product will be a scored story, for which the scores can be trusted because both scorers agreed that the scoring criteria were present and felt confident enough in their score that they could persuade someone else. Be sure to record your consensus score with an explanation, and keep your original score unchanged, so that you know where you started from.

So the three products are an elaborated scoring manual for yourself that helps you to manage your individuality and creativity so as to score reliably; useful learning for you; and scored data that can be used for research. Your learning includes how to make
fine distinctions about words and about a particular TAT scoring system. Through the process of learning this, you will likely also gain more ego strength, more humility, and some group collaboration skills that will help you in graduate school.

Some Common Conventions and Strategies

In scoring, a convention is a guideline adopted by a group of people to establish agreement in matters for which several options are about equally good and there is a need to settle, somewhat arbitrarily, on one of them. Conventions serve to establish agreement about scoring questions involving fine distinctions that come up relatively often, usually due to ambiguity or not enough information in the story. Like social conventions, scoring conventions are rules for handling minor decisions that might reflect chance fluctuations in a person’s word choices rather than important aspects of personality, but for which a line must be drawn to define an essentially arbitrary distinction between very similar words and phrases as to which should be scored and which not. For example, because each picture prompts stories with typical themes that appear often, sometimes fine distinctions must be made about these frequent themes to distinguish between stories that are merely recounting the picture and those that go beyond the picture far enough to be interpreted as representing the person’s internal world rather than only the picture.

General conventions that improve interscorer reliability include scoring conservatively—that is, scoring only confident decisions that have clear evidence, not merely possible or plausible inferences. Conventions that facilitate learning include learning to be concrete and specific, which is difficult for some intelligent and creative people. A second kind of convention is called a benchmark.

Benchmarking

Sometimes a scoring dilemma recurs for a particular story set because people tell similar stories about the same picture. Then it becomes helpful to review several stories together and to make a clear decision that will cover all of them and that can be applied consistently to the rest of the story set. Reading the manual and its examples, and the expert scoring of the practice stories, we can look for pivotal words indicating the strength of the statement being scored that fall on either side of a cut-point. These words can be called benchmarks.

A venerable line of research on the semantic mapping of language factor analyzed the structure of English vocabulary (Osgood, Suci, & Tannenbaum, 1957) and found three main dimensions that underlie and can be used to rate all words: evaluation (good–bad), potency (strong–weak), and activity (active–passive). If we had to, we could look up the ratings of the words in question on these dimensions. However, that is a lot of work, and most good scorers can probably make reasonably accurate judgments that also take into account the context of the word’s use, which the factor scores do not include. And usually all we need is one clearly defined point—the cut-point or benchmark—that separates what scores in a category from what does not.

For example, we can think of a set of emotion words as ranked on a continuum with terror, horror, fear, dread, apprehension, concern, and the like at one end and ecstasy, joy, happiness, pleasure, comfort, safety, and the like at the other end, with indifference in the middle. Scoring on that dimension then means making a decision about what is strong enough to score and what is not. If the manual says to score positive emotion words
and negative emotion words, then everything between ecstasy and comfort would probably score as positive, and everything between terror and apprehension would likely score as negative. But if there is an ordinal scale that ranks words as strongly positive, somewhat positive, somewhat negative, and strongly negative, then it is necessary to decide where to divide the numbers on the rating scale. Fear seems pretty strong, but how about dread? Is happiness strong, or only somewhat positive? Different scoring systems might define these benchmarks differently. In your manual notes, you will be gathering lists of what scores and what does not; then each new word should be scored in terms of whether it is more like the first list or the second list in terms of the degree of the dimension to be scored.

**Precedence**

In some systems, it is important to decide which scoring category of several mutually exclusive alternatives should take precedence. Three useful ways of doing this come from survey research: priority, intensity, and frequency. That is, respondents can indicate relative importance by mentioning something first, by mentioning it more often (e.g., saying two or more things that score in the same category), or by using stronger language (or saying, “Really, the most important is …”). In the context of how a story is typically organized, mentioning something first may be dictated by its chronology in the action rather than its importance, and mentioning something more often is likely to be influenced by verbosity, creating a bias related to story length. Thus, we often fall back on the strength or intensity of the words chosen, using benchmarking to indicate priority. However, in some systems, considering multiple mentions is a permissible and useful tactic.

**Strategic Guessing**

Scorer Erin Cobb (personal communication, July 15, 2005) found it helpful to begin with initial guesses about alternative categories that might apply in a particular situation. For each, she reviewed the criteria for scoring each category, tallying one point for each criterion that matched something in the story. This would include each similar example in either the manual or the practice stories. Then she added up the points and scored the one with the most points. This approach is not appropriate for all categories, as they may differ in the number of criteria involved; however, it should work for examples and practice stories. She noted that this approach reinforced her learning of the category criteria and her ability to make distinctions among them.

**Alternative Understandings**

One useful exercise when you are pondering a score about an ambiguous phrase and trying to decide how much of an inference is legitimate to make is to think of alternative understandings of what the storyteller might mean as a way of benchmarking your inferences. Look at the unambiguous words and phrases that come closest to the one you are trying to score that you would use as examples in arguing on one side or other. Then ask yourself what other explanations for these words and phrases might there be in the context of this story that would not score. For example, the phrase defending one’s honor (not otherwise clarified) might be a matter of professional ethics of a financial nature, personal morality in the eyes of God, or sexual chastity so as not to dishonor one’s family, depending on the context in the story and the cultural frame of reference of the
storyteller. Each of these might score differently in different systems, so it would be important to decide which meaning was closest to the storyteller's frame of reference.

Moving to Phase 3

Orientation to Phase 3

Moving on to your next set of stories, you will now be working in pairs. For your next meeting, each pair will score the first batch of the new set of stories, and then will meet and compare your scores as you did with the group. Take turns giving your scores the same way. Record each of your scores in the master file, with your agreements in the consensus score section. See how many of your disagreements you can reconcile, and record those under consensus too. Then bring your unreconciled disagreements, and any other questions that come up, to the next group meeting.

Typically, when you shift to a new picture, things change. Themes that came up in the familiar picture do not come up anymore, and new ones emerge that are phrased differently. Similar thoughts and feelings may be expressed differently, so new conventions might be needed to cover the different words and phrases that people often use. You might need a new set of benchmarks so that you can decide how strong or weak something is. In addition to improving accuracy and reliability, conventions can be used to simplify tedious arguments about scoring decisions that matter little in the final score.

Moving Right Along

As you become familiar with the system, the manual, and the scoring process, you will find that you can score more rapidly without increasing your error rate. Clinical psychologists learn a skill for attention management called evenly hovering attention. As applied to scoring, this means not necessarily focusing on each word separately but moving rapidly over a stream of them in a smooth continuous process of observation. It is still necessary to read the story carefully, but—having turned off your head movie!—scorable material will begin to jump out into the net of your waiting attention. The absence of scorable material will be easy to detect because nothing jumps out, reducing the rereading of stories without it.

Shifting to each new story set means scoring more slowly for a while as new benchmarks and conventions are established, but then you will gradually pick up speed again. Score as rapidly as you comfortably can while maintaining agreement with your teammate. Continue to score independently; that is, score all the assigned stories before discussing them with anyone else.

By now you are now all experienced scorers, with highly developed verbal skills that allow you to split the hairs of this scoring system very finely. As you go through the benchmarking process and the convention development process, part of the reason to have these is to impose cloture (cut off debate) on potentially interminable analyses of small differences of word choice that are likely due to psychological error variance or irrelevant aspects of the storyteller's personality. Conventions and benchmarks are one