A HANDBOOK OF

Clinical Scoring Systems for Thematic Apperceptive Techniques

EDITED BY
SHARON RAE JENKINS
Future Directions for Thematic Apperceptive Techniques

Sharon Rae Jenkins

Introduction

The bounty of divergent perspectives represented by the scoring systems in this volume invites a wealth of future research. It remains in this chapter to offer some priorities for the likely most important and fruitful pursuits. The authors in this volume have each included in their summary chapters a section on future research needed. However, there are some larger questions that, if explored, might make notable contributions to the literature by improving the clinical utility and scientific precision of most of these systems.

For parsimony’s sake, I propose a set of assumptions that I hope will facilitate discussion, if not consensus, on these questions across members of different theoretical and methodological communities:

(1) Picture stimuli sample the storyteller’s imagined interpersonal scenarios in hypothetical situations.

(2) These scenarios reflect the storyteller’s experiences, direct or vicarious, of role relationships, plausible activities and motivations of characters in those roles, and prevailing affects in situations like the one pictured.

(3) The experiences related in the scenarios generalize to similar internal experiences—*not necessarily similar actions*—in similar situations in the storyteller’s life.

(4) These similarities can be at high and unobvious levels of abstraction, as well as at lower levels of more concrete and literal representations.
(5) The relations between inner experience and overt action—and thus the predictive power of single scores on any measure—become more complex and less direct as a healthy person matures.

(6) Therefore, one general high priority for thematic apperceptive technique (TAT) research resides in studying the gap between inner experience and overt action, for example, by including demographic variables, other personality characteristics, and life circumstance influences as controls, moderators, and mediating variables.

(7) Therefore, one efficient strategy for such research involves multivariate studies using combinations of TAT scoring systems.

Careful conceptualization of how different aspects of inner and outer social experiences combine in the storyteller’s situation might forestall a host of negative findings using single systems and overly simplistic predictive models. We now have the technology to handle complex decision trees and multivariate hypotheses, and we should be applying it to these larger questions.

The larger questions begin with how best to strengthen the empirical foundations of the clinical systems presently available. Can any general statements be made about the best directions and caveats for establishing their validity, generalizability, and reliability? Then, what other scoring approaches should be tried for clinical use? What do we most need to know about these systems’ utility for the tasks of clinical practice assessment: establishing the diagnosis, estimating prognosis under varying circumstances, and treatment planning? How about studying the treatment process? What general methodological refinements should be considered? In particular, what about the need for more sets of pictures for both general and specific purposes? The effects of different picture characteristics on criterion-related validity merit attention and likely vary across systems. Finally, studies that combine two or more systems as complementary tools for the same purpose would give a more integrated picture of how story narratives might reflect storytellers’ typical social functioning.

□ General Agenda for Present Systems

The systems in this book were chosen for their promise; only a few—Communication Deviance, Objective Scoring, Transcendence Index (TI)—have yet accrued a substantial validational literature. For most, it will be easy to contribute to the basic research foundation because so much remains to be done. In some cases, conceptual replications of the initial validational studies would be useful, though for some systems this might be challenging for legitimate reasons of cultural change, as discussed in a later section. For nearly all, the content validity of their scoring categories is clear and consistent with theory, but additional aspects of validity need supporting evidence. The range of validity generalization across population—by age, gender, and educational level as well as by ethnicity and culture—is a high priority for all. Any questions of interscorer reliability in relation to the theoretical properties of the construct and the developmental stage of the scoring manual should be addressed. Examination of test–retest stability and cross-situational consistency versus situational specificity can provide clarity about the match between the measure’s properties and the theoretical properties of the construct.
Validity Studies

The specific validity studies needed vary across systems and are best addressed in each summary chapter. Some general strategies and caveats applying to most systems are given here. These involve identifying the minimal valid story, choosing appropriate criterion variables, improving our understanding of the structure and properties of the construct as measured in the population of interest, and addressing challenges in revalidating or rederiving measures under conditions of social change.

The question of the minimal valid story concerns the validity of the data sample (i.e., stories) rather than of the scoring system used, though systems likely differ in this area. How much information is enough to make a confident statement about the storyteller’s response to the pictured situation? This could be framed as the number of items (i.e., stories) or as the average size of each one (i.e., story length). Empirical studies of the Rorschach Comprehensive System (Exner, 2003) established a minimal protocol length of 14 responses, or items, for validity; few of the present TAT systems have comparable empirical evidence.

Story length has received much less attention in the clinical literature than in the academic studies, especially at the level of the individual story. The major concern has been evaluating the degree to which simply telling longer stories gets higher scores, rather than establishing a cutoff point for protocol validity. Dana (1959) examined and found correlations for his variables with number of words in the whole protocol, and research in the human motivation tradition routinely appraises and corrects scores for correlations with protocol length (Smith, Feld, & Franz, 1992).

In addition to that evaluation of possible measurement method error variance from response productivity, validity studies should routinely report exploratory analyses of the effects on validity coefficients of retaining or eliminating both short stories (say, arbitrarily, fewer than 25–50 words, depending on the complexity of the scoring system) and short protocols (arbitrarily, an average of 25–100 words per story). In chapters 12 and 13 in this volume on Barry Ritzler’s Picture Projection Test (PPT), he reported stronger hypothesized validity coefficients relating PPT and Rorschach variables in a subsample consisting of those participants having “open” protocols (Rorschach Lambda less than 1.00; PPT stories averaging more than 50 words in length) than for the sample as a whole. Many of the less central, unhypothesized, but theoretically reasonable associations also increased, especially those involving PPT Affect Suppression and the Rorschach Comprehensive System Depression Index and Texture scores. Graphic data analysis techniques such as scatterplots of TAT scores against story length, and comparing plots of criterion variables against scores corrected and uncorrected for story length, might reveal important nonlinear effects. Over time, cumulation of findings might indicate appropriate minimal story cutoffs for each system.

The question of story richness versus impoverishment, or the degree to which a story goes beyond merely describing the picture, as a measure of story validity has been little noted since Weisskopf’s (1950) publication of her TI (see chapters 4 and 5 in this volume), but that system should be explored as a possible counterpart to Lambda in the Rorschach Comprehensive System (Exner, 2003). Like Lambda, TI scores might reflect character style, a defensive response set, or neurological deficit, any of which might render questionable the usefulness of stories for some scoring systems.

Careful criterion variable selection links TAT scores to clinical practice. The best appraisal of a TAT system’s validity for clinical practice, beyond content validity, involves
choosing criterion variables that theoretically should be related to its scores within an appropriate population for a particular purpose under the right circumstances, and looking at the association between the system’s scores and the criterion variables. The most effective tests of such associations will involve, first, variables that are similar in frequency distribution and other properties described in any introductory statistics or measurement textbook. That is a given for any data analysis.

But second, TAT scores and the criterion variables should measure constructs that have similar structure and properties. TAT scores provide structured ways of looking at less structured data, and they correlate most highly with criterion variables that similarly provide structure for less structured behavior, such as experience sampling measures (McAdams & Constantian, 1983) and long-term career outcomes (McClelland, 1965, 1980, 1985a, 1985b). Regarding the properties of constructs, TAT scores for constructs that are theoretically trait-like should be compared with criterion variables that are likewise cross-situationally consistent (McClelland, 1981). Situation-specific constructs chosen to shed light on localized client problems should be measured using at least several pictures that evoke, or “pull for”, that specific situation (Blankenship et al., 2006; Tuerlinckx, DeBoeck, & Lens, 2002). Scoring systems for constructs that represent transient states or situation-specific responses might be studied with repeated-measures designs that compare scores from the same individuals studied under different administration conditions and instructional sets. Rather than overgeneralizing the results of such studies beyond particular systems to “the TAT,” they should be understood as unique to each individual scoring system until proven otherwise.

Now to the caveats for validity studies. Correlating TAT scores with similarly named constructs measured by structured self-report scales is rarely successful or useful as a validational strategy (Bornstein, 1998; McClelland, 1980, 1985a, 1985b; McClelland, Koestner, & Weinberger, 1989; Spangler, 1992; Winter, John, Stewart, Klohnen, & Duncan, 1998). As noted in chapter 1, construct names can be misleading, obscuring differences in theoretical origins and basic assumptions. Contrasting sources of systematic measurement error can also produce modest statistical associations and contrasting patterns of correlates (McClelland, Koestner, & Weinberger, 1989).

The effects of historical cultural change on validity studies should not be overlooked. Those who might wish to replicate a previous validational study to rederive or revalidate a system should consider carefully the ways in which the original studies for the older measures might not be reproducible with the necessary precision. What cultural changes might have taken place that might affect the findings? If the roles and opportunities available to a class of individuals have changed dramatically, the pattern of correlates for personality variables might change as well (Jenkins, 1987, 1996). If diagnostic categories are involved, some of these have changed considerably since the advent of the first psychiatric Diagnostic and Statistical Manuals (DSMs). For example, Dana’s Objective System (1959; see also chapters 6 and 7 in this volume) differentiated psychotic, neurotic, and normal groups. More precise distinctions among criterion groups would be both advisable and available with present-day DSM diagnoses, but the pattern of correlates for TAT scores might change.

For another case, Masling, Rabie, and Blondheim’s (1967) original derivation study for the Oral Dependency measures used obese and nonobese patients in Israel, as described in Steven Huprich’s summary chapter in this volume (chapter 20). To what extent might the causes of obesity in present-day American culture differ from those of the original project? Pervasive public health messages currently suggest that obesity is virtually epidemic in the United States; these implicate modern diets (e.g., starchy school lunches;
overconsumption of candy, sugar-filled drinks, and saturated fatty snacks; oversized restaurant meals) and chronically sedentary lifestyles. Few if any of these were operative in Israeli culture in 1967. Currently these cultural habits likely play a considerable role in obesity. Although oral dependency cannot be ruled out as an influence on who takes up these habits, there are also strong situational effects due to institutional factors that make fattening foods more available than healthy foods and sedentary office jobs more common than formerly. These other effects should then be measured and statistically controlled if the Oral Dependency scoring system is to be revalidated. It might be simpler just to choose criterion variables that might be more suitable to present-day American culture, as Huprich did.

**Generalizability Studies**

Each of the systems here was developed with a particular examinee population in mind, and the available research may or may not have been conducted exclusively on that population. Wide generalizability is not necessary for a system to be useful; narrowly targeted systems may fill unique niches in our assessment toolbox where nothing else will do. However, within or outside that niche, these questions should be addressed and answered by systematic studies: Are findings equally useful for examinees of all ages? All educational levels? Across languages? Across cultural backgrounds, including institutional and regional culture as well as ethnic, religious, and national cultures (Helms, 1992)? What does the theoretical foundation of the system lead us to expect? The face validity of the scoring categories? The literature?

Adequate information about population generalizability and its limitations may require innovation in our journals to encourage the publication of brief reports of null results for all measures. Without well-documented records in accountable reporting venues, measures are subject to gossip-like whispering campaigns that globalize the findings into a generalized failure of “the TAT” (e.g., “We used it and didn’t get anything”). Instead, a scientific attitude encourages specification of the conditions under which null findings were obtained and closer scrutiny of the methods used. Effective use of TAT measures requires thoughtful, theory-guided picture selection, carefully standardized administration, and reliable scoring, and failure of any one of these might yield invalid data (Lundy, 1985, 1988). Such reports are necessary to isolate the possible sources of null results in terms of the population assessed, the data gathered, the scoring system, statistical artifacts such as restriction of range or marginal interscorer reliability, and other data features discussed by Wilkinson et al. (1999). This useful information should be recognized with publication credit as well as to make the information available to future researchers.

**Reliability**

As noted in chapter 1, the major potential source of random error for TATs is low interscorer reliability. As described in chapter 2, appropriate scorer training and documentation of scoring conventions for data sets are important to secure scientific quality control. The relevance of internal consistency and test–retest forms of reliability for TATs is often questionable (1) because the nature of the storytelling task encourages storytellers to tell
differing stories; (2) because the number of stories gathered is usually relatively small compared with structured self-report scale items, which reduces reliability (Streiner, 2003); and (3) because the properties of the constructs being measured often include a more transient state or situational component, even when they are theoretically cross-situationally consistent traits (Atkinson, 1981; McClelland, 1980; Smith, 1992a; Smith et al., 1992).

However, recent studies that increase the number of “items” (Blankenship et al., 2006; Blankenship & Zoota, 1998; Hibbard et al., 2001), evaluate trait-like constructs (Hibbard et al., 2001), or that make the items (i.e., pictures) more homogeneous to pull for a single construct (Blankenship et al., 2006; Tuerlinckx et al., 2002) have found internal consistency reliabilities comparable to structured self-report measures of similar length and construct properties. In an older study, Dana’s (1959) Objective Scoring variables—Perceptual Organization, Perceptual Range, and Perceptual Personalization—showed high internal consistency reliability with only five cards for these theoretically trait-like narrow-band variables, with no notable loss of validity in discrimination between the criterion groups. Hibbard, Mitchell, and Porcerelli (2001) found that 10 TAT stories yielded good internal consistency for the Social Cognition and Object Relations Scale (SCORS; Westen, Lohr, Silk, Gold, & Kerber, 1990). These scales are somewhat wider band, more ideational-emotional-social, and perhaps a bit more influenced by situational differences than are Dana’s more perceptual variables.

These studies represent important demonstrations that the internal consistency reliabilities of TATs are much affected by the structural differences between TATs and the ability measures for which the idea of reliability was originally created (as noted in chapter 1 in this volume). Structuring the TAT data to be more like that from ability measures increases internal consistency reliabilities. Hopefully these results will settle questions about the reliability of “the TAT.” It is unclear whether these tactics will generalize to other scoring systems, or, more importantly, whether any gain in validity comes along.

Use of tactics that change the structure of the stories comes at a cost, however, especially for clinical applications and for scoring of archival TATs by new systems. Both of these benefit from having a larger sample of more heterogeneous pictures that elicits stories representing responses to a wider variety of the client’s life situations. It is this breadth that supports arguments for validity generalization from stories to clients’ lives, so narrowing the topics evoked by the picture sample also narrows the range of generalizability for the resulting scores. Further, what makes a sample of pictures more evocative of one construct is likely to make them less evocative of other quite different constructs, narrowing the range of constructs for which the stories yield adequate response frequencies.

Similarly, Blankenship et al.’s (2006) and Blankenship and Zoota’s (1998) tactic of separating stories into paragraphs directed by the usual standard probes is an ingenious solution to the problem of “participant burden” that would be induced by asking a sample of uninvested student volunteers to write 20 stories. This indeed solved the technical problem of reliability, although it required some alterations in the need for Power (n Pow) scoring system. For clinical purposes, though, because it imposes stronger constraints on the storyteller’s narrative, this data-gathering technique would make the stories useless—or the scores misleading—for scoring systems that evaluate spontaneously emergent narrative structure. These include at least the Adult Attachment Projective (AAP), Richard Dana’s Objective Scoring, Interpersonal Decentering, and the TI; all are sensitive to the breakup of narrative continuity and a stronger probe structure. The negative impact of the changed structure on scores for the other systems here is less obvious, but would require careful evaluation.
Clinical Research Needed

A major effort is needed to provide empirical support for the three classic tasks of clinical assessment—contributing to diagnosis, evaluating prognosis contingencies, and informing treatment planning—and for the emerging area of collaborative (Fischer, 1994) and therapeutic (Finn, 2007; Finn & Tonsager, 1992, 1997) assessment. The magnitude of the effort needed goes well beyond TAT assessment and the scope of this book. Still, a focused appraisal of these systems is a worthy goal for future research.

Clinical Utility for Diagnosis

Although helpful for identifying contributing risk factors in some cases, TATs are generally not instruments of choice for determining DSM Axis I diagnoses. TATs may show evidence of clients’ meeting diagnostic criteria for some kinds of thought disorder such as social delusions, for which Ritzler’s PPT form quality scores might be useful, or of more primitive problems via Robert Holt’s TAT primary process for psychotic disorders (chapters 16 and 17), but the Rorschach is probably a more sensitive indicator of most forms. For depressive disorders, Ritzler’s PPT emotional stress variables might be useful in establishing criteria related to negative mood among individuals who are reluctant to self-report these, but linkages between emotion scores and idiosyncratically interpreted story content are probably more useful for prognosis and treatment planning. The AAP in this volume shows promise for identifying and describing attachment disorders. Dana’s Objective Scoring in this volume shows validity for global appraisal of disorder severity, discriminating nonpatient, neurotic, and psychotic storytellers using much older diagnostic criteria than those used at present.

TATs’ best use for diagnostic purposes is probably for the Axis II personality disorders. These would include David Harder and Deborah Greenwald’s Ambitious-Narcissism system in this volume (chapters 24 and 25) for narcissistic personality disorder or features thereof; TAT Oral Dependency for dependent personality disorder or features (chapters 20 & 21); Summers’s Symbiosis in chapters 22 and 23 for borderline personality disorder and other disorders involving poor identity differentiation; or Westen et al.’s (1990) SCORS, which was developed to discriminate individuals diagnosed with borderline personality disorder from those with other disorders.

Research designs to test criterion validity hypotheses relating scoring systems to diagnoses would typically be cross-sectional, and might involve historical data and other sources of diagnostic information. One especially useful strategy for the development of new scoring systems for diagnostic purposes is empirical criterion keying (Anastasi & Urbina, 1997, pp. 350–351) or empirical derivation, the process used to develop the original Minnesota Multiphasic Personality Inventory (MMPI), the Oral Dependency system in this volume, and Westen’s SCORS, as well as some of the academic TAT scoring systems (McAdams, 1980; Stewart, 1992; Stewart & Winter, 1974; Winter, 1973, 1992a; Winter & Barenbaum, 1985). This strategy requires two or more clearly defined gold standard criterion groups, such as pure diagnostic groups (without comorbidities) that do not differ on other potentially confounding variables such as demographic characteristics. Stories told to the same stimuli by the two groups are compared to identify unique features that are relatively common in one group and rare in the other. Scoring category descriptions are then written to describe these differences as precisely as is feasible (perhaps using strategies in Boyatzis, 1998). As a check against differences that might appear...
by chance, the resulting categories should be cross-validated on a new sample of similar groups, as described by Anastasi and Urbina (1997, pp. 194–196).

**Clinical Utility for Estimation of Prognosis**

Not only might these scoring systems support diagnosis by giving evidence for diagnostic criteria, they also locate the criteria in a social context that might parallel the storyteller’s life experience. Does the problem appear localized to a particular kind of situation, or is it cross-situationally pervasive? In what story situations is the storyteller’s disorder most prominent or most disruptive? What seems to be the most important aspect of those circumstances: the role relationships of the people involved, the activity that the client says they are involved in, the thoughts or feelings that prevail, or some other aspect?

Read idiographically for consistent patterns of scores across types of stories, this information may be helpful in suggesting prognosis and in forming the treatment plan. What social role relationships are involved in the storyteller’s more positive stories compared with the more negative ones, and does this suggest an easier therapeutic alliance with an older or younger therapist of a particular gender? Under what stimulus conditions—and with what scores on which systems—do the characters do well compared with other conditions? In what ways, if any, might these conditions mirror the conditions in some aspect of the storyteller’s problematic life, either material or mental? Thus, what situations might make the client more likely to benefit, or more vulnerable to collapse or relapse of the sort identified by the system? Which should the client learn to avoid because they evoke malevolent object representations that the client is currently unable to manage? What are the key features of the client’s subjective situation to mimic for exposure therapy with response prevention? In what circumstances might the client more confidently pursue the treatment goals?

The usefulness of TATs for prognostication comes from the fact that each story can be read with content validity as a representation of a particular social context; the series of stories covers a range of same. Here, the variability among stories that is a challenge for internal consistency reliability takes on a positive value. This use of TATs can give hints about the life conditions under which the client might do better or deteriorate, and the context of the story might give hints as to why, which can be explored with the client in therapy.

The challenge for research testing hypotheses about prognosis is of course that longitudinal or time-series case-study designs are typically necessary. However, such designs are within the capability of practice research groups, university training clinics, internship sites, and some individual clinicians, if their clients consent and if appropriate ethical guidelines for research are followed. A double-blind prospective design would likely be beyond the reach of most clinicians who use the initial TAT scores as a guide to their own treatment plans. However, the stories gathered early in treatment perhaps could be rescored retrospectively for other systems not used by the clinician at the time, thus avoiding biases particular to those systems.

**Clinical Utility for Treatment Planning**

The role of TAT scores in treatment planning depends on whether the scores are related to the diagnosis that is the focus of treatment, or whether they represent ancillary
personality information that was gathered to suggest prognosis or treatment response, such as the likely developmental trajectory of a therapeutic alliance. TATs give unique information not only about the storyteller’s psychological strengths and weaknesses but also about the situations in which those might be most beneficial or problematic. There is evidence that treatments are more effective if they build on clients’ strengths rather than depending on their weaknesses. For example, Blatt and Felsen’s (1993) patients with cognitive strengths did better in cognitive behavioral therapy, whereas those with social strengths did better in interpersonal therapy.

Information about the conditions under which story characters do well—as defined by the relevant scoring system—suggests the conditions that might be most comfortable and familiar to the storyteller in psychotherapy. These are not necessarily the most therapeutically helpful conditions for the long run (if one is needed), but if there is evidence that the client might have difficulty forming a psychotherapy alliance, the creation of these conditions—to the extent feasible—might facilitate the initial bonding process.

If TAT scores are related to the diagnosis being treated, they might be useful for evaluating treatment progress in a pretest–posttest repeated measures research design. This relation might reflect either the presence of a problem, such as Primary Process (chapters 16 and 17) or Ambitious Narcissism (chapters 24 & 25), or the attainment of a treatment goal, such as an increase of Personal Problem-Solving ability (see chapters 10 and 11 in this volume), Empathy (chapters 30 & 31 in this volume), or Decentering (chapters 8 and 9 in this volume), or the maturation of defenses (DMM; Cramer, 2006). In such cases, successive administrations should include an instruction either to try to tell the same stories again, or to make explicit that it does not matter whether the stories are the same or different. Resolving this ambiguous implicit assumption for storytellers yields better data (Lundy, 1985; Smith et al., 1992; Winter & Stewart, 1977).

Uses in Therapy Process

Additional studies of key clinical processes are needed. To what extent do characteristic patterns of interaction in stories resemble the storyteller’s patterns of interaction in psychotherapy? The formation of an appropriate therapeutic alliance contributes significantly to outcome in both brief and long-term psychotherapy. TATs may be particularly useful for studies of psychotherapy processes because TAT stimuli elicit internal representations of interpersonal interactions, which may shed light on how a therapeutic alliance develops as well as on the client’s interpersonal presenting problems. Process-oriented measures such as Communication Deviance (chapter 26 in this volume), Interpersonal Decentering (chapters 8 & 9 in this volume), Pathogenesis (chapters 18 & 19 in this volume), Personal Problem-Solving (chapters 10 & 11 in this volume), and portions of Fine’s Scoring Scheme (chapters 14 & 15) might be useful in this regard.

Brief interpersonal approaches to psychotherapy (e.g., Levenson, 1995; Luborsky & Crits-Christoph, 1990; Strupp & Binder, 1984) share a focus on identifying the client’s central relationship difficulty in the form of a schematic outline called a cyclical maladaptive pattern (CMP; Levenson, 1995) or core conflictual relationship theme (CCRT; Luborsky & Crits-Christoph, 1990). These models assume that the client’s interpersonal scenario is reflected similarly in the client’s interpersonal behavior both with the therapist and in outside relationships. Probably at least some of the client’s TAT stories reflect important aspects of the same scenario. Other aspects might reflect the client’s theory of mind, or
why the client thinks other people do what they do. People often respond to what they believe are others' intentions, in addition to their own attitudes, beliefs, personality characteristics, and actions.

Scoring system categories might be useful for therapists to shape the focus of brief dynamic and narrative psychotherapy approaches that help clients rewrite their life stories (Lieblich, McAdams, & Josselson, 2004; White & Epston, 1991). Clients might be shown the scoring category description fitting their problematic storytelling pattern and asked to rewrite their stories to fit the scoring categories associated with more adaptive relationship patterns. In theory, “rewriting the story” should help to reshape both the client's internal interpersonal scenario and the problematic interpersonal behavior pattern. This approach exemplifies working with the client’s material and giving the client control. However, using stories as fictional analog behavioral self-observations might be less threatening and difficult for some clients than a more realistic context would be, and so might avoid the arousal of defenses.

Some useful interventions might combine the brief interpersonal and narrative therapy approaches (i.e., identifying and helping clients to rewrite their stories) with the collaborative and therapeutic assessment intervention strategy (i.e., inviting clients to examine and interpret their own data as a way of understanding and changing maladaptive patterns). Perhaps the best example in practice is Fischer’s (1994) approach to collaborative assessment feedback, in which some of the client’s responses in the assessment session are reflected back to the client as examples of the conclusions in the report. This way of making the output of the assessment process visible is an excellent example of demystification by information.

Methodological Refinements Needed

A full treatment of research methods and methodological considerations for TATs is beyond the scope of this book. Interested readers should consult Smith (1992a), Smith et al. (1992), Veroff (1992), Winter (1992b), and perhaps others in Smith’s (1992b) collection. Some of the most pressing issues are summarized here.

Data Collection

Without question, as Ritzler argues in chapter 12 of this volume, new picture sets are needed that are contemporary in design and content, and are suitable for a broad range of clinical purposes. Some might be tailored to the measurement of specific constructs (e.g., the AAP and PPT in this volume), others to particular cultural groups as well (e.g., the Contemporized Themes Concerning Blacks and the Tell Me A Story [TEMAS] in this volume; see also the Gerontological Apperception Test in Hayslip, Francis, Radika, Lambert, Panek, & Bosmajian, 2002). More important, however, is to open communication among clinical practitioners and researchers about strategic planning when it comes to the design of picture sets. What are the desirable features of a good picture? A good set? What kinds of pretesting are most useful in designing a set? Is it possible to develop a common language and goals for answering these questions?

Most clinical assessments would benefit from having available a large enough sample of diverse pictures to allow for oversampling of particular content areas. Murray’s
(1943) set is not balanced in this way; there are many pictures of individuals or same-gender groups that are seldom used either clinically or in research (Keiser & Prather, 1990). Few Murray pictures suggest work situations, and as noted by Ritzler in chapter 12 of this volume, the emotional tone ranges from neutral to gloomy. There should be a balanced selection of genders, ages, and gender-age combinations among the characters; more situations depicted that show people engaged with impersonal tasks; and a wider emotional range.

A second area that needs careful thinking involves the situational specificity of stimuli, especially pictures. This strength has been virtually untapped in research, though widely used by clinicians. A few studies have used pictures of a narrow range of situations for the purpose of predicting behavior in a specific circumstance (e.g., Libo, 1956, 1957, 1969) or to pull for particular constructs (Blankenship et al., 2006). Too many of these in one set might be ineffective for most clinical purposes given the typical alternating pattern seen in stories (Arnold, 1962; Atkinson, 1981; Reuman, 1982), or if the tactic is obvious to the storyteller.

A better approach would be to create a large enough selection of stimuli to allow for sampling a variety of situation types within the 9–10 pictures that are recommended for clinical or research purposes (Hibbard et al., 2001, cf. Smith et al., 1992), while also allowing for oversampling the one or two areas of interest to which the client’s problem might be localized. The scores for similar pictures can be used as a subscale, and predictions obtained from that subscale can be compared to the scores from the remaining pictures. Thus, subscales can be defined either by the features of the picture, by the content of the story, as done with Thomas’s Affective Scale (chapters 27 & 28), or by the life circumstances to which generalization of the findings is wanted.

Data Analysis

Data analysis is the pivotal link between stories and results from structured scoring systems and is perhaps the least comfortable for clinicians. But failures of conceptualization at this stage have brought down more than one worthy project. This section addresses subscale creation and error variance. This chapter and chapter 1 present arguments for the possible usefulness of subscales that aggregate the scores from similar pictures to represent clients’ likely responses to the type of situation involved.

One question for clinical data analysis then is when subscales are being formed, which is more important, the stimulus or the story? If the storyteller’s plot discusses siblings in response to a picture designated as a romantic couple, what then? The research answer to this empirical question about subscale creation depends on which classification of the story, by stimulus (picture type) or by response (e.g., the Symbiosis System’s dyad type; the roles and valences in Thomas’s Affective Scale), is more consistent with theory or produces the stronger validity coefficient in an atheoretical study. Does the criterion score assess a social structural variable or a subjective perception? The best answer might differ across systems and correlates. For the clinical answer, the clinician might look to see how many romantic stories had already been told and also check the Rorschach space score; was the storyteller being creative or oppositional?

The question of error variance is worth raising because it represents a source of inaccuracy in scores that might be controlled or avoided. Why do scores differ, other
than actual differences on the construct being measured? High on the priority list for the advancement of research on TAT data-gathering methods in general, as well as for each specific scoring system, is a better understanding of what each score represents.

Hopefully most individual differences in scores are due to differences between storytellers on the construct of interest. However, some part of that difference might be due to systematic measurement error (e.g., response style, verbal fluency) or to random error, which is not traceable to a specific systematic source. Probably there are some general sources of measurement error such as verbal fluency (McClelland, 1980; Smith et al., 1992), and response constriction versus dilation (Shatin, 1958; see also chapters 4 & 5 in this volume) that can be controlled statistically, as described by Smith et al. (1992). Random error is best minimized by careful attention to correct and consistent administration so that stories are gathered under standardized conditions for all storytellers, especially regarding instructional set (Lundy, 1988). In the event of doubt about variations in administration, it might be better to eliminate from a research data set those protocols for stories told under nonstandard conditions, as they might yield scores that are outliers on some systems.

It appears that TATs have no obvious social desirability response set as exists for self-report methods, but the question of a general positive–negative affectivity dimension is open for empirical test. It is possible that under some conditions for some systems, examinees may suppress certain kinds of stories and thus may lower their scores on variables involving more unpleasant or obviously threatening imagery. Such efforts at data manipulation may have differing impacts—or no impact—on scores from different systems. Evaluating this issue is an important validational step for TAT systems.

In general, we need a comprehensive, theory-informed picture of what sorts of further validity indices might be needed and for which systems, in which populations, for what purposes, in which circumstances. These may include an extension of the use of moderator variables proposed by Dana (1993, 1998, 2005) for work in evaluating the cross-cultural validity of assessment methods. This tactic can be applied broadly to check the validity generalization of scores to a variety of populations within the same culture as well.

Developing Further Systems

Recovering More of the Clinical Past, Developing the Future

Much interesting work has not been included here, as noted in chapter 1. In particular, my searches located several dissertations that began work on promising systems but remained unpublished. Undoubtedly there are more. The original documents are usually available through university libraries, and some might include adequate scoring manuals as appendices. (Max Prola’s inclusion of his adapted version of Weisskopf’s TI scoring manual in his dissertation’s appendix gave me a version I could use until I located the original in the collection of Weisskopf’s papers.) Some additional background information, perhaps even data usable for training purposes, might be available from the original researchers or their former academic institutions, or from the American Documentation Institute at the Library of Congress. That is where I—or rather, Myron Chace
in his careful sleuthing—located Melvin Feffer’s scoring system, thanks to the latter’s footnote in Feffer and Jahelka (1968). Enterprising graduate students, especially those with an interest in the field’s history, should continue to locate, retrieve, evaluate, use, adapt, and refine these scoring systems.

Arnold’s (1962) Story Sequence Analysis System would be a good example of a system too complex to include in this volume, but it merits further development. Arnold’s method has more in common with a grounded theory approach to open-ended data than with formal or content-analytic approaches. Thus, it captures better the individuality of the storyteller than do the other systems in this book, at some cost to the precision of systematic interindividual comparison. Recent advances in thematic analysis (e.g., Boyatzis, 1998) and narrative analysis (e.g., Josselson, Lieblich, & McAdams, 2003; Lieblich et al., 2004) might enable the refinement of Arnold’s system to allow such comparisons. For example, Arnold provided in her scoring manual for evaluation of stories along a gross positive–negative dimension for characters’ motivations. It seems likely that this dimension might underlie several of the scoring systems in this book.

The more interesting and unique aspect of her system, however, is the analysis of the sequence of imports across the stories in the protocol, viewing these as a series of scenes in a larger play with its own structure, interconnecting themes, and resolution. Here, as with Atkinson’s (1981) and Reuman’s (1982) work on the sequencing of motivation themes in successive stories, is a bigger picture than other approaches give, one that is more challenging to quantify—and probably nonlinear—but potentially fruitful for both clinical and research purposes. For example, the degree to which successive themes connect might reflect a broader openness or dilation versus defensive constriction, as well as integration versus compartmentalization of the client’s experience, in ways quite different from Weisskopf’s (1950) TI, which applies only to individual stories.

Repurposing the Past: Clinical Applications of Academic Systems

The gap between experience and action mentioned in assumptions (5) and (6) at the beginning of this chapter can have both mental and physical health consequences. Structured self-report measures cannot bridge this gap effectively (Shedler, Mayman, & Manis, 1993). Inhibition of action might play a role in the association between stress and mood (Langens & Stucke, 2005). Some of the most interesting TAT studies in the late David McClelland’s career are those linking measures of need for Power ($n_{Pow}$), combined with need for Affiliation ($n_{Aff}$), inhibition of action, and other variables, to health outcomes such as short-term immune functioning and longer-term hypertension and cardiovascular disorders (reviewed in McClelland, 1979, 1985a, 1989; McClelland, Davidson, & Saron, 1985). Although McClelland, not being a clinician himself, repeatedly insisted that motive measures were not designed for clinical use, and the existing literature includes few if any studies of clinical populations, these findings are certainly clinically relevant. It seems time to launch exploratory efforts in this direction using individual clinical administration and the larger picture sets that would allow for generalization to a wide variety of life circumstances.

Another excellent application is McKay’s (1991, 1992) Affiliative Trust–Mistrust System for isolating thought patterns associated with immunocompetence decrements under conditions of positive affiliative arousal. The use of object relations theory to shape the scoring categories of this system stands as a model for clinical health research in this
area that calls for follow-up by interdisciplinary teams including physiological and clinical psychologists. Astute theory development will be important for identifying the processes to study, the points at which measurement should best occur, and what (possibly nonlinear) form the expected associations might take.

Some of the other personality measures included in Smith (1992a) merit research exploration for possible clinical use. Indeed, two studies have found that people higher in the intimacy motive show evidence of better adjustment than those with lower scores (McAdams & Bryant, 1987; McAdams & Vaillant, 1982). Abigail Stewart’s measure of self-definition and social definition provides a view of how a client might navigate stressful situations—especially situations involving social pressure and differences from social conventions (Jenkins, 1996; Stewart, 1978, 1992; Stewart & Winter, 1974).

New Uses for Non-TAT Scoring Systems

Another advantage of less structured techniques is that some scoring systems are robust across data collection methods, supporting the conceptualization of these measures as tapping a person’s habitual thought patterns and ways of organizing and interpreting situations, which generalize across these different methods of sampling them. McClelland (1961) scored folk tales and children’s stories for motives; Winter (1987, 1991) scored presidential inaugural addresses; and LeVine, Strangman, and Unterberger (1966) scored recalled dreams. De Vos has applied his psychocultural scoring system similarly to a variety of ethnographic data such as folklore and ritual as well as literature and dreams (Ephraim, 2000).

This broad utility goes both ways. Several researchers, often graduate students (see Kuhn, 1962) have applied to TAT stories theory-grounded systems developed for other material. Sutton and Swenson (1983) scored TAT stories according to Loevinger’s (1970) ego development scoring system originally designed for a standard sentence completion task. Kalliopuska (1982) used Urist’s (1977) Mutuality of Autonomy (MOA) Scoring System, originated for the Rorschach, and found associations consistent with interpretation of MOA scores as measuring empathy. Black, Jenkins, Evans, and Shafenberg-Murry (2003) adapted for TAT scoring Benjamin’s (1993) Structural Analysis of Social Behavior method for scoring interview data and found these scores related to MOA scores from Rorschach protocols.

In summary, for those not wishing to initiate new scoring systems, as described in previous sections, there are numerous systems in varying stages of empirical development that merit further work. Some of these come from clinical contexts and need more validity evidence, and perhaps more detailed manuals, to improve interrater reliability. Others are from academic research settings and have well-supported validity for their current uses but need testing for applicability to clinical purposes. Additional options include translation to TATs of scoring systems developed for other narrative material. All these options could be tested more readily by increased sharing of data among researchers and clinicians.

Given the labor and expense of gathering and archiving TAT data, far-sighted researchers might gather such data for the initial purpose of evaluating one system, and then make the protocols available to those working with other systems for similar application. Maintaining an archive of such data sets would be a worthy enterprise for a research institute somewhere, facilitating the economical evaluation of scoring systems early in their development sequence. Such an archive already exists and has a large collection of data sets including TAT stories that have been scored for human motives.
and could be scored for other systems. That archive is the Henry A. Murray Research Archive at the Harvard–MIT Data Center. Further information is available on its Web site at http://www.murray.harvard.edu/mra/index.jsp. According to its site, “The Henry A. Murray Research Archive holds a significant collection of social science data for use in the study of human development in the context of social change. The Archive provides access to both quantitative data and qualitative data (case histories, open-ended interviews, and audio- and video-tapes)” for use by qualified researchers.

Science–Practice Gap, Reprise: Changes Needed in Research Process

Research must lead practice when it comes to systematizing and validating new scoring systems using controlled criterion groups, experimental designs, and quantitative methods. But clinical practice must lead research when it comes to defining problems that need solution, developing theory that can guide the scoring system development process, and defining criterion groups and variables for the development of those systems. Idiographic and qualitative methods are more useful in formulating questions and theory for scoring system development than are more structured and quantified measures. The lack of respect that systematic qualitative research methods have received in graduate training in clinical psychology in the last quarter century has been a major hindrance to both theory development and scoring system validation where TATs are concerned. Obviously, the best approach is Boulder model scientist–practitioner training!

The appropriateness of participant samples for particular research questions also is important to address. College student samples may be appropriate for obtaining psychometric and validity data in the middle phases of measure development when large samples are needed—assuming (a large, quite likely erroneous, but frequent assumption) that such samples include an adequate range of variation in the clinical features of interest. However, for the earliest phases of scoring category definition and refinement, and for the later stages of validation, samples of clients or therapists may be a requisite, depending on the nature of the measure and its theoretical foundation.

Finally, clinicians must educate researchers about improvements needed in areas that are usually considered the purview of researchers, but that are currently less productive for clinical practice than they might be—and researchers must listen. For example, journal reviewers need guidelines for making judgments that balance the clinical relevance of manuscripts against the degree of rigor possible given a new measure’s state of development. These guidelines are especially important for the less structured methods because of their greater sensitivity to administration conditions generally, and for TATs in particular because of the stimulus sampling and other technical issues described in chapter 1.

Attention to research findings and their application to practice is increasingly important for clinicians, but unless those findings are externally valid (Sue, 1999), clinically workable, and presented comprehensibly to clinicians, researchers’ efforts may not get the attention that they would merit otherwise. Active collaboration with practicing assessment clinicians can help to enrich psychological scientists’ grasp of how a TAT system’s validity can best be evaluated, as recommended by Westen (2002) for the development of empirically supported treatment manuals. Furthermore, such collaboration can encourage open-minded clinicians to have a broader grasp of what is possible in research, and perhaps to join one of the growing practice research networks.
Some of the very most pressing clinical research needs are those that historically have been caught in the scientist–practitioner gap. The continued viability of both assessment research and scientific professional practice depend on our closing this gap as quickly and solidly as possible. The reward of undertaking work in any part of the above enterprise is participating in the lively process of real clinical science, laboring in the woodshed, planting the vineyard, tracing the phenotypes of crop after crop to improve the breed. There is productive labor aplenty for everyone, from senior clinicians and researchers to undergraduate research assistants, from carrying water and planting the seeds to grafting and harvesting and pruning the unproductive. The final recommendation from this chapter is to continue the mapping of the landscape, with comments on the quality, unique characteristics, and depth of the soil in each area of the field so that those to come may find the most fertile spot for their particular labors.

These worthy activities toward the development and refinement of structured scoring systems may serve as a general foundation for clinical assessment practice. However, they should not threaten the established clinical wisdom of the seasoned practitioner, who may come to quite similar conclusions by a more implicit route. The wise assessor will not ignore these tools, but rather will build on and modify them to improve their fit to particular presenting problems, client populations, diagnostic groups, treatment programs, and institutional needs. Such systems may facilitate teaching and learning, make practice more efficient, and contribute to theory development, but they are merely a means to an end. The experience of the highly trained and sensitive psychologist integrating these systems with other test and non-test data is ultimately the best assessment technique.

References


References


