Journal of Personality and Social Psychology
2002, Vol. 82, No. 3, 448–461

0022-3514/02/$5.00 DOI: 10.1037//0022-3514.82.3.448

Measuring Needs With the Thematic Apperception Test: A Psychometric Study

Francis Tuerlinckx, Paul De Boeck, and Willy Lens
University of Leuven

Three apperception theories that explain how people respond to Thematic Apperception Test cards are proposed: a simple apperception theory, an apperception theory with a dynamic component, and an apperception theory with 2 types of responses. Each theory is translated into an item response theory model and is applied to need for achievement (nAch) data. The analysis indicates that the best fitting model is provided by the apperception theory with 2 types of responses, also referred to as the drop-out apperception theory. The 1st type of response predicted by this theory is determined by the nAch level of the person and the achievement-response-eliciting value of the card; this response is diagnostic for the nAch level of the person. The 2nd type of response is not determined by the 2 aforementioned characteristics and is therefore not diagnostic of the person's nAch level. The results are cross-validated for need for power and need for affiliation.

The measurement of individual differences in strength of needs, such as the need for achievement (nAch; the motive to succeed), need for affiliation (nAff; the motive to establish, maintain, and restore positive affective relationships with others), and need for power (nPow; the motive to control the means of influence) by content analysis of Thematic Apperception Test (TAT; Murray, 1943) stories has been a field of intensive research for many years. Several theories have been formulated about the underlying processes behind these fantasy-based measures (Atkinson, 1958, 1982; McClelland, Atkinson, Clark, & Lowell, 1953; Smith, 1992a). However, much of the controversy about the quality of the measurement instrument remains unresolved (see Lilienfeld, Wood, & Garb, 2000).

The research reported in this article has three objectives: general and specific objectives concerning the measurement of motives with the TAT as well as a methodological objective. The general objective is to capture the most prominent features of the response process and thus to shed light on how people respond to TAT cards and how they arrive at fantasy-based responses to pictures in general. For this purpose, we formulated three process theories on how a person's motive and the card's characteristics interact with each other and lead to a TAT response. In a next step, we linked each of these theories to a different item response theory (IRT) model so that the theory could be put to test. Of course, these IRT models are only approximations of the real response process, which is probably much richer and too complicated to be modeled with relatively simple mathematical models. However, the models might give indications about the main aspects of the underlying process.

Concerning the specific objective, we focus on the low internal consistency of the TAT as an instrument to measure needs. Some researchers (e.g., Entwisle, 1972) have considered the low reliability of the TAT to be an intrinsic problem of the test and a real impediment for its validity, whereas others (Atkinson, Bongort, & Price, 1977) have claimed that the low reliability is the direct consequence of a special kind of response process that does not imply poor validity. Like McClelland (1980), who called the alleged low reliability of the TAT "the issue which in the minds of many is decisive in determining that . . . operant measures [i.e., the TAT] ought not to be used" (p. 28), we feel that it is important to examine more closely the low reliability estimates of the TAT from a psychological process perspective. We evaluate the low reliability in light of the IRT models that represent possible process theories. Once a valid response process is found, one should be capable of deciding between the aforementioned competing perspectives on the low reliability of the TAT. We argue that such a decision also has practical consequences because the impact on validity may depend on the source of the unreliability.

The methodological objective of our article is to show the ease and the flexibility of applying and modifying IRT models to represent meaningful theories about the response process in a test. It is only recently that researchers began to apply these models to data from personality and social psychology (e.g., Fraley, Waller, & Brennan, 2000; Reise & Waller, 1990; Vansteelandt, 1999). As an exception, there are a few examples of older applications of IRT models to personality measures, notably in the field of projective techniques (Fischer & Spada, 1973; Kuhl, 1978), and we discuss these below.

A large part of the research concerning motive measurement with the TAT bears on nAch, and therefore we are also mostly concerned with nAch. However, we attempt to cross-validate our

Francis Tuerlinckx, Paul De Boeck, and Willy Lens, Department of Psychology, University of Leuven, Leuven, Belgium.

Francis Tuerlinckx was a research assistant of the Fund for Scientific Research—Flanders at the time this research was conducted. The research was funded by Grant GOA-2000/2 from the K. U. Leuven. We are grateful to the Interuniversity Consortium for Political and Social Research for the use of their data. We thank David Birch, Sam Sommers, and Tom Verguts for their critical reading and useful suggestions.

Correspondence concerning this article should be addressed to Francis Tuerlinckx, who is now at the Department of Statistics, Columbia University, 2990 Broadway, MC 4403, New York, New York 10027. Email: tuerlinckx@stat.columbia.edu

results for two other needs (nPow and nAff). The results of the analyses for these two needs are discussed briefly in the *Results* section.

We begin with an overview of the theoretical foundation of fantasy-based measurement of nAch. Next, the main criticisms of the validity and reliability of the TAT are reviewed. Further, we present three different theories about the response process for the TAT, all three having a common core. Subsequently, we formulate IRT models that correspond to the three basic theories, and this is followed by a section on testing these models for nAch data. We close the article with a discussion of our results.

Theoretical Foundation of nAch Measurement

Murray (1938) defined a need as a "construct . . . which stands for a force . . . which organizes perception, apperception, intellection, conation and action in such a way as to transform in a certain direction an existing, unsatisfying situation" (p. 124). The nAch was defined by Murray (1938) as the wish "to accomplish something difficult. . . . To overcome obstacles and attain a high standard. To excel one's self. To rival and surpass others" (p. 164). This need gives rise to achievement behavior. The TAT was developed by Morgan and Murray (1935) for the assessment of nAch and other needs. For McClelland et al. (1953), the nAch is a general and relatively stable personality disposition that is learned on the basis of affective experiences. The nAch is satisfied if the person is successful in competition with a norm of excellence, which results in a positive affect (i.e., pride in accomplishment). Failure leads to a negative effect (i.e., shame). McClelland, Atkinson, Clark, and Lowell (1958) refined the measurement of individual differences in nAch with the TAT by assuming that a TAT card may induce an expression of the need in the fantasy behavior of the person.

In a typical TAT administration situation, the person is presented with four to six pictures, one after the other. The person has to write down what comes up in his or her mind when seeing each picture. The respondent is invited to be creative, and four questions are given to help the respondent in writing a story about the picture.[1] The scoring system is based on a content analysis of the written stories (Atkinson, 1958). A story receives a score from -1 to 11, dependent on the amount of achievement imagery. When the content is not achievement oriented or is doubtful with respect to nAch, it is scored -1 or 0, respectively, and the analysis stops. When the content clearly contains achievement imagery, a score of 1 is given, and the score is raised by 1 for each of 10 additional categories that can be identified in the story.

Criticisms on the Use of the TAT for the Measurement of nAch

From the perspective of classical test theory, the measurement of nAch with the TAT has been attacked on two major psychometric aspects, reliability and validity. Classical test theory is about the total score (i.e., the sum of the scores on the items of a test), and, in this case, the cards are considered as the items.

Validity

Researchers have conducted several meta-analyses of studies that investigated the predictive validity of the TAT by correlating nAch scores with performance in achievement or performance tasks such as school or job success. Klinger (1966) concluded that correlations between TAT measures and performance outcomes were significant only in 50% of the studies. Fineman (1977) reviewed studies in which more than one measure of nAch was involved (TAT, questionnaires, other projective techniques) and found no evidence for convergent validity.

McClelland (1980), McClelland, Koestner, and Weinberger (1989) and Winter, John, Stewart, Klohnen, and Duncan (1998) have argued that there is a difference between needs or implicit motives (as measured by the TAT) and self-attributed or explicit motives (as measured by questionnaires) and that they correlate with different outcomes. In a recent meta-analysis of 105 studies, Spangler (1992) found support for these ideas. The validity issue is not the main topic of this article, but we take it up briefly again in the Discussion section.

Reliability

Another psychometric property of the TAT that has been a point of intensive debate is its reliability. The reliability of the scoring system is usually sufficiently high (see Entwisle, 1972; Fineman, 1977; Smith, 1992b; Veroff, Atkinson, Feld, & Gurin, 1960), but the measurement reliability is often viewed as problematic.

Two types of reliability estimates are mainly used with respect to the TAT: internal consistency estimates and test–retest correlations. We discuss both, starting with the test–retest correlation. From Entwisle (1972) and Fineman (1977), it is concluded that test–retest correlations for the TAT fluctuate around .30. Winter and Stewart (1977) hypothesized that the instructions of the TAT prevent the test–retest correlation from being high because the participant is asked to be creative. In the case of measuring nPow, Winter and Stewart (1977) instructed their participants during the second test session to write the same stories as the previous ones or to write whatever they would like. These instructions led to an increase in the test–retest correlations for the measurement of nPow. However, Kraiger, Hakel, and Cornelius (1984) could not replicate these results for nPow. These studies have not been replicated for nAch, either. Also, one might wonder why allowing people to change the content of the story (i.e., the appearance of the motive) has an influence on the motive itself. If a person has a high nAch, this should be revealed under any circumstance, whether or not the person is allowed to tell the same story again.

Entwisle (1972) concluded that the fantasy-based measure of nAch has low reliability, mostly below .30 or .40, when estimated with internal consistency measures such as Cronbach's alpha. From the studies reviewed by Fineman (1977), it is concluded that the median internal consistency is .32. Entwisle (1972) warned that the low reliability may explain the low predictive validity of the nAch TAT scores because in classical test theory, the square root of the reliability is an upper bound for the predictive validity.

Because the low internal consistency problem results from low intercorrelations among the cards (Entwistle, 1972; Jensen, 1959;

[1] The questions are as follows: (a) Who are these people? [Who is this person?] What are they [is he/she] doing? (b) What has led up to this—what went on before? (c) What do they [does he/she] want—how do they [does he/she] feel? (d) What will happen? How will it end?

Mitchell, 1961), greatly increasing the number of cards could be a solution to the problem. However, this is often not practically possible because a person may become tired after too many pictures (Fineman, 1977; Smith, 1992b) and, according to Atkinson (1954), the nAch score may drop considerably after four cards. Entwisle (1972) and Smith (1992b) claimed that six pictures are probably an upper limit.

Smith (1992b) reviewed the critiques regarding the reliability of thematic apperceptive measures of motives. He concluded that the reliability of many of these motive measures is "indeed relatively low, but not as low as critics have alleged" (p. 126). He reported Cronbach's alphas in the .50 and .60 range in studies with high interscorer agreement, an appropriate picture selection, and six or more pictures. The last issue contradicts earlier statements (Atkinson, 1954; Smith, 1992b) that a single test session with more than six pictures should be avoided.

The views on the low internal consistency of the TAT can be summarized (and thereby inevitably simplified) by dividing them into two groups. On the one hand, there are the TAT nonbelievers (e.g., Entwisle, 1972; Fineman, 1977), who reject the TAT as a measurement instrument by saying it is unreliable. On the basis of the classical test theory, they claim that the TAT cannot be valid if it has a low reliability because it only measures error. On the other hand, there are the TAT believers (e.g., Atkinson, 1981; Atkinson et al., 1977; Blankenship & Zoota, 1998; Cramer, 1996, 1999; McClelland, 1980, 1985; Reuman, 1982), who say that the low internal consistency shows that the classical test theoretical framework is not appropriate for assessing the quality of the TAT as a measurement instrument and that the TAT can be a valid instrument despite its low internal consistency. The chief argument in this reasoning comes from the dynamics of action theory of Atkinson and Birch (1970). This theory and its implications for internal consistency are discussed in the next sections.

Arguments other than those based on the dynamics of action are also used to dismiss classical test theory as being inappropriate for the TAT. However, we do not believe that those arguments are strong enough to reject the classical test theoretical framework. For instance, Lundy (1985) argued that from the classical test theoretical perspective, the internal consistency cannot be smaller than the test–retest correlation. Next, he observed in his sample that the internal consistency was smaller than the test–retest correlation and concluded from this that the assumptions of classical test theory are not valid for the TAT. However, as already noted by Cronbach (1951), coefficient alpha "may be either higher or lower than the coefficient of stability [i.e., test–retest correlation] over an interval of time" (p. 309). Another example is McClelland (1985), who claimed that card scores in a test are practically unrelated and suggested using the separate card scores in a multiple regression analysis to optimize the prediction of a criterion. However, using the separate card scores as predictors simply shifts the reliability problem to another level, because one can ask whether these individual card scores are reliable. Furthermore, it is left unexplained why the cards measure one underlying trait although they are not related.

We believe that the questions regarding the low reliability of the TAT cannot be solved using a classical psychometric framework. It is our goal to find an appropriate model for the TAT that reflects a theory about the TAT response process. Several such theories are outlined in the next section.

Three Theories on the Response Process in the TAT

In this section, the theories about how people respond to a TAT card are discussed from the perspective of nAch, but they apply to other needs as well (e.g., power or affiliation).

A Basic Apperception Theory

The first theory that we outline is the simplest one (perhaps too simplistic). It is the building block of the two other theories that follow, and therefore it is discussed in greater detail. A starting point for the basic apperception theory is the programmatic formula of Lewin (1935):

$$B = f(P, E), \tag{1}$$

where B stands for the imaginative achievement behavior of a person in a given situation, P stands for the relevant characteristics of the person for that behavior, and E stands for relevant features of the environment. The expression $B = f(P, E)$ means that the imaginative achievement behavior is a function of both features of the person and features of the situation.

To derive a workable hypothesis from Equation 1, we need to delineate the relevant person and situation features that determine achievement fantasy behavior. The relevant person characteristic for achievement-oriented behavior is the strength of the person's achievement motive, called the nAch level. The environmental term E refers to the TAT cards being different in the extent to which they elicit the achievement-oriented fantasy behavior. These differences in elicitation potential of cards, cue value, press (Murray, 1938), or card pull (Cramer, 1996) are recognized in nAch research. Veroff et al. (1960) noted that "pictures will differ, depending on their content, in the average amount of motivational imagery they elicit from any group of subjects" (p. 2). Atkinson (1965) asserted that the nature of this eliciting force is the arousal of an expectancy of goal attainment (leading to satisfaction) in the test taker when emitting the imaginative achievement behavior. The eliciting force or the card pull of the pictures is called the instigating force in the remainder of the article.

Finally, we need to explain how person and picture features integrate and drive the achievement imagery; this is the task of clarifying the function f in Equation 1. The instigating force of the picture arouses the motive so that their joint effect becomes the tendency or motivation to emit achievement fantasy behavior. Thus, the tendency or motivation is the motive that is instantaneously aroused through the instigating force of the card. A motive is conceived of as a stable characteristic of the person, whereas a motivation refers to a nonstable, temporally aroused motive (see also Atkinson, 1957).

If the momentary tendency or the joint effect of the instigating force of the card and the nAch level of the person is large enough, a threshold will be exceeded and an achievement response is likely to follow. This means that if a card has a low instigating force, it is not too likely that the threshold will be exceeded for persons with a moderate nAch level and, therefore, only those persons with a high nAch may be expected to give an achievement-related response. By contrast, if a card has a high instigating force, the threshold for giving an achievement-oriented response will be easily exceeded; therefore, many people (even those with a low nAch level) will show achievement imagery in their response.

Such a threshold mechanism has already been proposed by Tyler, Tyler, and Rafferty (1962; see also Vislie, 1972) for the measurement of needs, although not in the context of the TAT.

Concerning the variables that are essential in the response process, a simplification is made. Implicitly, we assumed that only the nAch level of the person and the instigating cues in the card pertaining to the nAch are important in the process. However, it may be the case that the aroused achievement motive competes with all other aroused motives and that the strongest tendency determines the actual response (Atkinson & Birch, 1970). We neglected the possibly competing nature of the tendencies because, in fact, a single story can contain elements that originate from different tendencies. The same protocol can obtain a positive score for several different needs, and this contributes to the relative independence of the tendencies in their expression on the TAT.

A Dynamic Apperception Theory

The theory presented in this section is based on the dynamics of action theory of Atkinson and Birch (1970). In this theory, a dynamic perspective toward motivation is taken instead of the usual static one, and the main issues concern the change of behavior and the influence of behavior on the underlying tendencies. The dynamics of action is a strongly formalized and general theory about motivation, but in this article only the most basic principles are used.

As before, we assume that the achievement motive underlies the achievement imagery and that the motive is aroused by the cues in the TAT cards. The extent to which the cues arouse the motive is again called the instigating force. The new element is that the emitted fantasy behavior itself can change the tendency strength. Atkinson and Birch (1970) assumed that acting in some way reduces the tendency to act in the same way later. Thus, a tendency is a dynamic concept in which strong values are likely to be followed by weaker ones. Behaviors satisfy their underlying tendencies, and the force by which this happens are called the consummatory force.

The dynamics of action theory was implemented as a computer program, and Atkinson et al. (1977) showed that the computer model of the theory predicted low internal consistencies for TAT nAch scores. This can be explained as follows: When on the first card a nAch-oriented response is given, that response will satisfy the underlying nAch so that it becomes less probable that a nAch-oriented response will be given to the next card, because the tendency strength has declined so much that the threshold will not be exceeded. On the following card, the motive can be aroused again in the usual way. This phenomenon, called the sawtooth effect by McClelland (1980), leads to low correlations between subsequent cards and thus to a low internal consistency.

It remains an open question whether a consummatory effect is truly present in the nAch measurement with the TAT. Therefore, we need to develop an appropriate theory for the response process under the assumption that satisfying effects are present. In a first step, we expand Equation 1 along the lines of the principles of the dynamics of action:

$$B_j = f(P, E, B_{j-1}), \tag{2}$$

where B_j stands for the behavior of a person in a given situation on card j. The symbol P denotes the nAch level of the person, and E the instigating force of the picture cues. Moreover, B_{j-1} is the behavior of the person on card $j - 1$. In Equation 2, the joint effects of the person characteristics, the relevant environmental features, and the influence of the previously emitted behavior constitute the tendency to emit achievement-oriented behavior.[2] Note that we only consider the behavior that happened just before the current one as important. This restriction is common in many theories of time-dependent behavior (see, e.g., Wickens, 1982).

The achievement-oriented behavior on card $j - 1$ has a consummatory influence if the underlying tendency for card j decreases compared with the case in which there was no achievement-oriented behavior on card $j - 1$. If such a mechanism is active, one says that achievement-oriented behavior has a refractory phase (Cramer, 1996; McClelland, 1980; Murray, 1938; Winter & Stewart, 1977), because after the behavior is emitted, it is less likely that it will be emitted at the next occasion because of a decline in the tendency. Such a periodicity may result in the sawtooth effect (McClelland, 1980), in which achievement and nonachievement fantasy behavior alternate.

A Stochastic Drop-Out Apperception Theory

For the last hypothesis we refer to Murray (1965), who stated that

> only a fraction—as a rule a relatively small fraction—of the aggregate of words, phrases, and sentences that make up a set of stories represent important constituents . . . of the patient's past or present personality. As a rule, most of the obtained material consists of statements that are not representative of anything that needs to be included in a formulation of his personality. In short, the larger fraction of the protocol is chaff; the smaller fraction, grain. (p. 430)

Thus, according to Murray, a substantial part of a person's responses to a set of TAT cards does not reveal anything about the strength of his or her needs. Murray (1943) estimated that "under average conditions about 30 per cent of the stories . . . will fall in the impersonal category" (p. 15). This impersonal category is defined as psychologically irrelevant and composed of story elements that are not representative of the motives in an individual's personality (e.g., things shown in the picture).

In our reformulation of Murray's (1965) claim, it is assumed that not every time a person responds to a picture do the nAch level (P) and the instigating force from the picture (E) determine the content of the fantasy behavior. This means that stories told in response to a card do not necessarily contain information about the nAch level. One way of understanding this assumption is to conceive of the answering process as a two-step process. First, either the picture appeals to the apperception by the person or it does not: In the context of nAch, this means that the need-related fantasy set of the person is either activated or not. If it is activated, then the second step is similar to what has been described in the

[2] To be fully in line with the dynamics of action theory, the time scale should be continuous. However, in the context of the TAT it is much simpler to consider the administration of one card as one time point, so that a discrete time scale results, with as many points as there are cards. Also, we assume in the following that the number of the card also determines its position in the order of presentation.

basic apperception theory: Dependent on the instigating force of the card and the nAch level of the person, an achievement-oriented answer may be given or not. If the need-related fantasy set of the person is not activated, then a non-achievement-oriented response follows (i.e., the person responds with an irrelevant story for nAch). Then the response of the person does not reveal anything about his or her nAch level because the nAch was not involved in the response-generating process. The response of the person to the card may then be explained by other factors—for example, actual but irrelevant events that one remembers.

This hypothesis is called the *stochastic drop-out hypothesis*. We speak of a drop-out because sometimes the person's achievement level is not reflected in the response, rendering it nondiagnostic. Moreover, this drop-out is stochastic because in some cases the nAch level and the instigating force determine the response of the person, but in other cases they do not and the response is not fully predictable. Note that when the content of a story does not contain achievement-oriented fantasy behavior, this may have two different causes: Either there was a failure to activate the need-related fantasy set of the person or the need-related fantasy set was activated but the combination of the need and the instigating force was not strong enough to emit achievement imagery.

In accordance with stochastic drop-out theory, the formula of Lewin (1935) can be extended as follows:

$$B = f(P, E) \text{ or } B' = f(P', E'), \tag{3}$$

where P' and E' denote some causes for the nonachievement imagery B' that are related neither to the achievement motive of the person nor to the achievement-instigating force of the card. The behavior is nondiagnostic in those cases, whereas it is diagnostic for the case in which $B = f(P, E)$.

Three IRT Models

Each of the three theories that were proposed in the previous section can be translated into an IRT model. A recent introduction to IRT was provided by Embretson and Reise (2000). IRT models have also been applied to projective techniques in the past. Fischer and Spada (1973) applied some IRT models, which are similar to the first model proposed in this article, to Rorschach and Holtzman inkblot tests. A special version of the TAT was also subjected to an analysis with the same kind of model (Kuhl, 1978).[3] However, the projective method studied by these authors differs from the one studied in this article. Furthermore, there are two problems with the earlier analyses. First, in all the cases, only some basic models were estimated, and more complicated models that may represent more truthfully the underlying response process for projective tests were not considered. Second, at the time the research was done, there were few possibilities to evaluate the appropriateness of the proposed models.

All models in this article are designed for binary (0–1) data. As mentioned in the Theoretical Foundation of nAch Measurement section, the scoring system for nAch assigns a value between −1 and 11 to the content of a story. For our analysis of the data, we have dichotomized the scores as follows: If a score of −1 or 0 is obtained in the first scoring step, a new score of 0 is assigned, and if a score of 1 is obtained in the first scoring step, a new score of 1 is assigned. Remember that in case a 1 is assigned, this can be augmented to a maximum of 11 in a second step. We do not take this second step into account. This dichotomization procedure is far from arbitrary given that the dichotomized score corresponds to the absence or presence of achievement imagery. Although some information is lost by dichotomizing, Entwisle (1972) reported correlations around .90 between dichotomized and nondichotomized scores, implying that the information loss is small and that the full and dichotomized scores are nearly equivalent.

To facilitate the presentation of the technical part of the IRT models, we need to introduce some notation. An arbitrary person is denoted by v, and an arbitrary card by j. The presence or absence of achievement imagery for person v and card j can be represented in a straightforward way with a (random) variable X_{vj}, which takes the value 1 if achievement imagery is present and the value 0 if achievement imagery is absent. Hence, if Person 1 has emitted achievement fantasy on Cards 1, 2, 5, and 6 but not on Cards 3 and 4, the (random) variables X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, and X_{16} take values 1, 1, 0, 0, 1, and 1, respectively.

The models that we present assume that a person v can be assigned a value that denotes the nAch level that is symbolized by θ_v. The instigating force of a card j can also be assigned a value, and it is denoted by β_j. Next, we discuss the three IRT models that correspond to the earlier proposed theories.

The Basic Apperception Model (BAM)

In the basic apperception theory, the joint effect of only the nAch-level θ_v of the person and the instigating force of the card β_j underlies achievement imagery, and the joint effect is considered to be the simple addition of both quantities. Hence, we can say that the tendency of person v to give an achievement-related response on card j, denoted as T_{vj}, equals

$$T_{vj} = \theta_v + \beta_j. \tag{4}$$

The tendency to achieve T_{vj} maps onto the probability for person v to give an achievement-related response on card j in the following way:

$$\Pr(X_{vj} = 1) = \frac{\exp(T_{vj})}{1 + \exp(T_{vj})}, \tag{5}$$

where $\Pr(X_{vj} = 1)$ is the probability that person v gives an achievement-oriented response on card j. If the tendency T_{vj} exceeds the threshold zero, then the probability of an achievement imagery becomes larger than .50, so that it is likely that achievement imagery follows. However, if the threshold is not exceeded, it is not very likely that an achievement fantasy story will be told. The model from Equation 5 is known in the literature as the Rasch model (Embretson & Reise, 2000).

[3] Kuhl (1978) performed analyses on data from the Heckhausen TAT (Heckhausen, 1991), which is essentially a two-dimensional concept because of the different coding system used.

The Dynamic Apperception Model (DAM)

The second IRT model is an elaboration of the BAM according to the dynamic apperception theory.[4] First, we discuss what the BAM predicts about dynamic effects. For simplicity, it is supposed for the moment that there are only two TAT cards, numbered 1 and 2, and that they are administered in that order.

To study predictions about dynamic effects in a series of TAT cards, it is necessary to consider the response given on the previous card. For the BAM, the tendency strength for person v on Card 2 is as follows:

$$\begin{aligned} T_{v2} &= \theta_v + \beta_2 \text{ if } X_{v1} = 0 \\ T_{v2} &= \theta_v + \beta_2 \text{ if } X_{v1} = 1. \end{aligned} \tag{6}$$

Equation 6 indicates that the tendency strength for person v on Card 2 is independent of the response that is given on Card 1. Regardless of the preceding response, the aroused tendency to achieve is simply $\theta_v + \beta_2$. Hence, the BAM predicts that there are no dynamical phenomena during test taking.

Next, we discuss the DAM, which incorporates dynamic effects. The key difference from the BAM is that given that an achievement imagery occurred on Card 1, the tendency to achieve on Card 2 changes compared with the case in which no achievement imagery occurred on Card 1:

$$\begin{aligned} T_{v2} &= \theta_v + \beta_2 - \beta_{12} \text{ if } X_{v1} = 0 \\ T_{v2} &= \theta_v + \beta_2 + \beta_{12} \text{ if } X_{v1} = 1. \end{aligned} \tag{7}$$

Thus, from Equation 7 it can be seen that the tendency to achieve for person v on Card 2 depends on the response that the person has given on Card 1. The parameter β_{12} is called the interaction parameter because it quantifies the interaction between the two TAT cards. Dependent on the value of β_{12}, three important cases can be distinguished.

First, if $\beta_{12} = 0$, Equation 7 becomes equal to Equation 6, implying the absence of dynamic effects. Second, if $\beta_{12} < 0$, the behavior has a consummatory force, because the tendency to achieve on the second card will become smaller if an achievement-oriented response was already given on the first card. Third, an intensification of the tendency occurs if $\beta_{12} > 0$. Although the dynamics of action theory is about satisfaction, our formal model does not exclude the opposite (a behavior stimulating the same behavior). If $\beta_{12} > 0$ and an achievement-oriented response on the first card is given, the tendency to give one on the second card increases.

Again, the tendencies in Equation 7 can be transformed into probabilities, as is done in Equation 5 (but now conditional on the response to Card 1). The resulting conditional probabilities capture the critical part of the DAM needed for our purposes. However, these probabilities are not sufficient to characterize the full IRT model, and therefore we have given the complete model in Appendix A. For an extensive discussion of the model, see Hoskens and De Boeck (1997).

As indicated before, the dynamics of action theory of Atkinson and Birch (1970) leads to the prediction of a so-called sawtooth effect in TAT data due to the assumed consummatory force of achievement fantasy behavior. This sawtooth effect means that an achievement response has a greater probability of being followed by one that is not nAch oriented than by one that is nAch oriented. Conversely, if no achievement-oriented behavior is emitted on Card 1, the probability that a person will respond in an achievement-oriented way on Card 2 is higher than one would expect independent of the previous response. In our binary case, the sawtooth should be revealed (ideally) in response patterns such as 101010 or 010101.

To illustrate that such response patterns are predicted by the DAM, we conducted a small simulation study and give the results in Table 1. For 10 samples of 1,000 hypothetical participants, six responses according to the BAM and the DAM were simulated, and the mean relative frequencies of (complete or partial) sawtooth related patterns are shown. For each of the 10 replications, the nAch levels of the persons (θ_vs) were drawn from a standard normal distribution, as were the eliciting forces (β_js). The interaction parameters for consecutive cards are all equal to -1.5. From Table 1, it can be concluded that if the interaction parameters of the DAM are negative, this model predicts more occurrences of alternating response patterns than does the BAM.

The Stochastic Drop-Out Apperception Model (SDAM)

The last model we present is derived from the stochastic drop-out apperception theory. The theory states that some responses are not diagnostic for the nAch level of the person because the nAch level and the instigating force are not involved in the generation of the response. For person v and card j, this can be expressed formally as follows:

$$\begin{aligned} T_{vj} &= \theta_v + \beta_j, \text{ with probability } \lambda_j \\ T_{vj} &\text{ is not active, with probability } 1 - \lambda_j. \end{aligned} \tag{8}$$

so that with probability λ_j the response is diagnostic and with probability $1 - \lambda_j$ it is not.

The translation into an IRT model is not as straightforward as for the first two models. However, the model can be explained easily by means of a tree model (see Figure 1) representing the two-step process as discussed in the theoretical section. The left-most split of the tree represents the first step: The card may or may not activate the need-related fantasy set of the person. If the need-related fantasy set is not activated (lower branch of the tree), the given response has no achievement-related elements and is coded as 0. However, if this fantasy set is activated (upper branch), the second split of the tree represents the second step in the process: The tendency may or may not lead to an achievement-oriented response to the card, resulting in a coded response of 1 or 0, respectively.

As can be seen from Figure 1, the probability of giving achievement imagery equals

$$\Pr(X_{vj} = 1) = \lambda_j \frac{\exp(T_{vj})}{1 + \exp(T_{vj})}, \tag{9}$$

[4] The mathematical model that we derive in the next section from our version of the dynamics of action is different than the one of Atkinson and Birch (1970).

and the probability of a nonachievement imagery is simply the complement: $1 - \Pr(X_{vj} = 1)$. In Equation 9, λ_j is the probability for card j of taking the branches in the tree that lead to a response that reflects something of the person's nAch level in the given response (a diagnostic response). The probability of a drop out on card j is equal to $1 - \lambda_j$. Therefore, λ_j is denoted as the non-drop-out parameter for card j. The other parameters have the same interpretation as in the original BAM. If $\lambda_j = 1$, the model simplifies to the BAM (Equation 4) for card j. Hence, the BAM is a special case of the SDAM model with no drop-outs.

The observed responses are achievement oriented or not (i.e., they are scored 0 or 1). According to the SDAM, there is a contamination in the observed data between non-nAch-oriented responses that are the result of taking the non-nAch-related path in Figure 1 (so-called nondiagnostic responses) and non-nAch-oriented responses that are the result of taking the nAch-related path and ending up in a failure to express the need in overt behavior (diagnostic responses). This means that when a response is observed that contains no achievement imagery, it may be an informative response for the nAch level or not. The model does not allow us to tell for each individual response whether it is diagnostic, but for different cards the drop-out probability can be estimated.

The model in Figure 1 assumes that the non-drop-out probabilities differ across items. However, an alternative assumption is that the probability of giving a diagnostic response is the same for all pictures. This simpler version of the SDAM is called the restricted SDAM, and it has a non-drop-out probability parameter λ, which is common for all items and which gives an overall estimate for the probability of an achievement diagnostic response.

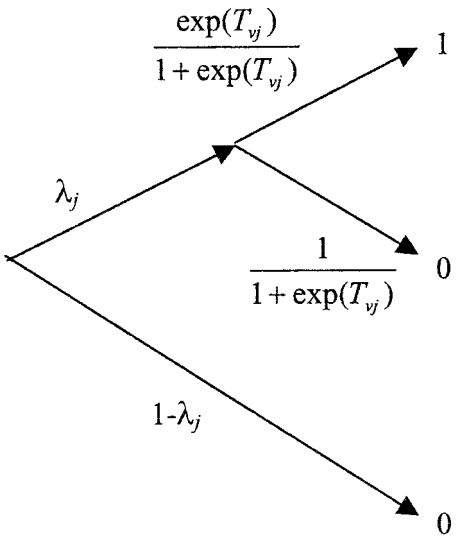

Figure 1. Schematic representation of the stochastic drop-out apperception model. λ_j = the non-drop-out probability for card j; $1 - \lambda_j$ = the dropout probability. T_{vj} = the tendency to achieve for person v on card j.

Empirical Study: Testing the Theories

Method

The data used in this study come from a large survey study called Americans View Their Mental Health from 1957 (Gurin, Veroff, & Feld, 1975). The purpose of the study was to evaluate how Americans in the 1950s perceived their mental health and which actions they undertook to handle problems of mental illness (Gurin, Veroff, & Feld, 1960). Next, a summary of this study is given (see also Veroff et al., 1960).

Participants. From a total of 2,460 interviewees, 1,619 (randomly selected) were given a set of TAT pictures. The sample consisted of 904 women and 715 men, interviewed by 159 interviewers. The mean age was estimated to be approximately 44.1 years.[5]

Materials. Two test forms were used, each containing six 4-in. × 6-in. (10.16-cm × 15.24-cm) cards, with one common card for the two forms. One form was given to the men and the other to the women; hence, men and women received five different cards and one common card. Besides nAch, nAff and nPow were also measured with this set of pictures (see Gurin et al., 1957, Study 2). The pictures are described by Veroff et al. (1960) and are reprinted in Smith (1992a, Appendix II). The cards were selected from a larger pool of cards to have equal meaning in different social groups (to avoid bias) and to have a strong cue to one need and minor cues to the other two. After a pretest, the researchers concluded that the order of presentation had no systematic effect on the nAch score (Veroff et al., 1960; Veroff, Feld, & Crockett, 1966).

Procedure. In the actual interview setting, the instructions given by the interviewers were highly standardized. The pictures were shown for 20 s, and then the four guiding questions were asked. The total response time was limited to 22 min for the six pictures.

The scoring of the protocols for nAch was done by three trained coders who each coded two pictures for both sexes, using the detailed scoring manual (McClelland et al., 1958). The total number of stories was divided into three parts, and from each part a subset of the protocols was coded by a fourth expert coder. Hence, a coding reliability check could be performed. The Spearman rank-order correlations between the total scores from the expert coder and the other coders were .89, .77, and .81 for the three parts, which is sufficiently large.

Analysis. We checked the dichotomization of the data by computing correlations between the total nAch score for the dichotomized and non-dichotomized scores. We also performed a classical psychometric analysis in which the proportion of achievement-oriented answers on each card and the reliability of the test were computed, separately for men and women.

On two points, our analysis was not in line with the original one by Veroff et al. (1960). First, there were some minor, although statistically significant, correlations between length of the protocol and the nAch score ($r = .28$ for men and $r = .25$ for women; see Veroff et al., 1960). However, the dichotomization was not based on the corrected scores because it would not make any difference. The correction for protocol length would only affect scores of 1 or higher, but it would not replace a clear achievement-

Table 1
Mean Relative Percentages of Alternating Response Patterns for the Basic Apperception Model (BAM) and the Dynamic Apperception Model (DAM)

Response patterns	BAM	DAM
01 and 10	34.41	62.53
101 and 010	18.91	43.73
1010 and 0101	9.20	29.21
10101 and 01010	5.25	21.15
101010 and 010101	2.77	15.37

[5] Because we only had age intervals and the number of people within such an interval available (Veroff et al., 1960), we had to approximate the mean age by weighting the midpoint of each interval with the number of people in it.

oriented response with a non-achievement-oriented response or vice versa. Second, in the original study, some protocols were considered to be inadequate for further analysis because the imaginative content could not be scored for any of the three needs. In our analysis, all protocols were used.

To begin with, separate analyses were performed on the male and female samples. We decided to do this for three reasons. First, men and women received only one common card. Second, it is possible that there are differences in the response mechanism between men and women. Separate analyses can reveal such qualitative differences. Third, if the response processes appear to be similar, an analysis on two data sets is a kind of cross-validation.

The three IRT models (BAM, DAM, and SDAM) were fitted to the data. Model fitting implies two stages: estimation of the parameters (finding appropriate values for the unknown quantities like β_j or λ_j appearing in the equations) and testing the appropriateness of the model. In the following, both points are discussed in a nontechnical manner. A more complete discussion of the matter is given in Appendix B, but it can be skipped without loss of continuity.

For the estimation of the models' parameters, the SAS procedure nonlinear mixed models (NLMIXED) was used, as included in SAS V8e (SAS Institute, 1999). We refer to Rijmen, Tuerlinckx, and De Boeck (2001) for an introduction to estimating IRT models with SAS. Following the estimation, the models were tested on four criteria. First, an ordinary chi-square test statistic was computed to test the global fit of the model. The result of the chi-square test can be summarized in a *p* value, which gives the probability that this value or one larger than the test statistic will be observed given that the model is true. If the *p* value is very small (e.g., below .05), it is unlikely that the model is correct.

Second, we wanted to compare different models that are fitted to the same data set. Two models may have comparable chi-square statistics and both may fit the data, but one of the models may have more parameters than the other one. In that case, the more parsimonious model (with fewer parameters) is to be preferred. A statistic that allows for comparison of different models is Akaike's information criterion (AIC; Akaike, 1977). The AIC penalizes the goodness of fit of the model for the number of parameters. The model with the smallest value of the AIC is preferred because it has the best equilibrium between the model fit and the number of parameters.

Third, a more specific test was carried out. To assess the problem of the low internal consistency, we checked whether the proposed model could explain the observed value of Cronbach's alpha of TAT data that were analyzed. For this purpose, 2,000 new data sets were simulated on the basis of the model, and from these new data sets, 2,000 predicted values of Cronbach's alpha were obtained. Then we determined the proportion of predicted alphas that were smaller than the observed alpha. Hence, the result of the test on the internal consistency can also be expressed in a *p* value. If the model is correct, we expect that approximately half of the simulated values will be smaller than the observed one.

Fourth, the test–retest correlations of the model were predicted. Because we had no observed test–retest correlations for our data, the predicted test–retest correlations were compared with the values found in the literature.

In a following stage, the data from men and women were analyzed together, which was possible because they had one TAT card in common. The models applied for this joint analysis were only compared with respect to their AIC value, and no further tests were performed. Modeling both sexes together allows us to compare achievement levels of men and women. One could have many expectations about the difference in nAch levels between men and women, but Stewart and Chester (1982) concluded that there are no indications of genuine gender differences in nAch level.

Results

Descriptive statistics. For the male sample the correlation between the original and dichotomized scores was .94, and for the female sample it was .95, indicating that not very much information was lost. Table 2 contains the proportion of nAch answers for each card and Cronbach's alpha separately for the male and female samples. It can be seen that the pictures did not elicit many achievement-oriented responses. The low proportions of nAch responses is not startling because there are many needs active in a person and some cards are chosen to cue needs other than nAch (e.g., nAff and nPow). Nevertheless, there is enough intercard variation in the proportion of nAch responses that has to be taken into account by the models. From Table 2, it can also be concluded that the internal consistency of the TAT was indeed very low (in both samples), confirming the results of Entwisle (1972) and Fineman (1977). These very low internal consistencies were probably also a consequence of the low proportion of achievement imagery.

Table 2
Descriptive Statistics for nAch: Proportion of Achievement Imagery and Cronbach's Alpha

Card no.	Men	Women
1	.28	.41
2	.21	.02
3	.02	.03
4	.11	.30
5	.11	.28
6	.02	.01
α	.25	.18

Note. nAch = need for achievement.

Results of the IRT analysis. Although in an IRT analysis, estimation comes logically before model checking, when presenting the results, it is clearer to discuss the best fitting model first and only then the parameter estimates of the chosen model and their implications.

Table 3 contains the results for the male sample of the four different tests for four models: the BAM, the DAM, the SDAM with card-specific non-drop-out parameters, and the restricted SDAM with a common non-drop-out parameter. Let us start with two remarks about the fitted models. First, it was necessary to set the interaction parameters β_{23} and β_{34} of the DAM equal to each other to prevent estimation problems. Second, for the SDAM, all drop-out parameters were made card specific, and in a second step the model was made more parsimonious by restricting λ_1, λ_2, and λ_6 to 1 as their values approached 1 (this implies that for Cards 1, 2, and 6, the BAM applies).

From Table 3, it can be seen that none of the models were rejected by the data for the men according to the chi-square statistic.[6] Considering the AIC criterion, the SDAM with card-specific λ_js fit the data best with the fewest parameters (because of the lowest AIC value). Looking at how well the models predicted the observed Cronbach's alpha, one can see that only both SDAMs were successful. The BAM and DAM overestimated the internal consistency. For example, for the BAM, only

[6] The models even tended to overfit the data (*p* values were close to 1). This was possibly caused by the fact that many response patterns were not observed and did not contribute very much to the chi-square statistic. When the models were fitted to the nAff and nPow data, we did not encounter overfitting of the data.

Table 3
Fit Statistics for Four IRT Models for the Male nAch Sample

Model	Chi-square			AIC	Cronbach's alpha	Test–retest correlation[b]		
	χ^2	*df*	*p*		*p*[a]	5%	*Mdn*	95%
BAM	40.39	57	.953	2,808.2	.02	.31	.37	.43
DAM	35.16	53	.972	2,810.1	.01	.31	.37	.43
SDAM	29.80	54	.997	2,801.9	.60	.19	.28	.36
Restricted SDAM[c]	40.18	56	.945	2,806.6	.50	.15	.27	.37

Note. $N = 715$. IRT = item response theory; nAch = need for achievement; AIC = Akaike's information criterion; BAM = basic apperception model; DAM = dynamic apperception model; SDAM = stochastic drop-out apperception model.
[a] For Cronbach's alpha, the *p* value is equal to the number of times the Cronbach's alphas of the simulated data sets were smaller than the Cronbach's alpha of the observed data set. [b] For the simulated test–retest correlations, the lower 5% and upper 95% values are shown with the median. [c] Restricted SDAM refers to the SDAM with a common non-drop-out parameter λ for all items.

2% of the predicted values of Cronbach's alpha were smaller than the observed one; this rate was 60% for the best fitting SDAM. The predicted test–retest correlations ranged from .19 to .36, with a median of .28. The median value was very close to the median test–retest correlation (.32) that was found in the review study of Fineman (1977). However, note that our simulated test–retest correlations probably tended to underestimate the real test–retest correlations somewhat, because in reality, memory effects and other person-specific factors may spuriously magnify this correlation.

The model fitting results for the female sample are presented in Table 4. Concerning the estimated models, there are two differences from the male sample. First, two interaction parameters of the DAM now also have to be constrained ($\beta_{23} = \beta_{34}$). The fact that these are the same parameters as in the male sample is a coincidence, because these cards differ for men and women. Second, for the SDAM, the drop-out parameters were constrained as follows: $\lambda_1 = \lambda_2 = 1$, $\lambda_3 = \lambda_6$, and $\lambda_4 = \lambda_5$, on the basis of earlier runs of the program. These are different constraints than in the male sample because of the different card sets that were used. The conclusions concerning the best fitting model are largely the same as for the male sample. All chi-square tests indicate that the models were not rejected by the data, and the model with the lowest AIC (and therefore the best fitting model) was again the SDAM with card-specific λ_js. Again, the very low observed Cronbach's alpha (.18) was best predicted by both SDAMs. The test–retest correlations for the best fitting model were somewhat lower than for the male sample and also lower than what can be found in the literature. However, we have to keep in mind that these predicted test–retest correlations are underestimations.

Because the SDAM with card-specific λ_js is chosen as the best fitting model for both the male and female sample, we look at its parameter estimates in detail (see Table 5). For the male sample, it can be concluded that for some cards all responses have a diagnostic value for the nAch level of the person (i.e., Cards 1, 5, and 6 because $\lambda_1 = \lambda_5 = \lambda_6 = 1$). However, the diagnostic value of Card 3 is extremely low (only 2% useful responses, or approximately 14 responses). This implies that it is very difficult to make claims about the instigating force of Card 3. This can be derived from the large standard error of the estimated β_3, which expresses the large uncertainty about its exact value. If we ignore the card specificity of the non-drop-out probabilities by assuming a common λ for all cards, the probability of useful responses is estimated at 66%. Conversely, there is an estimated drop-out of 34%, a number remarkably similar to Murray's (1943) estimate of 30% unusable fantasy material. From the results, it can also be concluded that the instigating force is not related to the non-drop-out probability. Intuitively, one could think that cards with a weak instigating force always have high drop-out probabilities. The results for Card 6 for men show that this is not true, because the card has a weak instigating force ($\beta_6 = -4.46$) and 100% of the responses are valid. The instigating force only comes in after the fantasy set for achievement motivation has been drawn on. Card 4 (for men) also illustrates that a strong instigating force ($\beta_4 = 0.24$) and a high drop-out (λ_j is only 0.20) can occur together.

For the female sample, similar conclusions can be drawn. Under the restricted SDAM, the estimated number of diagnostic responses (λ) was 62%, which is once again close to Murray's (1943) guess that about 30% of the responses are not diagnostic. The percentage of diagnostic responses was somewhat lower for the women than for the men, but the difference was not significant. As for the men, there was no relationship between the instigating force of the cards and the drop-out probabilities.

As a further illustration of the difference between the non-drop-out probabilities and the instigating forces, we discuss the estimated results from two cards. First, Card 1 in the male form (two men working on a machine in a shop) elicited in all cases responses that were diagnostic for the level of nAch. As stressed before, this is not the same as saying that all generated stories contained an achievement element, because only 28% of the stories contained such achievement elements. But the remaining 72% of the stories with no achievement content were also informative about the instigating force of the card and about the nAch level of the person. Second, Card 5 in the female form (two women preparing food) appealed in 48% of the cases to the achievement fantasy set. Thus, only 48% of the stories contained useful information with respect to the instigating force of the card and the nAch level of the persons. However, of all stories, 28% of them contained achievement imagery; thus, this card has a large instigating value, and that is exactly what is shown in Table 5. The

Table 4
Fit Statistics for Four IRT Models for the Female nAch Sample

Model	Chi-square			AIC	Cronbach's alpha	Test–retest correlation[b]		
	χ^2	*df*	*p*		*p*[a]	5%	*Mdn*	95%
BAM	40.74	57	.949	3,960.0	.00	.32	.37	.42
DAM	29.92	53	.996	3,957.5	.00	.32	.37	.41
SDAM	24.13	55	1.00	3,943.2	.62	.15	.22	.28
Restricted SDAM[c]	22.23	56	1.00	3,943.8	.46	.13	.21	.29

Note. $N = 904$. IRT = item response theory; nAch = need for achievement; AIC = Akaike's information criterion; BAM = basic apperception model; DAM = dynamic apperception model; SDAM = stochastic drop-out apperception model.
[a] For Cronbach's alpha, the *p* value is equal to the number of times the Cronbach's alphas of the simulated data sets were smaller than the Cronbach's alpha of the observed data set. [b] For the simulated test–retest correlations, the lower 5% and upper 95% values are shown with the median. [c] Restricted SDAM refers to the SDAM with a common non-drop-out parameter λ for all items.

estimated value of β_5 is large (0.45) because 28%/48% = 58% of the diagnostic responses contain achievement imagery, which is a large number.

We also checked whether there were individual differences in non-drop-out probabilities. For this purpose, we allowed the parameter λ of the restricted SDAM to vary over persons by assuming that λ was distributed normally in the population. However, the estimated variance of this population distribution was very small in both the male and the female subsamples. Moreover, the AICs of this model were larger than those of the best fitting model, indicating that there is no reason to assume that λ is related to individual differences.

The DAM was the least appropriate for both the male and female sample, but, nevertheless, we checked the estimated interaction parameters (β_{12}, β_{23}, etc.) to see whether a consummatory effect of an achievement imagery response was revealed. As explained when the model was presented, the interaction parameters should be smaller than zero if there is a consummatory effect. For the men, only two interaction parameters were smaller than zero ($\beta_{23} = \beta_{34} = -0.50$, with a standard error of 0.30) but not significant. For the women also, two interaction parameters were smaller than zero ($\beta_{23} = \beta_{34} = -0.26$, with a standard error of 0.08), and this time the interaction parameter differed significantly from zero, $t(903) = -3.12$, $p < .01$. However, in general, we must conclude that there were no indications of a general consummatory effect of achievement imagery. Furthermore, neither for men nor for women were there indications of a tendency intensification, as none of the positive interactions were significantly larger than zero.

Next, the data of men and women were modeled together. This was possible because the two samples had one common card (Card 6). For the joint modeling, only the BAM and the restricted SDAM (a common λ) were estimated. The DAM was not estimated because it was the worst fitting model in the separate analyses and it would not perform better in a joint analysis. The SDAM with card-specific λ_js was not fitted because it would involve a lot of computing time because of the many parameters. Furthermore, we did not perform an extensive testing; only the AIC was checked, and it was found that the SDAM fit better than the BAM (AIC = 6768.7 for the BAM, and AIC = 6749.2 for the SDAM). We note that the parameters (instigating forces and non-drop-out parameter λ) for the cards were very close to those from the separate analyses. The overall proportion of diagnostic responses is estimated at 63%.

Before discussing the next step in the joint analysis of the results of men and women, we need to stress that gender differences may be revealed in three instances. First, there may be a difference between men and women in the empirical proportion of achievement imagery (see Table 2). However, it is not clear whether differences would stem from the proportion of diagnostic responses (λ) or from differences in the level of nAch (θ). Furthermore, the differences in empirical proportions in Table 2 are confounded with the different set of cards, and therefore it is impossible to derive anything about gender differences in nAch

Table 5
Parameter Estimates for the SDAM for Men and Women

Sample	β_1	β_2	β_3	β_4	β_5	β_6	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
Men												
Estimate	−1.15	−1.08	1.98	0.24	−2.44	−4.46	1.00	0.72	0.02	0.20	1.00	1.00
SE	0.10	0.73	5.17	1.14	0.13	0.29		0.31	0.02	0.09		
Women												
Estimate	−0.44	−4.25	0.73	0.59	0.45	−1.86	1.00	1.00	0.05	0.48	0.48	0.05
SE	0.08	0.23	1.93	0.50	0.47	0.94			0.03	0.07	0.07	0.03

Note. For values where λ was restricted to 1.00, no standard error is estimated. SDAM = stochastic drop-out apperception model.

level by relying solely on Table 2. Second, gender differences may show up in the case of different proportions of diagnostic responses (i.e., differences in λ). From the results of the separate analyses, it must be concluded that there were no reliable differences in non-drop-out probability concerning nAch (see above). Note, however, that in this analysis the difference was also confounded with the different set of cards. Third and finally, men and women may differ in their nAch level (θ_v) as derived from the diagnostic stories. Men may have, on average, a larger nAch than women, or vice versa. These gender differences in nAch level were not confounded with the particular set of cards presented to each gender group. It is important to distinguish between the second and third case of gender differences. Both are unrelated, because if the fantasy set is activated, the person can still give a lot of achievement-unrelated stories, displaying a low nAch level, or a lot of achievement imagery and a high nAch level.

Because the data of men and women were analyzed in one model, it is possible to test whether there were gender differences in nAch level. Using a nonparametric Wilcoxon rank sum test (Rice, 1995), we conclude that men had a higher median nAch level than did women ($z = 3.99$, $p < .0001$) under the SDAM. This result deviates from the conclusion of Stewart and Chester (1982) that there are no gender differences in nAch. Thus, men and women do not differ on their overall proportion of diagnostic responses (see nonsignificant difference between λs from separate analyses) for the nAch level, but they do differ with respect to their median nAch level.

Conclusion. From the results, we conclude that the SDAM with card-specific λ_js is the best fitting model for the achievement data both for men and women. The model is an acceptable approximation to the data as measured by the chi-square statistic. It is also both an appropriate and a parsimonious model, as indicated by the AIC. Moreover, it predicts the low Cronbach's alphas of the test better than do the BAM or DAM, and, finally, the expected test–retest correlations correspond to values that are found in the literature. We derived a different version of the SDAM for men and women, but that is because a different set of cards was used for men and women, with only one overlapping card. The single common card allowed a joint analysis of the male and female data, and it appeared that men had a significantly higher nAch level than did women.

The results of these nAch analyses were replicated with the nAff and nPow data (measured using the same set of cards and sample). In those cases, the SDAM was also by far the best model for the data. Again, the low internal consistencies were explained each time by the SDAM but not by the other two models (BAM and DAM). Concerning sex differences in the non-drop-out probability, women seemed to be more likely to have an activated affiliation imagery set than did men, whereas the reverse was true for power imagery, but none of these results in non-drop-out probability was significant. Concerning the level of the needs, there was no difference in nAff levels between men and women, but women had a higher nPow than men did. By and large, the conclusions regarding nAff and nPow data cross-validate the conclusions from the nAch data.

Discussion

In this article we have tried to model the process behind achievement imagery responses on a series of TAT cards. For this purpose, we formulated three apperception theories, each of which we linked to an IRT model, and we then tested the IRT models. In the introduction, three objectives of this research were set forth: a general, a more specific, and a methodological objective. We now discuss what we can conclude with respect to these three objectives.

Understanding the Response Process of Need-Related Fantasy

The general objective of this research was to formulate and test a theory about the response process on a series of TAT cards. The theory that seems most plausible in light of the data is the stochastic drop-out apperception theory. Thus, we conclude that persons respond to some cards by telling a story with a content that is not influenced by the achievement motivation. This may happen, for example, because irrelevant facts about the card or the person's life are given, or other motives than the achievement motive might determine the content of the story. These responses do not have a diagnostic value for the nAch. On the other hand, there are stories that are influenced by the nAch level of the person and the instigating force of the card. Such responses have a diagnostic value, meaning that in those cases the achievement fantasy set is activated and the story content reflects the strength of the motive and the cue value of the card. However, the achievement fantasy may still be activated if the content of story does not contain any achievement elements. This can occur when the joint influence of the motive and the card is not strong enough.

The results we obtained also show that it is unlikely that a consummatory mechanism, as proposed by the dynamics of action theory, is active in a series of TAT cards that measure nAch. This implies that achievement fantasy behavior has no satisfying effect on the underlying tendency that controls it, at least not over different cards. It is still possible that within one story there is consummatory influence, but our data do not allow us to investigate this. In the presentation of the DAM, it has also been mentioned that achievement imagery may have an intensifying effect on the underlying tendency to achieve, but this effect is also not present in the data. Such an intensifying effect would correspond to a fantasy escalation process that occurs if the tendency strength increases each time the achievement imagery is emitted. In that case, a fantasy achievement element activates similar fantasy achievement elements, as in an associative chain of activated fantasy elements. However, this does not seem to be present in the data.

Instead of showing a consummatory or intensifying trend, achievement imagery seems to be somewhat erratic and irregular, as implied by the drop-out model. Thus, fantasy behavior is far from lawful and predictable. Individual stories can be quite irrelevant and not inspired by the important dynamic sources within the person that we want to measure. It is unpredictable what will happen, although some cards are better triggers of the diagnostic imagery than others are (as can be seen in the differences in the non-drop-out probabilities).

The drop-out apperception theory borrows elements from several other theories. First of all, there is the well-accepted idea that cards differ in how relevant they are for eliciting imagery connected to some motive. The theory and corresponding IRT model that we proposed here explicitly acknowledge this fact by allowing cards to differ in instigating force. Second, we have built into our

theory and model the idea of a threshold that has to be exceeded, even if the fantasy is already activated. The idea of a threshold is less common in theories about projective techniques, but it has been used more often in perception and personality psychology. Third, the idea of an irrelevant story content was already formulated by Murray (1943, 1965).

The non-drop-out probability can be estimated at the level of the card, but unfortunately it is impossible to classify individual stories as diagnostic or nondiagnostic. If the non-drop-out percentage for some card is estimated at 60%, nothing more can be said about an individual story than that it is diagnostic with a probability of .60. From a statistical point of view, it is impossible to have model parameters connected to each individual story.

Reliability of the TAT

Concerning the specific objective of our study on the low internal consistency of the TAT, the validity of the stochastic drop-out apperception theory implies that some stories do not reveal anything about the person's motive strength. If some stories just drop out from the test in this way, this could be bad news, because the test would actually be shortened, thereby narrowing the basis for statistical and psychological inference. Of course, a given set of cards may be more reliable for one motive than for others. For instance, the six cards used in this study were clearly more suited for measuring nAch in men (we estimated about 34% drop-outs) than for measuring nAff (although this result is not shown, we estimated about 49% drop-outs). The estimated drop-out probabilities for each separate card are a good basis for the selection of an optimal set of cards.

The low internal consistency resulting from the SDAM also has consequences for the validity of the TAT. The only way to reduce the influence of the drop-outs is to increase the number of cards, but it has been explained before that this is not advisable from a practical point of view. More than six cards would lead to fatigue of the test takers, and the motive scores may decrease after too many cards (Atkinson, 1954). The hypothesized decrease could stem from an increasing drop-out after a certain number of cards (which could not be tested in our application given that only six cards have been presented).

Psychometric Models

Concerning our third objective, we have illustrated in this article that verbal theories about response processes can be translated into IRT models, and we have demonstrated how those IRT models can be tested. Moreover, we apply the IRT models in the domain of projective techniques, an area in which they are usually not applied. The assumptions of classical test theory may not be valid for projective techniques, so an alternative psychometric framework needs to be considered. From our research, it appears that classical test theory certainly does not match the complexity of the projective response process in the TAT and that a better match is provided with IRT models of a more complex kind.

We should add to this that even the more sophisticated stochastic drop-out model is without any doubt too simple, although it is empirically valid. Almost by definition, psychometric models do not fully cover a complicated psychological reality. But what these models do instead is make assumptions that constrain the range of possible observations. A model is either rejected, and hence its assumptions may be considered wrong, or it turns out to be tenable, and then the model does capture some important features of the data. But one should realize that this is not the same as saying that the model is the underlying truth.

Of course, one may wonder what is gained in estimating such complex models. We see three main advantages. First, the IRT models we tested each correspond with a psychological theory, and therefore testing these models is a way of testing and better understanding psychological theories. Second, one needs a valid psychometric theory for the interpretation of classical psychometric measures (e.g., internal consistency). Both the DAM and the SDAM contradict classical test theory, such that the common meaning of the classical psychometric measures changes and the classical formulas no longer apply. Third, IRT modeling allows for a great deal of flexibility to test many hypotheses even if the data do not appear to allow for this at first glance. For instance, in this article we have checked whether there are gender differences when men and women have only one common card administered. A second example is the test for individual differences in non-drop-out probabilities. This kind of differentiation in testing group differences is not possible within the classical test theoretical framework.

Although we advocate the use of IRT modeling to find a correct psychometric model for personality data, it must be stressed that there are inevitably also some disadvantages. First, most of the literature on these models is not easily accessible to mathematically nonskilled psychologists. Second, the programs necessary to apply these models are not largely available, although some improvement may be expected now that SAS has a flexible procedure for these kind of models (Rijmen et al., 2001; SAS Institute, 1999). Third, the models often require large sample sizes for a reliable estimation of the parameters. Especially in the field of projective techniques, large sample sizes are often not available. The existence of the data set we have used (Veroff et al., 1960) was very helpful in this respect. But despite these difficulties, we hope that further progress is made in the search for valid psychometric models for the more dynamic kind of data encountered in the domain of personality.

The models discussed in this article are applicable to types of data other than projective techniques. As a first example, suppose that a very long personality checklist is administered on a computer. The first part of the test could be analyzed with the BAM (or the Rasch model), whereas for the second part, the restricted SDAM with a common drop-out probability could be fitted. The SDAM allows for the possibility that some of the responses of the test takers are not generated by the BAM because of test fatigue during a long test. The drop-out can be explained as a consequence of attention loss, random pressing, or anticipatory responses (pressing too fast without having read the question). As a second example, consider the case when two questions in a personality questionnaire are very much alike. Then it is very likely that if a person responds in some way to the first question, he or she will respond in the same way to the second question. In such cases, the DAM with positive dependencies may be applied to analyze the responses.

References

Akaike, H. (1977). On entropy maximization principle. In P. R. Krishnaiah (Ed.), *Proceedings of the symposium on application of statistics* (pp. 27–47). Amsterdam: North-Holland.

Atkinson, J. W. (1954). Explorations using imaginative thought to assess the strength of human motives. In M. R. Jones (Ed.), *Nebraska Symposium on Motivation* (pp. 56–112). Lincoln: University of Nebraska Press.

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review, 64,* 359–372.

Atkinson, J. W. (Ed.). (1958). *Motives in fantasy, action, and society.* Princeton, NJ: Van Nostrand.

Atkinson, J. W. (1965). Thematic apperceptive measurement of motives within the context of a theory of motivation. In B. I. Murstein (Ed.), *Handbook of projective techniques* (pp. 433–455). New York: Basic Books.

Atkinson, J. W. (1981). Studying personality in the context of an advanced motivational psychology. *American Psychologist, 36,* 117–128.

Atkinson, J. W. (1982). Motivational determinants of thematic apperception. In A. Stewart (Ed.), *Motivation and society* (pp. 3–40). San Francisco: Jossey-Bass.

Atkinson, J. W., & Birch, D. (1970). *The dynamics of action.* New York: Wiley.

Atkinson, J. W., Bongort, K., & Price, L. H. (1977). Explorations using computer simulation to comprehend thematic apperceptive measurement of motivation. *Motivation and Emotion, 1,* 1–27.

Blankenship, V., & Zoota, A. L. (1998). Comparing power imagery in TATs written by hand or on the computer. *Behavior Research Methods, Instruments, & Computers, 30,* 441–448.

Cramer, P. (1996). *Storytelling, narrative, and the Thematic Apperception Test.* New York: Guilford Press.

Cramer, P. (1999). Future directions for the Thematic Apperception Test. *Journal of Personality Assessment, 72,* 74–92.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Hillsdale, NJ: Erlbaum.

Entwisle, D. E. (1972). To dispel fantasies about fantasy-based measures of achievement motivation. *Psychological Bulletin, 77,* 377–391.

Fineman, S. (1977). The achievement motive construct and its measurement: Where are we now? *British Journal of Psychology, 68,* 1–22.

Fischer, G. H., & Spada, H. (1973). *Die psychometrischen Grundlagen des Rorschachtests und der Holtzman Inkblot Technique* [The psychometric foundation of the Rorschach Test and the Holtzman Inkblot Technique]. Berne, Germany: Huber.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78,* 350–365.

Gelman, A., Carlin, B. P., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Gurin, G., Veroff, J., & Feld, S. (1960). *Americans view their mental health: A nationwide interview survey.* New York: Basic Books.

Gurin, G., Veroff, J., & Feld, S. (1975). *Americans view their mental health, 1957* [Computer file]. Conducted by University of Michigan, Institute for Social Research, Social Science Archive. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Heckhausen, H. (1991). *Motivation and action.* Berlin, Germany: Springer-Verlag.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local item dependence among test items. *Psychological Methods, 2,* 261–277.

Jensen, A. R. (1959). The reliability of projective techniques: Review of the literature. *Acta Psychologica, 16,* 108–136.

Klinger, E. (1966). Fantasy need achievement as a motivational construct. *Psychological Bulletin, 66,* 291–308.

Kraiger, K., Hakel, M. D., & Cornelius, E. T., III. (1984). Exploring fantasies of TAT reliability. *Journal of Personality Assessment, 48,* 365–370.

Kuhl, J. (1978). Situations-, reaktions- und personbezogene Konsistenz des Leistungsmotivs bei der Messung mittels des Heckhausen-TAT [The situation-, response-, and person-related consistency of the achievement motive as measured by the Heckhausen-TAT]. *Archiv für Psychologie, 130,* 37–52.

Lewin, K. (1935). *A dynamic theory of personality.* New York: McGraw-Hill.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1,* 27–66.

Lundy, A. (1985). The reliability of the Thematic Apperception Test. *Journal of Personality Assessment, 49,* 141–145.

McClelland, D. C. (1980). Motive dispositions: The merits of operant and respondent measures. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 1, pp. 10–41). Beverly Hills, CA: Sage.

McClelland, D. C. (1985). *Human motivation.* Glenview, IL: Scott, Foresman.

McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). *The achievement motive.* New York: Appleton-Century-Crofts.

McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1958). A scoring manual for the achievement motive. In J. W. Atkinson (Ed.), *Motives in fantasy, action, and society* (pp. 179–204). New York: Van Nostrand.

McClelland, D. C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review, 96,* 690–702.

Mitchell, J. V., Jr. (1961). An analysis of the factorial dimensions of the achievement motivation construct. *Journal of Educational Psychology, 52,* 179–187.

Morgan, C. D., & Murray, H. A. (1935). A method for investigating fantasies. *The Archives of Neurology and Psychiatry, 34,* 389–406.

Murray, H. A. (1938). *Explorations in personality.* New York: Oxford University Press.

Murray, H. A. (1943). *Thematic Apperception Test manual.* Cambridge, MA: Harvard University Press.

Murray, H. A. (1965). Uses of the thematic apperception test. In B. I. Murstein (Ed.), *Handbook of projective techniques* (pp. 425–432). New York: Basic Books.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter logistic model to personality data. *Applied Psychological Measurement, 14,* 45–58.

Reuman, D. A. (1982). Ipsative behavioral variability and the quality of thematic apperceptive measurement of the achievement motive. *Journal of Personality and Social Psychology, 43,* 1098–1110.

Rice, J. A. (1995). *Mathematical statistics and data analysis.* Belmont, CA: Duxbury Press.

Rijmen, F., Tuerlinckx, F., & De Boeck, P. (2001). *Estimating psychometric models with SAS.* Unpublished manuscript, University of Leuven, Leuven, Belgium.

SAS Institute. (1999). *SAS/STAT (experimental) user's guide* (Version 8e). Cary, NC: Author.

Smith, C. P. (Ed.). (1992a). *Motivation and personality: Handbook of thematic content analysis.* Cambridge, MA: Cambridge University Press.

Smith, C. P. (1992b). Reliability issues. In C. P. Smith (Ed.), *Motivation and personality: Handbook of thematic content analysis* (pp. 126–139). Cambridge, MA: Cambridge University Press.

Spangler, W. D. (1992). Validity of questionnaire and TAT measures of need for achievement: Two meta-analyses. *Psychological Bulletin, 112,* 140–154.

Stewart, A. J., & Chester, N. L. (1982). Sex differences in human social motives: Achievement, affiliation, and power. In A. J. Stewart (Ed.), *Motivation and society* (pp. 172–218). San Francisco: Jossey-Bass.

Tyler, F. B., Tyler, B. B., & Rafferty, J. E. (1962). A threshold conception of need value. *Psychological Monographs: General and Applied, 76,* 1–28.

Vansteelandt, K. (1999). A formal model for the competency-demand hypothesis. *European Journal of Personality, 13,* 429–442.

Veroff, J., Atkinson, J. W., Feld, S. C., & Gurin, G. (1960). The use of thematic apperception to assess motivation in a nationwide interview study. *Psychological Monographs: General and Applied, 74,* 1–32.

Veroff, J., Feld, S., & Crockett, H. (1966). Explorations into the effects of picture cues on thematic apperceptive expression of achievement motivation. *Journal of Personality and Social Psychology, 3,* 171–181.

Vislie, L. (1972). *Stimulus research in projective techniques.* Oslo, Norway: Scandinavian University Books.

Wickens, T. D. (1982). *Models for behavior: Stochastic processes in psychology.* Hillsdale, NJ: Erlbaum.

Winter, D. G., John, O. P., Stewart, A. J., Klohnen, E. C., & Duncan, L. E. (1998). Traits and motives: Toward an integration of two traditions in personality research. *Psychological Review, 105,* 230–250.

Winter, D. G., & Stewart, A. J. (1977). Power motive reliability as a function of retest instructions. *Journal of Consulting and Clinical Psychology, 45,* 436–440.

Appendix A

Probability Formula for the Dynamic Apperception Model (DAM)

To introduce the DAM, define the observation x_{vj} as 1 if person v showed achievement imagery on card j and as 0 otherwise. The total probability on the response pattern (x_{v1}, x_{v2}) equals

$$\Pr(X_{v1} = x_{v1}, X_{v2} = x_{v2}) = \frac{\exp(x_{v1}(\theta_v + \beta_1) + x_{v2}(\theta_v + \beta_2 - \beta_{12}) + 2x_{v1}x_{v2}\beta_{12})}{1 + \exp(\theta_v + \beta_1) + \exp(\theta_v + \beta_2 - \beta_{12}) + \exp(2\theta_v + \beta_1 + \beta_2 + \beta_{12})}. \quad \text{(A1)}$$

From Equation A1, the probabilities on the different possible response patterns (0, 0), (0, 1), (1, 0), and (1, 1) can be defined. The model in Equation A1 can be generalized easily to the case of six cards. In the psychometric literature, the model in Equation A1 is called the *constant dependency model* (Hoskens & De Boeck, 1997). The model as presented in Hoskens and De Boeck (1997) uses a different parametrization but is fully equivalent with Equation A1.

Appendix B

Details on Testing the IRT Models

Four different model tests were performed, each with a specific purpose. First, the construction of the chi-square test is explained. With six binary items, there were $2^6 = 64$ different possible response patterns (all possible patterns of zeros and ones). For each response pattern, there was an associated observed frequency and a predicted frequency that could be computed from the estimated model parameters. Using the observed and predicted frequencies, we calculated the familiar Pearson chi-square statistic. The computed value of the chi-square statistic was compared with a chi-square distribution with degrees of freedom equal to $2^6 - 1 - npar$, where *npar* is the number of parameters pertaining to the items of the tested model.

Second, the AIC (Akaike, 1977) is defined as follows:

$$\text{AIC} = -2\log(L) + 2 \times npar, \quad \text{(B1)}$$

where $-2\log(L)$ is -2 times the log likelihood of the model (when the model is fitted, this quantity is minimized) and *npar* is the number of parameters pertaining to the items of the model. As can be seen from Equation B1, the AIC penalizes a model for having too many parameters. The model with the lowest AIC is to be preferred.

The third test used Cronbach's alpha, but because we had binary items, this reduced to the well-known KR-20 index (Cronbach, 1951). We have chosen to evaluate whether the models could predict the internal consistency of the data set. For this purpose, we used a Bayesian testing procedure (Gelman, Carlin, Stern, & Rubin, 1995). This required that new data be simulated under the model, which was a two-step process.

First, we generated a set of parameter values using the estimated parameters and their standard errors. The generated parameter values were drawn from a normal distribution with a mean equal to the estimate from the SAS program and with a standard deviation equal to the standard error, also derived from the program output. By using the normal distribution for generating new parameter values, we approximated the Bayesian posterior distribution (see Gelman et al., 1995).

Second, using the generated set of parameter values, we simulated a new data set from the model and computed Cronbach's alpha for this new simulated data set. These two subsequent steps were repeated 2,000 times, so that we ended up with 2,000 simulated values of Cronbach's alpha. Next, Cronbach's alpha of the original data set was compared with the set of 2,000 Cronbach's alphas from the simulated data sets, and we counted how many times the simulated ones were smaller than the observed one. If Cronbach's alpha for the observed data was always smaller (or larger) than Cronbach's alpha for data sets under the model, then the model was not capable of explaining a low Cronbach's alpha. The *p* value was defined as the proportion of Cronbach's alphas from the replicated data that were smaller than the observed Cronbach's alpha. Hence, if the *p* value was small, this indicated that the model was not capable of explaining the low observed internal consistency.

The fourth test was related to the third one, except that we now focused on the test–retest correlation. We followed the same procedure as described for Cronbach's alpha, except that this time we simulated for each set of parameter values two new data sets (representing the two different testing occasions). Next, the correlations between the total raw nAch scores for the two data sets were computed, and this was seen as a simulation of the test–retest correlation. Because we had no test–retest correlation for the observed data at our disposal we compared the simulated values with values found in the literature.

Received April 21, 2000
Revision received August 24, 2001
Accepted August 30, 2001 ■