Running head: CONTENT CODING METHODS

Content Coding Methods in Implicit Motive Assessment: Standards of Measurement and Best Practices for the Picture Story Exercise

Joyce S. Pang

Nanyang Technological University, Singapore

In *O.C. Schultheiss & J.C. Brunstein (Eds.), Implicit Motives. New York, NY: Oxford University Press.*

Author note:

All picture stimuli described in this chapter are available on request from the author.

Correspondence concerning this chapter should be addressed to Joyce S. Pang, Division of Psychology, Nanyang Technological University, Singapore 639798, or sent by email to (joycepang@ntu.edu.sg).

Content Coding Methods in Implicit Motive Assessment: Standards of Measurement and Best Practices for the Picture Story Exercise

Thematic content analysis refers to a set of assessment methods where written or oral responses to open-ended questions and naturally-occurring narrative material are analyzed in order to reveal complex personality processes such as motivation (e.g., McClelland, Atkinson, Clark, & Lowell, 1953), cognitive complexity (e.g., Suedfeld, Tetlock, & Streufert, 1992), and stereotypes (e.g., Taylor, Lee, & Stern, 1995). The types of content analysis methods are numerous and range from archival analysis (e.g., Winter, 1992) to the thematic analysis of survey responses (Atkinson, 1958).

Although the content-coding method of assessing implicit motives is labor-intensive and time-consuming it also possesses the advantage of richness of information. Specifically, the researcher can rescore the same set of protocols for a number of diverse constructs and thus investigate complex personality forces happening at the same time, e.g., different motives acting in the same situation. Additionally, researchers are able to measure motives at-a-distance, thus gaining access to a pool of otherwise unavailable data from subjects, who, for instance, are deceased or live in remote locations. Finally, since the methodology is open-ended and non-reactive; researchers are less worried about response sets that commonly occur in self-report methods (c.f., Nisbett &Wilson, 1977).

The goal of this chapter is to inform non-expert researchers who are interested in assessing implicit motives of the proper and possible uses of content-coding methods. In this chapter, I will reiterate and elaborate on important points from similarly instructional sources on motivational content analysis, such as Schultheiss and Pang (2007) and Smith (1992), as well as provide additional recommendations from previously unpublished data. Specifically, I focus on the Picture Story Exercise (PSE; Koestner & McClelland, 1992), because it is the most widely used method for assessing implicit motives, especially the Big Three motives of *n* Power (the need to have impact on other people), *n* Affiliation (the need to establish and maintain positive relations with others), and *n* Achievement (the need to do things better).

After a discussion of the background of the PSE and two standards of measurement—internal consistency and validity—that have been widely discussed in the evaluation of the PSE as an instrument, the first part of the chapter will introduce various topics such as picture cue selection, administration, coding systems and coder training, data processing, and other pragmatic issues that are relevant to using the PSE in research. In turn, these topics are introduced systematically according to the stages of measurement of pretest, test, posttest, and retest. Readers who are interested in measuring implicit motives with the PSE should be able to embark on such research after reading and carrying out the recommendations contained in this section. The latter part of this chapter discusses the development and choice of picture sets for multi- versus single-motive measurement. This section provides previously-unpublished cue strength statistics for a single-motive picture set that "pulls" mainly for *n* Achievement.

Background of the PSE

The PSE is a research version of the Thematic Apperception Test (TAT; Morgan & Murray, 1935) and it is the most widely-used tool for measuring implicit motives. It is based on an underlying assumption that needs can be inferred from imaginative material generated in response to sufficiently arousing pictorial, verbal, or textual cues. Accordingly, most non-clinical research on implicit motives use the PSE, which requires participants to write imaginative stories in response to ambiguous picture stimuli that depict people in various "everyday" situations; the imaginative story protocols are then analyzed for motive imagery using experimentally-derived coding systems. Researchers have developed and used PSE-based coding systems for a variety of motives on a variety of materials, such as political speeches, diaries, and literary sources (Cramer, 1996; Smith, 1992; Winter, 1994).

Traditional notions of internal consistency are based on the idea that items (or, in the case of the PSE, picture cues) on a test measure related aspects of the same construct and the larger interitem correlations are the more reliable a test is (Nunnally, 1978). Hence, each item on a trait inventory needs to be sufficiently correlated with other items on the inventory. Internal consistency estimates are suitable tools for evaluating self-report measures such as trait inventories and measures of self-attributed motives (such as Jackson's 1984 Personality Research Form) that tap into people's need to maintain a consistent self-concept. Questions about the typically-lower internal consistency of the PSE (around .20 to .50) have been raised in early (Entwisle, 1972) as well as recent critiques (Lilienfeld, Wood, & Garb, 2000).

One reason for the typically-low internal consistency coefficients for PSE motive measures is given by a theory about the dynamic and continuous nature of the motivational sequence. According to the Dynamics of Action theory (DOA; Atkinson & Birch, 1970), a motivational drive sets off a behavioral sequence that culminates in the fulfillment of the underlying need, and the fulfillment of the need acts to decrease the strength of the motivational drive. In other words, the sheer act of expressing a motive reduces its strength or intensity. This decrease in the strength of a motive tendency after that tendency has been expressed is referred to as the consummatory value of the activities/events (Atkinson & Birch, 1970).

Since each PSE cue is assumed to possess consummatory value—a subject's motivation is expressed in the form of motive-relevant imagery in each PSE story—we would expect the motive imagery of pictures that immediately follow motive expression to be decreased; having satisfied her motivational need by responding to one PSE cue, the subject reacts less strongly to the next picture and writes less motive-relevant imagery on the second picture. However, over time, the motivational drive eventually returns to its initial state, thus raising the consummatory value of pictures less immediate in time to the initial motive expression. This waxing and waning nature of the motivational sequence is reflected in the resulting irregularity in the amount of motive imagery in successive PSE stories. Hence, internal consistency estimates may not be suitable tools for assessing PSE motive measures that tap into a dynamic motivational process. Rather, it is the combined motives scores over the entire span of the PSE that should more accurately represent the expression of a person's motive.

However, Tuerlinckx, De Boeck, and Lens (2002) recently tested Atkinson and Birch's theory using models from item-response theory and they found no evidence for a consummatory mechanism. Since Atkinson and his colleagues only tested the DOA theory using computer simulations (Atkinson, Bongort, & Price, 1977), there is a need for future studies to empirically test Atkinson and Birch's (1970) widely accepted conventional wisdom about the drive-reducing effect of PSE-measured motives.

Schultheiss and Pang (2007) also pointed out another reason that PSE measures have lower internal consistency. They refer to the issue of *motive extensity*, which is defined as the range of motivationally relevant situations and contexts that are depicted in a given battery of picture cues. There are many situations which are could be typically associated with need satisfaction and these situations are represented in different picture cues settings. For instance, although a courtroom and a boxing ring may be equally obvious settings for *n* Power displays, they arouse themes associated with different aspects of power manifestation (e.g., covert versus overt, verbal versus physical).

One important characteristic of motives is the variability of behaviors associated with a motive. Specifically, motives have no fixed repertoire of instrumental behaviors. Instead, motivated behaviors are very variable, depending on a variety of factors, such as what is required by the immediate situation to achieve the desired incentive, whether there are obstacles to reaching the goal, the intelligence and skills of the person, and the presence of other conflicting motives and desires. For instance, if a person is motivated by *n* Affiliation, she can achieve her goal of establishing friendly relations by doing any of a number of things. Perhaps the person might go to a party to increase the opportunities for being around people. Alternatively, she might stay at home to write letters to her old friends (McAdams & Constantian, 1983). Knowing that someone is affiliation-motivated does not increase the likelihood of knowing whether or not the person will express their *n* Affiliation by going out to a party or by staying home to write letters.

This *behavioral variability* in motivated action manifests itself in the PSE partly in the *motive extensity* of PSE picture cues. PSE picture sets frequently depict people in a wide variety of situations. The characters in the picture cues can be seen as acting out various modes of motive consumption and motive expression (e.g., partying versus writing letters). Accordingly, a picture set designed to study *n* Achievement may depict individuals in competitive sporting situations as well as in isolated work environments.

Indeed, McClelland and his colleagues have reasoned that motive extensity in the PSE is an important factor in motive measurement as motivational strength is not only reflected in the intensity of individual's responses to motivational incentives and cues, but also in the variability and range—the extensity—of cues that elicit a motivational response. According to McClelland (1987), people learn to associate naturally rewarding feelings with situations, events, or behavior that elicit such feelings. Consequently, the more motivated a person is, the greater the range of situations in which she will learn to associate with attaining the naturally pleasurable feeling. The issue of motive extensity has two implications for the internal consistency of the PSE.

First, a PSE measure that contains highly dissimilar pictures would have lower internal consistency but broader validity. Thus, lower internal consistency on the PSE may be partly due to the dissimilarity of different picture cues. A highly motivated person who has developed her motive only in a small range of situations will inject motive imagery unevenly across a set of highly dissimilar pictures. However another equally highly motivated person who has a more extensive motive will respond to a wider variety of picture cues and will have higher inter-item correlations. As Schultheiss and Pang (2007) suggest, there is a problem of bandwidth-fidelity, which refers to the tradeoff between highly specified, multidimensional, and highly variable assessment of personality constructs on one hand and more simplistic but reliable prediction on the other.

Second, the validity of the PSE can be increased if a researcher carefully chooses a PSE picture set that encompasses a sufficient variety of situations in which the motive can potentially be expressed. Encouragingly, Schultheiss, Liening, and Schad (2008) have recently studied test-retest reliability for handwritten and typed PSEs and found that motive scores show substantial ipsative stability. In other words, participants responded similarly to the same picture cue across testing occasions. The authors' findings corroborate the idea that although the PSE may not necessarily have high inter-picture correlations, it still provides reliable and valid readings of motive levels especially if pictures with similarly motivationally-relevant thematic contents are being correlated with each other. This point of sampling a range of motivationally-relevant settings to increase PSE extensity will be elaborated on later during the discussion about picture cue selection.

Guide for Using the PSE

Our goal in implementing the PSE methodology is to maximize the validity and reliability of the measure, as well as to minimize the sources of error in assessment. Essentially, we want to make sure that we are measuring the motive of interest, rather than any of its conceptual relatives or artifacts from various demand characteristics. Thus, the researcher should be extra careful to select appropriate picture cues and scoring systems, and to adopt standardized administration instructions and conditions in order to minimize any confounds.

Consequently, the following section provides a step-by-step guide to administering and scoring the PSE. Many of the main points have been covered in detail by Schultheiss and Pang (2007) and Smith (1992; 2000) so I will highlight the important issues here. While Schultheiss and Pang (2007) organized their guide to the PSE around key issues and major questions, such as “Which motives to assess?” and “Which pictures to use?,” the step-by-step guide contained in this chapter is organized systematically by the major stages of measurement of pretest, test, posttest and scoring, and retest. Furthermore, while Schultheiss and Pang (2007) offers a greater depth of discussion for certain issues (e.g., on various types of coding systems and test-retest reliability), the range of topics covered here is wider and incorporates all the major points of Schultheiss and Pang’s as well as some important points (e.g., regarding administration setting and pretesting of new pictures) that were not dealt with in that chapter.

In Table 1, the major standards (internal consistency, validity, and test-retest reliability) and stages of PSE measurement are cross-referenced with citations of relevant sources; the interested reader is referred to these external sources should more detail be desired than is covered in this chapter.

Pretest

During the pretest stage, the researcher should first be clear about the problem, hypotheses, and goals of the research. This conceptual step is important because it guides the selection of materials such as picture cues and scoring systems. It is also during the pretest stage that the researcher can collect pilot data for the development of new picture sets for motives that are not commonly studied or for which picture sets are not publicly available.

The researcher should also decide if the PSE is the most expedient method of data collection. Smith (2000) suggests that archival or naturally occurring materials such as personal documents, broadcasts, and email exchanges may sometimes be more informative than motive scores from a laboratory-administered PSE; the researcher who is trying to devise cues to measure a heretofore under-studied construct or to use with an unusual population may find that archival and naturally occurring materials provide more meaningful data because these offer greater ecological validity.

Once the researcher has decided to use the PSE, she needs to make a number of other additional decisions during the pretest stage, the main ones being the number, selection, pretesting, and order of presentation of picture cues.

Number of Pictures. Researchers should use picture sets of eight pictures or fewer because longer tests tend to decrease in validity as a result of the effects of fatigue (Reitman & Atkinson, 1958). However, Schultheiss and Pang (2007) have shown that variance of motive scores is inversely related to size of the test battery, so that motive score distributions of picture batteries with fewer than four pictures are considerably skewed to the left. Motive scores start to resemble a normal distribution once a five-picture battery is used. Hence, it is recommended that researchers use between 5-8 pictures in a PSE.

Selecting picture cues. The selection of PSE cues is based on the principle that evocative images trigger motive-relevant emotions and cognitions. Accordingly, cue selection is a very important topic for valid PSE measurement because picture sets differ in their ability to elicit different motives, their suitability for measuring a single motive versus several motives at once, and in the range of motive-relevant settings that they depict. Amongst the cue characteristics that a researcher should consider while selecting pictures are, cue strength, cue ambiguity, universality, relevance, and extensity.

Cue strength—sometimes referred to as stimulus pull—is the average amount of imagery for a particular motive that is elicited by a picture cue. While some pictures have good cue strength for only one motive, others are capable of eliciting imagery for more than one motive, and still other pictures may elicit very little imagery at all. In order to maximize the effectiveness of the PSE and obtain scores that contain adequate variance, researchers should select pictures that have sufficiently high cue strength for the motive(s) of interest.

Recently, there has been renewed research interest in discovering the cue strength characteristics of commonly used PSE picture cues (Schultheiss and Brunstein, 2001; Langan-Fox & Grant, 2006; Pang and Schultheiss, 2005) for measuring the Big Three motives of *n* Power, *n* Affiliation, and *n* Achievement at once. Table 2 presents commonly-studied pictures that have been collected from previously published research (Smith, 1992; McClelland & Steele, 1972; McClelland, 1975) and their cue strength statistics, as studied by various researchers (e.g., Schultheiss & Brunstein, 2001; Pang & Schultheiss, 2005; Langan-Fox & Grant, 2006).

Generally, a picture cue is designated as having high pull for a motive when it elicits at least one codeable image from at least 50% of the participant pool (c.f., Schultheiss & Brunstein, 2001). Researchers interested in measuring all three motives at once are recommended to use the data represented in Table 2 by selecting pictures with moderately high to high pull for the motive(s) that they are interested in. For instance, as shown in Figure 1, *boxer*, *ship captain*, *trapeze artists*, *nightclub*, and *women in laboratory* all qualify as high pull pictures for *n* Power. A researcher interested in developing a picture set solely for measuring *n* Power should include these five pictures in her picture set. However, if the researcher was interested in measuring only *n* Affiliation, she would use *nightclub* and *bridge* in addition to other pictures that show high to moderately high pull for *n* Affiliation.

Cue ambiguity refers to the ability of a picture to evoke multiple motives (see Smith, Feld, & Franz, 1992, and Murstein, 1972 for a broader discussion of cue ambiguity). Haber & Alpert (1958) suggested that researchers select pictures with low ambiguity, that is, pictures that clearly evoke a particular motive. This is because pictures that are more ambiguous also tend to have lower cue strength compared to less ambiguous pictures. Additionally, according to Haber and Alpert (1958), pictures with low ambiguity and high cue strength tend to have greater test-retest reliability (r = .59) than pictures with high ambiguity and low cue strength (r = .36).

However, there are also good reasons to ensure that picture cues should have sufficient ambiguity, rather than be overpoweringly representative of a particular motive. First, cue ambiguity ensures an adequate range of scores, thus increasing the variance of the measure and its ability to discriminate between highly-motivated and less-motivated participants. If a picture has too low cue ambiguity, then all participants would be prompted by the overwhelming cue strength into inserting certain motive imagery in their stories, resulting in a limited degree of variance in motive scores.

The second reason cue ambiguity is desirable is related to the indirect manner in which implicit motivational intents are expressed in PSE and other free-response assessment formats. Two studies by Clark and Sensibar (1955) are instructive: In the first study, which was conducted in a classroom setting, they showed slides of nude pinup girls to one group of male undergraduates and slides of landscaping and architecture to another group. The stories each of these groups wrote following exposure to the slides were then coded for manifest sexual imagery, which included explicit references to sexual acts such as fondling, kissing, and sexual intercourse. Surprisingly, the students exposed to the nude pinup pictures had significantly lower amounts of sexual imagery in their protocols than the students in the landscape condition.

However, when the same study was conducted at a fraternity party, Clark and Sensibar (1955) found different results. The male undergraduates who were shown pictures of nude pinups wrote significantly higher amounts of manifest sexual imagery than students at the same party who saw the landscape slides. Additionally, whether they belonged to the nude pinup or the landscape condition, students at the party wrote significantly more explicit sexual imagery than the students in the classroom administration. Finally, the authors rescored all the protocols from the party and classroom administrations and found that the protocols of students in the nude pinup conditions who had both extremely high *and* extremely low manifest sexual imagery scores, also included significantly higher amounts of sexual symbolism (e.g., round objects = breast; long object = penis) than students in either of the landscape-control conditions.

McClelland (1987) has interpreted Clark and Sensibar's study as indicating that some intents—such as sexual arousal— that are typically socially-inappropriate cause more anxiety in some settings than in others. In those settings where sexual arousal do not cause anxiety (fraternity party), participants' sexual drive are expressed in both explicit sexual imagery and indirect sexual symbolism. However, in settings where sexual arousal causes anxiety (classroom), participants' sexual drives are expressed more covertly in metaphoric or symbolic imagery.

In addition to McClelland's interpretation, I would like to add another implication of Clark and Sensibar's (1955) findings. The more explicit a picture cue is (e.g. nude pinup girls) in referring to specific motivational intents, the more likely it is to lead to the defense mechanisms seen in Clark and Sensibar's (1955) study. Thus, it is important to vary the content of picture cues so as not to arouse suspicion. This is particularly relevant when one is assessing motives that are expected to arouse anxiety in populations. Hence, a researcher needs to select picture cues that elicit a sufficient amount of ambiguity—so as to provide enough variance in motive scores within a population—while also having a high to moderately high cue strength—so as to effectively evoke the motive(s) of interest.

Another important criterion for selecting picture cues is universality, which is the tendency of pictures to have similar motivational significance to almost all members of a population (Smith, Feld, & Franz, 1992). Universality contributes to the face validity of a picture cue—the picture should clearly evoke certain motivational themes in most members of the population, to the extent that the prototypical response will produce average amounts of motive imagery. By promoting a baseline level of motive imagery, universality allows researchers to more easily observe individual differences in motive tendencies from the variation of motive scores (Atkinson, 1958).

Relevance refers to the ability of pictures to reflect current concerns and experiences of participants. Since the principle behind PSE measurement is that participants will project their predominant wishes, desires, and motives onto the characters in the pictures, the contexts depicted in the picture cues should be representative of current experiences. For this reason, it may be advisable to update pictures that have become out of date with respect to features such as clothes and hair styles. Additionally, it may not be advisable to use pictures showing characters that are significantly younger than participants as these pictures may either elicit past recollections of scripted or actual events rather than participants' current motivations (c.f., McClelland et al., 1953).

Generally, pictures used should be representative of common situations in which motives of interest are aroused. For instance, to assess achievement motivation, one might include pictures of work, school, and other performance settings. Furthermore, as discussed earlier, a person's motive can be expressed in a variety of situations. To date, the *extensity* of a picture set has not been studied systematically, however, there have been recommendations by previous motivational researchers (e.g., Smith, 1992; Schultheiss & Pang, 2007) to select picture sets which represent a broad spectrum of motivationally relevant contexts and situations which have a broader range of validity.

Pretesting New Pictures. Instead of using previously established picture sets (e.g., those mentioned here or contained in Smith, 1992), some researchers may wish to develop and pretest their own picture sets for measuring particular motives. Clearly, this enterprise is labor intensive and time consuming, however, researchers may choose to do so either because they would like to fine-tune pictures to suit particular situations or contexts, or because they are interested in studying some under-researched motives for which cue strength statistics have not previously been published.

During pretest, researchers should include amongst new picture cues, commonly used pictures for which cue strength have previously been established, in order to provide a standard of comparison for convergent and divergent validity, and to provide a control in ensuring that the findings on cue strength of the new pictures are not due to some spurious artefacts of the experimental condition or some other extraneous factor.

Pretest pictures may come from a variety of media outlets, such as print and television advertisements, newspapers and magazine, and even screenshots from films that are of sufficient resolution quality to be presented clearly. There are a number of pointers that the researcher can follow in identifying suitable pretest pictures:

1. Pictures included in the pretest battery should depict one or more persons engaging in or preparing to engage in instrumental behavior related to the motive. The researcher should also apply the criteria of cue strength, ambiguity, universality, relevance, and extensity during the selection of cues for pretesting.
2. Conduct thematic searches for usable pictures by considering prevailing theory about the motive or construct in question. For instance, when looking for cues for the *n* Achievement picture set described in the latter part of this chapter, I typed in

search terms such as “workplace”, “competition”, “achieve”, and “excellence” into an internet search engine. These search terms were arrived at after careful perusal of the available literature on achievement motivation and after prolonged reflection of the contexts and situations through which the achievement motivated process, as described in *The Achievement Motive* (McClelland, Atkinson, Clark, and Lowell, 1953), is expressed.

3. To increase motive extensity, broaden the pretest picture set to include pictures that depict as many motive-relevant behaviors and situations as possible. For instance, pictures for which cue strength statistics have previously been published and which appear to arouse related motives can be included in pretest battery. Thus, for the *n* Achievement picture set, I included in the pretest battery pictures that were used by Heckhausen (1963) to develop his scoring system for hope of success and fear of failure motivation. Additionally, include pictures that depict motive-relevant contexts which may not have been represented in commonly-available picture cues. For instance, some pictures in the *n* Achievement pretest battery also depicted individuals in competitive group sporting events settings, such as skating and soccer.

4. Favor contemporary sources of pictures over dated ones and contexts that are familiar but not so well known as to evoke scripted responses. Additionally, pictures that showcase famous events and/or persons should not be used, since there is the danger that participants will inject historical, biographical, or general knowledge about these people and events into their stories rather than writing an imaginative account. For instance, I once pretested a picture with a distinguished-looking lady in Victorian attire holding a test-tube. This picture cue was quickly excluded from future testing because nearly every story written in response to it made some reference to Marie Curie and radiation poisoning.

5. In order to increase the generalizability of the pictures across different participant samples (e.g., across different genders, age groups, races, social strata, etc.), select pictures that are either sufficiently light on detail as to be relatively ambiguous, or which do not depict recognizably extreme samples (e.g., characters in the pictures are neither very young or the very old).

The pretest administration procedure and conditions should be as similar to the main testing conditions as possible. In other words, researchers should try to exert minimum social, instructional, and time pressure on participants. Similarly, although there may eventually be a sizeable number of pictures included in the pretest battery, the number of picture cues presented to each participant in pretest conditions should be limited to the number that would be presented to each participant in actual test conditions (i.e., between 5-8). Although this means that each participant in the pretest may not view all the pictures that are being pretested, it is preferable to compromising the validity of data due to the effects of fatigue. The number of participants required in a pretest is small (about 20-30 per cue).

Once protocols from the pretested pictures have been collected and coded using the same coding system that will be used in the actual study, the researcher should select picture cues while bearing in mind the principles of ambiguity, cue strength, universality, relevance, and test extensity. For single-motive measurement, the main issue is making sure that there is sufficient pull for the motive in question. Thus, pictures selected for single-motive measurement should demonstrate high cue strength for the motive of interest. However, for multi-motive measurement, the balance between ambiguity and cue strength is more important. Picture cues selected for multi-motive picture sets will naturally have greater cue ambiguity, which brings down cue strength. For this reason, it is advisable when administering a multi-motive picture set to either choose a picture set containing pictures which all possess moderate to high cue strength for several motives, or to alternate picture cues of high cue strength for one motive with picture cues of high cue strength for another motive.

Picture Order and Placement. Pang and Schultheiss (2005) have explored the influence of picture position on motive scores and found that varying a picture's position in a sequence of pictures has little effect on total amounts of *n* Power, *n* Affiliation, and *n* Achievement imagery obtained for that picture cue. In general, picture order effects are minimal, especially if picture order can be randomized across participants. However, if picture serial position cannot be randomized, Smith, Feld, and Franz (1992) recommend that pictures with low and high pull for a given motive should be alternated within a battery of pictures. Their recommendation is based on the rationale that a picture cue with high cue strength will decrease the consummatory value of subsequent picture cues. The alternating of high- with low-pull pictures mimics the waxing and waning motivational process and increases overall validity of the instrument. In the same way, pictures with similar thematic contexts (e.g., showing competitive situations; showing potential disruption of familial relations) should be alternated with pictures with dissimilar contexts.

Test

During the test stage, care must be taken to ensure that all PSE administration conditions are standardized. If possible, coders can start to be trained in the coding system of choice since the coders will require sufficient time to achieve satisfactory intercoder agreement with practice or pilot materials, and –in the case of multiple coders—with each other. Among the various administration conditions that the researcher should pay attention to, are, the setting and experimenter characteristics, mode of administration (group versus individual; hand-written versus typed), instructions, and participant characteristics.

Administration Setting. The main aim of PSE administration is to exert as little pressure as possible on participants. For this reason, situational factors such as locality, mood of surrounding atmosphere, experimenters, instructions, and timing and tone of the experiment should be standardized as much as possible and presented in a way that does not exert too much social or time pressure on participants. Hence, experimenters should be as inconspicuous as possible, avoid referring to the PSE as a "test," and avoid openly using a stopwatch or any kind of time-tracking device (c.f. Murstein, 1965; Lundy, 1988).

Because of the assumed sensitivity of motives to situational incentives, the PSE should be administered before other components of the experiment, e.g., mood induction or any other experimental manipulation (Veroff, 1992; Lundy, 1988).

The PSE can be administered individually or in small groups of fewer than eight participants (c.f., Schultheiss & Brunstein, 2001). Although previous research suggests that individual and group testing produce similar motive scores (Lindzey & Heinemann, 1955), Schultheiss and Brunstein (2001) showed that groups larger than eight are not optimal because the presence of others starts to become more obvious to the participant and may prompt social evaluative and/or affiliative concerns. While limiting group size makes it easier to control for potential interactions (either between participants or between the experimenter and participants) that may have unintended motive-arousing effects, individual testing is also not desirable, because it is difficult to establish the same testing situation and rapport between the participant and the experimenter for all participants.

One way out of this dilemma is to use personal computers for the presentation of PSE instructions and picture stimuli and the recording of responses. Blankenship and Zoota (1998) utilized computerized administration and compared typed responses with handwritten ones. They found no significant differences in motive imagery between the two data collection formats. Similar findings were also reported by Schultheiss, Liening and Schad (2008), who recommended the use of computer-based PSE administration for future research on implicit motives. Where computer administration is not possible, experimenter-administered PSE sessions can be conducted quite effectively—based on personal experience and research convention, sessions with 4-8 participants allows the experimenter to interact with participants in a way that avoids most of the problems posed by too-large or too-small groups.

Instructions. Lundy (1988) showed the importance of instructions in ensuring that an administration setting has neutral connotations for participants. When instructions provided an ego threat, emphasized that the TAT was a personality measure, or stressed the importance of following rules and instructions carefully, resulting motive scores failed to correlate with criterion measures. Thus, care must be taken to ensure that instructions are as neutral and nonthereatening as possible since nonneutral instructions elicit less valid motive scores.

The following instructions, adapted from Schultheiss and Pang (2007), have been used successfully in numerous studies. They are standard instructions that have been compiled from various sources (Lundy, 1988; Atkinson, 1958; Smith, 1992) and may either be spoken by an experimenter, printed on an instruction page and distributed to each participant, or displayed on a computer screen:

> "PICTURE STORY EXERCISE
>
> "In the Picture Story Exercise, your task is to write a complete story about each of a series of [number of] pictures—an imaginative story with a beginning, a middle, and an end. Try to portray who the people in each picture are, what they are feeling, thinking, and wishing for. Try to tell what led to the situation depicted in each picture and how everything will turn out in the end.
>
> "On your desk are [number of pictures] sheets of paper for you to write your stories on. They are labeled PSE 1 through PSE [number of pictures]…In the upper left hand corner of each writing sheet there are some guiding questions—these…are only guides to writing your story. You do not need to answer them specifically.

> "Each picture will be presented for 10 seconds. After that, please write whatever story that comes to mind. Don't worry about grammar, spelling, or punctuation—they are of no concern here. You will have about 5 minutes for each story. I will tell you when there are 20 seconds remaining and when it is time to move on to the next picture."

The guiding questions referred to in the above instructions are printed at the top left hand corner of every writing page (c.f., Schultheiss & Pang, 2007):

What is happening? Who are the people?

What happened before?

What are the people thinking about and feeling?

What do they want?

What will happen next?

In a typical PSE administration, after participants have read or been given the abovementioned general instructions, they are allowed to view each PSE picture for between 10-15 seconds, after which they are prompted to start writing. Participants should not be allowed to refer back to the picture after they have started writing, because constant reference encourages stories that contain purely descriptive elements of a picture rather than imaginative material. Generally, the writing time allocated to each picture is 5 minutes—participants are alerted when they have about 20-30 seconds remaining and then asked to move on to the next picture at the end of 5 minutes. If researchers wish to test more pictures within the same overall time frame, writing time can be shortened to 4 minutes per picture without dramatically sacrificing resulting amount of codeable material. Additionally, Schultheiss, Liening, and Schad (2008) found that people generally produce about 30% more material when typing their responses, so writing-time per story in computer-administrations can be reduced to 4 minutes without substantially compromising amount of codeable data.

Blankenship and Zoota (1998) recommend that researchers should not set a time limit during individual computer administration, since participants vary in their typing abilities, however, in group settings, researchers may find it easier to standardize administration conditions if they set a time limit, even for computer administrations. In experimenter administered studies, the experimenter can monitor and be sensitive to the writing speed of all participants in the group, perhaps encouraging the group to move on to the next picture, sometimes even before the allotted time of 5 minutes is up, if all participants have clearly finished writing.

For computer administrations, researchers may use a detectable but low-key message to prompt participants to move on to the next phase in the PSE. For instance, Schultheiss and Pang (2007) recommend pairing a blinking message such as "Please finish your story and press the [previously specified] key to move on to the next picture" with a short beep every 10 seconds in order to remind participants to move on to the next picture. Another precaution that should be taken during computer administrations is to program the software to allow participants to move on to the next picture only after enough time (e.g., 4 min) has elapsed for them to produce enough story material.

Participant gender considerations. Prior research has shown that different samples are differentially responsive to different picture cues (e.g., Bellak, 1975). Accordingly, researchers have been interested in whether the validity of the PSE can be increased by tailoring it to the population of interest. For instance, researchers typically use either gender-balanced picture sets or different pictures for males and females (Smith, Feld, & Franz, 1992).

Previous research studying whether different PSE pictures pull for different stories as a function of the participant's sex has confounded participant sex with sex of persons depicted in the picture cue (Worchel, Aaron, & Yates, 1990). Stewart and Chester (1982), in their comprehensive review of research on the question of sex differences in PSE stories, concluded that research conducted over the 25 years that they surveyed was inconclusive about gender differences in *n* Affiliation. Chusmir (1983; 1985) conducted a study in which male and female subjects were given a picture set that was balanced for sex. The results indicated that sex of the pictures had no significant main or interaction effects for scores on *n* Achievement or *n* Power although cues that depicted female characters produced higher *n* Affiliation scores than cues that depicted male characters, and this effect was greater for women than for men. Some recent work has either found no significant gender differences for *n* Achievement, *n* Power, or *n* Affiliation (e.g., Langan-Fox & Grant, 2006; Tuerlinckx, De Boeck, & Lens, 2002) or significantly higher *n* Affiliation score in women's protocols (Pang & Schultheiss, 2005; Schultheiss & Brunstein, 2001). The pattern of findings suggests that participant and picture gender may affect, if anything, *n* Affiliation scores, albeit not consistently so. Thus, although women generally write more than men do and their affiliation scores are slightly higher (Schultheiss & Brunstein, 2001; Pang & Schultheiss, 2005), no other motive differences exist between men and women.

Posttest

During the posttest stage, identifying information should be removed from protocols before they are coded. Additionally, coders should score material only after having obtained sufficient interrater reliability with practice materials of the chosen scoring system, and they should at minimum be blind to the study conditions in which participants are in, if not to the study hypotheses. Blind coding reduces coder biases which may compromise validity.

Other major considerations during the posttest stage have to do with processing of protocols to prepare for coding, the selection of a suitable coding system and training coders to the selected system, and the calculation and reporting of inter-rater reliability.

Protocol transcription and Processing. Ideally, handwritten entries should be transcribed in order to increase accuracy of coding and ease of data storage and sharing. If the protocols cannot be transcribed, then coders should make their markings on photocopies rather than originals. This precaution of preserving the originals allows future re-analyses or re-coding without the danger of leaving coding marks that will bias future coders.

In preparation for scoring, all identifying information about the participant and experimental condition should be removed from the protocols, in order to preserve the participant's anonymity as well as to prevent the "halo effect," which happens when the coder may have formed impressions about the participant during the process of scoring that influence the future scoring of subsequent, ambiguous responses. In order to prevent halo effect, the coder may wish to score stories randomly either within or across participants. However, I would recommend against following Smith, Feld, and Franz's (1992) suggestion that the coder score all stories for one cue before moving on to all stories for another cue, as my personal experience is that this strategy increases the likelihood of *scorer drift*, which is the tendency of forming implicit rules-of-thumb after continuously scoring many stories with similar content. Scorer drift is minimized by scoring stories by participant instead of by cue, particularly if picture cue sequence is randomized within participants.

At this point, it may be necessary to drop certain participants who either display low verbal fluency or whose stories clearly reflect a lack of cooperation. Specifically, stories with fewer than 30 words have been found to be unscorable and should thus be excluded (c.f. Walker & Atkinson, 1958). Additionally, through an initial reading of the protocols and experiment logs, it may become obvious to the coder when a participant has clearly misunderstood the task, or is uncooperative. For instance, an uncooperative subject might intentionally include rampant verbal insults, violent or facetious imagery, or nonsensical storylines, or they may repeatedly and explicitly state their disdain for the story writing task.

These uncooperative participants should be dropped from coding and further analyses. However, if numerous participants in a sample show lack of cooperation, the researcher should investigate whether there was some problem with the data collection procedures (e.g., a disturbing public event, an overly directive experimenter) and whether the entire group's data should be discarded.

Coding systems. There are numerous content coding systems available to the interested researcher. The most commonly used coding systems are for *n* Achievement, *n* Affiliation, *n* Intimacy, and *n* Power. These coding systems are summarized in more detail Schultheiss and Pang (2007) and their full versions are compiled in Smith (1992).

Briefly, these coding systems share two common traits. First, all of these systems were constructed through the examination of differences in story themes in experimentally-manipulated or naturally-occurring groups that differ in degrees in strength of a given motive. Second, the subcategories in these systems all follow a general framework for a hypothetical sequence of motivated behavior. Specifically, the motivated behavioral sequence is initiated when the person or persons experiences a *Need* and then engages in goal-directed *Instrumental Activity* in order to fulfill this need. This instrumental behavior could result in either *Positive or Negative Affect* as well as *Positive or Negative Goal Anticipation*, depending on either successful or unsuccessful goal progress. A scoring decision is first made for the absence or presence of motive-relevant imagery before moving on to score subcategories; typically, one point is awarded for the presence of each category or subcategory and each category or subcategory can only be scored once per story.

For instance, in the *n* Achievement scoring system (McClelland, Atkinson, Clark, & Lowell, 1953), stories are scored for achievement-related imagery if there is a mention of *competition with a standard of excellence*. Having scored presence of achievement-related imagery, the coder then goes on to score for subcategories of *need, instrumental activity*, *positive and negative anticipatory goal states*, *blocks to goal progress*, *positive or negative affect*, *achievement thema* (a weighting category given when the achievement imagery is the central theme of the story).

Similarly, in the *n* Affiliation scoring system (Heyns, Veroff, & Atkinson, 1958), *n* Affiliation related imagery are scored whenever there is concern for establishing, maintaining, or restoring positive relations with others. Once the *n* Affiliation imagery is scored, the coder goes on to look for subcategories that are similar to those for *n* Achievement, e.g., *need*, *instrumental activity,* etc.

In the *n* Intimacy scoring system (McAdams, 1980), the coder first decides if there is *n* Intimacy related imagery by looking out for either dialogue or any verbal or nonverbal exchange between story characters. After scoring for the presence of *n* Intimacy imagery, subcategories such as *psychological growth and coping* and *time- or space-transcending quality of a relationship* can be scored.

In the *n* Power scoring system (Winter, 1973), the coder scores *n* Power related imagery whenever a story character is concerned about having an impact on other people. Once *n* Power content is identified, then subcategories such as *prestige of actor* and *stated need for power* can be scored.

Finally, Winter's (1994) Manual for Scoring Motive Imagery In Running Text distills the four coding systems described above into its major coding categories only. Winter's (1994) integrated system is popular with researchers because it is an abbreviated version and hence less time-consuming to learn and to use. The integrated system also combines *n* Affiliation and *n* Intimacy into a single conjoint category because of the presumed theoretical overlap between the two concepts (see chapter by Weinberger et al, this volume)

When selecting a coding system, the researcher should consider the complexity and the conceptual background of the system. Some systems are similar conceptually but may differ in complexity (e.g., original *n* Achievement system and the *n* Achievement sub-system of the Winter system for running text). There will be a tradeoff between complexity and construct validity on one hand and interrater reliability on the other. The less complicated system will be easier to learn and achieve greater levels of interrater reliability but the more complicated system will include a more comprehensive and differentiated representation of the construct.

Coder Training. Before coding any of the study materials, the coder needs to spend sufficient time reading the coding definitions and categories associated with each coding system. Practice materials consisting of sample protocols and “expert” coding answer keys are usually provided with each coding manual. The amount of practice required depends on the size and complexity of the coding system. Generally, novice coders should undergo at least 12 hours of scoring practice material before moving on to scoring any PSE protocols (c.f., Smith, Feld, & Franz, 1992). In the event that coders fail to achieve at least 85% agreement with coding materials, even though they have completed all available practice materials in the coding manual, they should re-read the coding manual carefully and re-score practice materials (restarting with the earliest practice stories) until the 85% criterion is reached.

An efficient way of improving accuracy of coding is to encourage coders to construct a ‘crib’ sheet during training, to which each coder should refer liberally during the actual coding stage. This document—preferably not more than 2 pages long to maintain ease of reference—represents a condensation of the coding system and contains the main points, definitions, and descriptions of each coding category, as well as exceptions, examples, and other important coding conventions that the coder has discovered from experience. By referring liberally to this customized crib sheet during coding, the coder will find it easier to remain faithful to coding categories and thus minimize scorer drift.

Coding. It is very important to make coding decisions based on information available in the data only, and not from inference. It is equally important that the coder should be able to justify every coding decision by noting which coding category an image falls under; an image that does not fall into any coding category should not be scored, no matter how obvious a manifestation of the motive it may seem to the coder. It is recommended that coders err on the side of caution and refrain from scoring any imagery that is marginal. When coding, a useful maxim is, "When in doubt, leave it out."

Another method for reducing coding errors is to enter motive scores at as fine-grained a level as possible. Specifically, reliability is improved when motive scores are entered at the subcategory level rather than the category level, since coders are forced to refer explicitly to each subcategory when making coding decisions. Thus, for the *n* Achievement coding system, coders should enter scores for *Need*, *Instrumental Activity*, *Goal State*, and so on.

Once motive scores are entered, researchers should check for outliers, examine score distributions to total scores for each motive, and make adjustments using statistical transformations to correct any scores that do not conform to a normal distribution.

If motive scores correlate with protocol length, they should be subjected to word count correction before being used together with other variables in data analysis. There are two ways of correcting for word count. A simple method is to multiply the total motive score by 1000 and then divide the result by the total word count for each participant. The resulting score can be easily interpreted as motive images per 1000 words. While image per 1000 words is readily interpretable and allows for easy between-sample comparisons, this correction method sometimes creates artifacts because the resulting motive scores are not necessarily 0-correlated with word count. Thus, researchers who use this method should always check again for the variance overlap between corrected motive scores and total word count. Another commonly used word count correction method is to residualize motive scores for word count and use the resulting residuals in subsequent analyses. However, while this method is effective at removing the influence of verbal fluency on motive scores, the residualized motives scores are not as easily interpreted and will not be directly comparable between different samples and studies.

If there is more than one coder, an adequate degree of agreement (at least 85%) must be obtained between the coders before scoring of raw data can commence. Coders should score stories from a pilot study or, in lieu of that, from a small subset of stories from the actual study in order to establish interrater agreement before going on to score stories from the rest of the respondents. This is because there are some coding conventions that may be idiosyncratic to each study and which have to be agreed through discussion between the coders. Coding discrepancies should be resolved through consensus, in the view of establishing coding guidelines that are specific to the idiosyncrasies of the particular participant sample but which are still consistent with the coding categories of the scoring manual. Once coders have established the requisite 85% reliability with each other, however, they should score the remaining protocols independently, without discussion or collaboration.

Typically, most participants produce stories around 100 words in length and an experienced coder will need 2-5 min to score a PSE story; hence a typical study with 6-picture protocols from 100 participants will take between 20 and 50 hours to code (with additional time needed to review the scores and to enter scores into the data spreadsheet).

Inter-rater reliability reporting. In published work, an index of interrater agreement between two or more independent coders should be reported. There are a number of methods to determine interrater reliability: Spearman’s rank correlation, index of concordance, and the intraclass correlation coefficient. Specifically, the index of concordance between two coders, A and B, is calculated by using the following formula:

(2 x agreements on motive imagery)/ (total motive imagery score for coder A + total motive imagery score for coder B).

The intraclass correlation coefficient (ICC) is a chance-corrected reliability coefficient for continuous data that is equivalent to kappa under appropriate conditions (Meyer et al., 2002). The one-way random effects ICC is especially desirable as a measure of interrater reliability because it calculates correlations between observations that do not have an obvious order (i.e., who is “rater 1” and “rater 2” is irrelevant). The one-way random effects ICC is calculated by the following formula:

[MS (between coder) – MS (within coders)]/ [MS (between coders) + MS (within coders)].

Retest

As previously noted, if a researcher is interested in stability and change in motive scores, a second PSE will be administered to participants in the retest stage. The retest stage is largely similar to the test stage with the exception of a slight modification in the instructions to participants should the same picture cues be used in subsequent sessions. As with the test stage, administration conditions should be neutral and standardized so as to minimize the influence of demand characteristics.

Researchers have previously recommended some strategies for modifying retest instructions so as to improve test-retest reliability (Reumann, 1982; Winter & Stewart, 1977). Winter and Stewart (1977) found that test-retest reliability coefficients for *n* Power after one week were relatively high when participants were either given explicit instruction to write the same story that they had at the previous testing (r = .61) or when they were told not to worry about the degree of similarity between the two sets of stories (r = .58). In contrast, participants asked to write a different story in the second testing session had significantly lower test-retest reliability coefficients of .27.

The authors concluded that during retesting, participants feel the pressure to be original in their storytelling; when instruction is given to ignore the similarity between their stories in both sessions, test-retest reliability increased. For this reason, researchers have since recommended that retest instructions should include a statement which assures participants that they are not expected to produce stories that are different from those in a previous PSE testing session (Schultheiss & Pang, 2007):

> "You may remember seeing some of these pictures before. If you do, feel free to react to them as you did before, or differently, depending on how you feel now. In other words, tell the story the picture makes you think of now, whether or not it is the same as the one you told last time."

Additionally, in a meta-analysis of published and unpublished data from studies employing empirically derived coding systems and standardized PSE administration conditions, Schultheiss and Pang (2007) showed that average motive stability coefficients are .71 for a 1-day retest interval, remain fairly high at .52 for a 1-month interval, and drop to .25 if the retest interval is extended to10 years. These levels of retest stability are adequate over time and show a rate of decrease that is similar to that found for self-report trait measures (e.g., Schuerger, Zarrella, & Hotz, 1989).

Introducing an N Achievement Picture Set

Using the abovementioned procedures for pretesting new picture cues, I pretested a set of eleven picture cues selected specially for measuring *n* Achievement in three different samples (total $\underline{N}$ = 81) of American college-aged students. Of these pictures, *director*, *man-at-desk*, and *workers* were obtained from Heckhausen (1963), *women in laboratory* is a well-used picture that was originally used by McClelland and his associates in developing practice sets for their scoring system (c.f., Smith & Franz, 1992), and the other pictures come from numerous print and media searches. All the pictures were scored for *n* Power, *n* Affiliation, and *n* Achievement using Winter's (1994) manual for scoring motive imagery in running text. Table 3 presents average motive scores for all three motives, organized by picture cue.

As shown in Figure 2, six pictures—*women in laboratory*, *skaters*, *pianist*, *footballers*, *chemist*, and *gymnast*—with low cue ambiguity and moderately high to high cue strength were eventually chosen to become part of this *n* Achievement picture set. These six pictures had high pull for *n* Achievement, with the percentage of participants having at least one instance of *n* Achievement imagery per story ranging from 54% to over 90%. Additionally, in each of these six pictures, *n* Achievement scores were higher than compared to the other two motives.

While cue strength relates to the intensity of the motive evoked by a picture, picture set extensity relates to the ability of a picture set to depict different motive-relevant situations. The six pictures in the *n* Achievement picture set illustrate motive-relevant behavior in at least three distinct contexts: in competition, at work, and during sports and leisure, thus demonstrating the picture set's degree of motive extensity.

Summary

This step-by-step guide offers a systematic description of different aspects of preparing, administering, scoring, and processing PSE data. Some important points such as picture set extensity and administration conditions have been highlighted because these elements are easily standardized but if they are not standardized, can greatly affect the reliability and validity of the data. By following this guide the interested researcher should be able to successfully carry out PSE-based motivational research. Additionally, this chapter provides recommendations for developing motive-specific picture sets and gives *n* Achievement, *n* Power, and *n* Affiliation cue strength statistics for numerous commonly-used picture cues as well as cue strength statistics for a dedicated *n* Achievement picture set.

The reader may wonder whether it is worth it to undergo so much preparation and careful handling to administer the PSE. As Veroff (1992) argues, the PSE enables us to have richer data to study how different life experiences affect motivational striving in different ways. Also, the conceptual motivational process folded into the measure combined with the coherent storyline of PSE protocols ensures that we can test general propositions about the nature of motives, especially how motives are resolved over the life span and immediate time frame. By means of content analysis, large amounts of qualitative information can be reduced to smaller and more manageable forms of representation of quantitative categories, frequencies, and ratings. Finally, as McClelland et al. (1989) have shown, PSE and other related measures are capable of assessing implicit motivation into which self-report measures are unable to tap.

Moreover, as recent empirical work reviewed in this chapter has shown, it is possible to improve the reliability and validity of the instrument by reducing errors in data collection, ensuring adequate scorer training, developing standardized and neutral experimental conditions, and creating new picture cues that efficiently tap into the motive(s) of interest. Nonetheless, more research is needed to investigate other important issues. Some ideas for future work include: producing validation of picture sets by testing the ability of motive scores derived from the sets to predict motive-relevant behavior; embarking on more systematic investigations of the gender difference issue as well as the issue of participant gender-cue gender interactions; customizing the PSE method to investigate more and other motives (e.g., sex, hunger).

The PSE is a complex instrument to learn and to use. However, based on personal experience and review of recent literature, steep learning curves have not prevented the discovery of meaningful findings. Given the significant advantages of using the PSE, our goal for using this rich instrument is to maximize sources of validity and minimize any sources of error. It is hoped that this chapter has contributed to the effort for better measurement of implicit motives and that researchers will continue to develop and refine the instrument towards this cause.

References

Atkinson, J. W. (1958). Motives in fantasy, action, and society: a method of assessment and study. Oxford, England: Van Nostrand.

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.

Bellak, L. (1975). *The Thematic Apperception Test, the Children's Apperception Test and the Senior Apperception Technique in clinical use*. New York: Grune & Stratton.

Blankenship, V., & Zoota, A. L. (1998). Comparing power imagery in TATs written by hand or on the computer. Behavior Research Methods, Instruments & Computers, 30(3), 441-448.

Chusmir, L. H. (1983). Male-oriented vs. balanced-as-to-sex thematic apperception tests. Journal of Personality Assessment, 47, 29-35.

Chusmir, L. H. (1985). Motivation of managers: Is gender a factor? . Psychology of Women Quarterly, 9, 153-159.

Clark, R. A., & Sensibar, M. R. (1955). The relationship between symbolic and manifest projections of sexuality with some incidental correlates. The Journal of Abnormal and Social Psychology, 50, 327-334.

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

Entwisle, D. R. (1972). To Dispel Fantasies About Fantasy-Based Measures of Achievement Motivation. Psychological Bulletin, 77, 377-391.

Haber, R. N., & Alpert, R. (1958). The role of situation and picture cues in projective measurement of the achievement motive. In J. W. Atkinson (Ed.), Motives in fantasy, action, and society (pp. 644-663). New York: Van Nostrand.

Heckhausen, H. (1963). *Hoffnung und Furcht ub der Leistungsmotivation [Hope and fear components of achievement motivation]* Meisenheim am Glam: Anton Hain.

Heyns, R. W., Veroff, J., & Atkinson, J. W. (1958). A scoring manual for the affiliation motive. In J. W. Atkinson (Ed.), *Motives in Fantasy, Action, and Society* (pp. 205-218). Princeton, NJ, USA: Van Nostrand.

Jackson, D. N. (1984). Personality Research Form (3rd Ed.). Port Huron, Saginaw, MI: Sigma Assessment Systems, Inc.

Koestner, R., & McClelland, D. C. (1992). The affiliation motive. In C. P. Smith (Ed.), Motivation and personality: Handbook of thematic content analysis (pp. 205-210). New York: Cambridge University Press.

Langan-Fox, J., & Grant, S. (2006). The Thematic Apperception Test: Toward a Standard Measure of the Big Three Motives. Journal of Personality Assessment, 87(3), 277-291.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The Scientific Status of Projective Techniques. *Psychological Science in the Public Interest, 1(2),* 27-66.

Lindzey, G., & Heinemann, S. H. (1955). Thematic Apperception Test: individual and group administration. *Journal of Personality and Social Psychology, 24,* 34-55.

Lundy, A. (1988). Instructional set and Thematic Apperception Test validity. Journal of Personality Assessment, 52, 309-320.

McAdams, D. P. (1980). A thematic coding system for the intimacy motive. *Journal of Research in Personality, 14(4),* 413-443.

McAdams, D. P., & Constantian, C. A. (1983). Intimacy and affiliation motives in daily living: An experience sampling analysis. *Journal of Personality and Social Psychology, 45(4)*, 851-861.

McClelland, D. C. (1975). Power: The Inner Experience. New York: Irvington.

McClelland, D. C. (1987). *Human Motivation*. New York, NY, US: Cambridge University Press.

McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). The achievement motive. East Norwalk, CT, US: Appleton-Century-Crofts.

McClelland, D. C., Clark, R. A., Roby, T. B., & Atkinson, J. W. (1949). The projective expression of needs. IV. The effect of the need for achievement on thematic apperception. Journal of Experimental Psychology, 39, 242-255.

McClelland, D. C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? Psychological Review, 96(4), 690-702.

McClelland, D. C., & Steele, R. S. (1972). Motivation workshops. New York: General Learning Press.

Meyer, G. J., Hilsenroth, M. J., Baxter, D., Exner, J. E., Fowler, J. C., Piers, C. C., et al. (2002). An examination of interrater reliability for scoring the Rorschach, comprehensive system in eight data sets. Journal of Personality Assessment, 78, 219-274.

Morgan, C. D., & Murray, H. A. (1935). A method for examining fantasies: The Thematic Apperception Test. Archives of Neurology and Psychiatry, 34, 289-306.

Murstein, B. I. (1965). Projection of hostility on the TAT as a function of stimulus, background, and personality variables. *Journal of Consulting Psychology, 29(1)*, 43-48.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review 84, 231-259.

Nunnally, J. C. (1978). *Psychometric theory*. NY: McGraw-Hill.

Pang, J. S., & Schultheiss, O. C. (2005). Assessing implicit motives in U.S. college students: Effects of picture type and position, gender and ethnicity, and cross-cultural comparisons. Journal of Personality Assessment, 85, 280-294.

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

Schuerger, J. M., Zarrella, K. L., & Hotz, A. S. (1989). Factors that influence the temporal stability of personality by questionnaire. *Journal of Personality and Social Psychology, 56(5)*, 777-783.

Schultheiss, O. C., & Brunstein, J. C. (2001). Assessment of implicit motives with a research version of the TAT: Picture profiles, gender differences, and relations to other personality measures. Journal of Personality Assessment, 77(1), 71-86.

Schultheiss, O. C., & Brunstein, J. C. (2005). An Implicit Motive Perspective on Competence. In A. J. Elliot & C. S. Dweck (Eds.), Handbook of competence and motivation (pp. 31-51). New York: Guilford Publications.

Schultheiss, O. C., Liening, S., & Schad, D. (2008). The reliability of a Picture Story Exercise measure of implicit motives: Estimates of internal consistency retest reliability, and ipsative stability. *Journal of Research in Personality,* 42, 1560-1571.

Schultheiss, O. C., & Pang, J. S. (2007). Measuring implicit motives. In R. W. Robins, R. C. Fraley & R. F. Krueger (Eds.), Handbook of research methods in personality psychology (pp. 322-344). New York: Guilford Press.

Smith, C. P. (Ed.). (1992). Motivation and personality: handbook of thematic content analysis. Cambridge [England]; New York, NY, USA: Cambridge University Press.

Smith, C. P. (2000). Content analysis and narrative analysis. In H. T. Reis & C. M. Judd (Eds.), Handbook of research methods in social and personality psychology (pp. 313-335). New York: Cambridge University Press.

Smith, C. P., Feld, S. C., & Franz, C. E. (1992). Methodological considerations: steps in research employing content analysis systems. In C. P. Smith (Ed.), Motivation and Personality: Hand-book of thematic content analysis (pp. 515-536). Cambridge, MA: Cambridge University Press.

Smith, C. P., & Franz, C. E. (1992). Appendix I: Practice materials for learning the scoring systems. In C. P. Smith (Ed.), (pp. 537). New York: Cambridge University Press.

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.

Suedfeld, P., Tetlock, P. E., & Streufert, S. (1992). Conceptual/integrative complexity. In C. P. Smith (Ed.), Motivation and personality: Handbook of thematic content analysis (pp. 393-400). New York: Cambridge University Press.

Taylor, C. R., Lee, J. Y., & Stern, B. B. (1995). Portrayals of African, Hispanic, and Asian Americans in magazine advertising. American Behavioural Scientist, 38, 608-621.

Tuerlinckx, F., De Boeck, P., & Lens, W. (2002). Measuring needs with the Thematic Apperception Test: A psychometric study. *Journal of Personality and Social Psychology, 82(3)*, 448-461.

Veroff, J. (1992). Thematic apperceptive methods in survey research. In C. P. Smith (Ed.), Motivation and personality: Handbook of thematic content analysis (pp. 100-109). New York: Cambridge University Press.

Walker, E. L., & Atkinson, J. W. (1958). The expression of fear related motivation in thematic apperception as a function of proximity to an atomic explosion. In J. W. Atkinson (Ed.), *Motives in Fantasy, Action, and Society* (pp. 143-159). Princeton, NJ, USA: Van Nostrand.

Winter, D. G. (1973). *The power motive*. NY: Free Press.

Winter, D. G. (1992). Content analysis of archival materials, personal documents, and everyday verbal productions. In C. P. Smith (Ed.), Motivation and personality: Handbook of thematic content analysis. New York: Cambridge University Press.

Winter, D. G. (1994). Manual for scoring motive imagery in running text. Unpublished instrument, University of Michigan, Ann Arbor.

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.

Worchel, F. T., Aaron, L. L., & Yates, D. F. (1990). Gender bias on the Thematic Apperception Test. Journal of Personality Assessment, 55, 593-602.

Table 1. Factors, stages, and standards of measurement for the Picture Story Exercise (PSE).

		Measurement	Stage of	
Factors		standard	Measurement	Selected references
Administration				
	- Mode of administration (verbal, written, typed)	V, TRR	PRE, T	Blankenship & Zoota, 1998; Schultheiss & Pang, 2007
	- Size of group (individual, group)	V,	PRE, T	Schultheiss & Pang, 2007; Schultheiss & Brunstein, 2001
	- Instructions	V, TRR	PRE, T, RT	Lundy, 1988; Winter & Stewart, 1977; Murstein, 1965; Schultheiss & Pang, 2007
	- Situational factors	V, TRR	PRE, T	Smith, Feld, & Franz, 1992; Veroff, 1992
	- Experimenter characteristics (gender, authority, formality)	V, TRR	PRE, T	Smith, Feld, & Franz, 1992; Atkinson, 1958; Klinger, 1967; Veroff, 1992
Participant characteristics	- Social context (culture, race, and gender)	V	PRE, T	Stewart & Chester, 1982; Bellak, 1975; Murstein, 1972
	- Degree of sample heterogeneity	V	PRE, T	Cramer, 1996
	- Recent life changes	V	T, RT	Koestner, Franz, and Hellman,1991
Data processing and analysis	- Data entry format	IR	POST	Schultheiss & Pang, 2007
	- Word count correction	V	POST	Schultheiss & Pang, 2007
	- Selection of scoring scheme	V	SCORE	Smith, 1992; Schultheiss & Pang, 2007
	- Coder training	IR	SCORE	Schultheiss & Pang, 2007
	- Calculating interrater reliability	IR	SCORE	Winter, 1994; Meyer, Hilsenroth, Baxter, Exner, Fowler, Piers, & Resnick, 2002
Picture cues				
	- Ambiguity	V, IC, TRR	PRE, T	Smith, Feld, & Franz, 1992
	- Universality	V, IC, TRR	PRE, T	Smith, Feld, & Franz, 1992
	- Cue strength	V, IC, TRR	PRE, T	Smith, Feld, & Franz, 1992
	- Content/themes	V, IC, TRR	PRE, T	Langan-Fox & Grant, 2006; Schultheiss & Brunstein, 2005; Pang & Schultheiss, 2005
	- Number of pictures (fatigue, variability in scores, motives being measured)	IC, TRR	PRE, T	Reitman & Atkinson, 1958; Schultheiss & Pang, 2007
	- Picture position	IC	PRE, T	Schultheiss & Pang, 2007
	- Number of motives being measured	V, IC, TRR	PRE, T	Schultheiss & Pang, 2007
	- Gender depicted	V, IC	PRE, T	Worchel, Aaron, & Yates, 1990

Note: Test-retest reliability (TRR); Interrater reliability (IR); Internal consistency (IC); Validity (V); pretesting (PRE); Testing (T); Retest (RT); Posttest (POST); Scoring (SCORE)

Table 2

Means and Standard Deviations of Raw Motive Scores by Picture Cue and Country

	N Power		*N* Achievement		*N* Affiliation	
Picture	*M*	*SD*	*M*	*SD*	*M*	*SD*
Women in Laboratory						
U.S.[a]	0.77	0.85	1.08	0.93	0.19	0.50
German[b]	0.80	0.84	0.66	0.77	0.19	0.48
Australian students[c]	1.32	1.60	1.66	2.53	0.18	0.63
Australian managers[d]	0.55	1.34	1.91	2.24	0.22	0.60
Ship Captain						
U.S.[a]	1.01	0.88	0.14	0.47	0.21	0.53
German[b]	1.16	0.92	0.11	0.37	0.20	0.53
Australian students[c]	1.63	1.71	-0.70	1.01	0.33	0.96
Australian managers[d]	1.08	1.51	-0.88	0.51	0.32	0.84
Couple by River						
U.S.[a]	0.23	0.54	0.00	0.21	2.06	1.07
German[b]	0.43	0.72	0.03	0.17	1.84	1.05
Australian students[c]	0.72	1.26	-0.90	0.58	2.34	1.72
Trapeze Artists						
U.S.[a]	0.70	0.79	0.76	0.83	0.49	0.80
German[b]	0.79	0.85	0.78	0.84	0.43	0.71

Table 2 (continued)

	N Power		*N* Achievement		*N* Affiliation	
Picture	*M*	*SD*	*M*	*SD*	*M*	*SD*
Trapeze Artists						
Australian students[c]	0.99	1.38	1.00	2.44	0.25	0.44
Australian managers[d]	0.51	1.05	0.73	2.18	0.54	1.00
Nightclub Scene						
U.S.[a]	0.75	0.82	0.01	0.30	1.32	1.10
German[b]	0.86	0.83	0.09	0.31	1.29	1.08
Boxer						
U.S.[a]	0.79	0.90	1.14	1.06	0.17	0.51
Architect at desk						
German[b]	0.22	0.46	0.29	0.55	1.16	0.84
Australian students[c]	0.82	1.32	0.11	2.01	1.09	1.40
Bicycle Race						
U.S.[e]	0.95	0.95	1.65	1.08	0.16	0.41
Woman and Man Arguing						
U.S.[e]	1.70	0.94	0.31	0.66	0.18	0.53
Hooligan Attack						
U.S.[e]	2.26	0.95	0.01	0.14	0.15	0.43
Lacrosse Duel						
U.S.[e]	0.69	0.84	1.22	0.96	0.16	0.50

Table 2 (continued)

	N Power		*N* Achievement		*N* Affiliation	
Picture	*M*	*SD*	*M*	*SD*	*M*	*SD*
Men on Ship Deck						
U.S.[f]	1.82	1.26	0.48	0.63	0.20	0.51
German[g]	1.14	0.95	0.47	0.73	0.29	0.63
Soldier						
German[g]	0.94	1.01	0.35	0.64	0.17	0.36
Soccer Duel						
German[g]	0.74	0.77	1.79	0.81	0.07	0.27
Couple Sitting Opposite a Woman						
U.S.[h]	1.05	0.98	0.40	0.79	0.73	0.90
Girlfriends in Café With Male Approaching						
U.S.[h]	0.53	0.65	0.38	0.76	1.62	1.43
Man and Woman with Horses and Dog						
Australian students[c]	0.61	1.15	-0.63	1.16	1.08	1.55
Conference group: Seven Men Around a Table						
Australian managers[d]	1.13	1.55	-0.59	1.12	0.68	1.09

Note. Underlined scores indicate that more than 50% of participants have responded with at least one instance of codeable motive imagery to the picture cue. Motive imagery for the U.S. and German samples was scored using Winter's (1994) coding system for scoring motives in running text.

Table 2 (continued)

Motive imagery for the Australian samples was scored using McClelland, Atkinson, Clark, & Lowell’s (1953) scoring manual for *n* Achievement (scores range from -1 to +11 and mean scores are based on a 6-picture set for the student sample and a 4-picture set for the manager sample), Heyns, Veroff, & Atkinson’s (1958) scoring manual for *n* Affiliation (scores range from 0 to +7), and Winter’s (1973) revised scoring manual for *n* Power (scores range from 0 to +11).

[a] From Pang & Schultheiss, 2007 (*n*= 323).

[b] From Schultheiss & Brunstein, 2001 (*n*= 428).

[c] From Langan-Fox & Grant, Study 1, 2006 (*n*= 334).

[d] From Langan-Fox & Grant, Study 2, 2006 (*n*= 213).

[e] From Wirth, Welsh, & Schultheiss, Study 2, 2006 (*n*= 109).

[f] From Schultheiss, Campbell, & McClelland, 1999 (*n*= 42; male participants only).

[g] From Schultheiss & Rohde, 2002 (*n*= 66; male participants only).

[h] From Schultheiss, Wirth, & Stanton, 2004 (*n*= 60).

All picture stimuli are available upon request from the author.

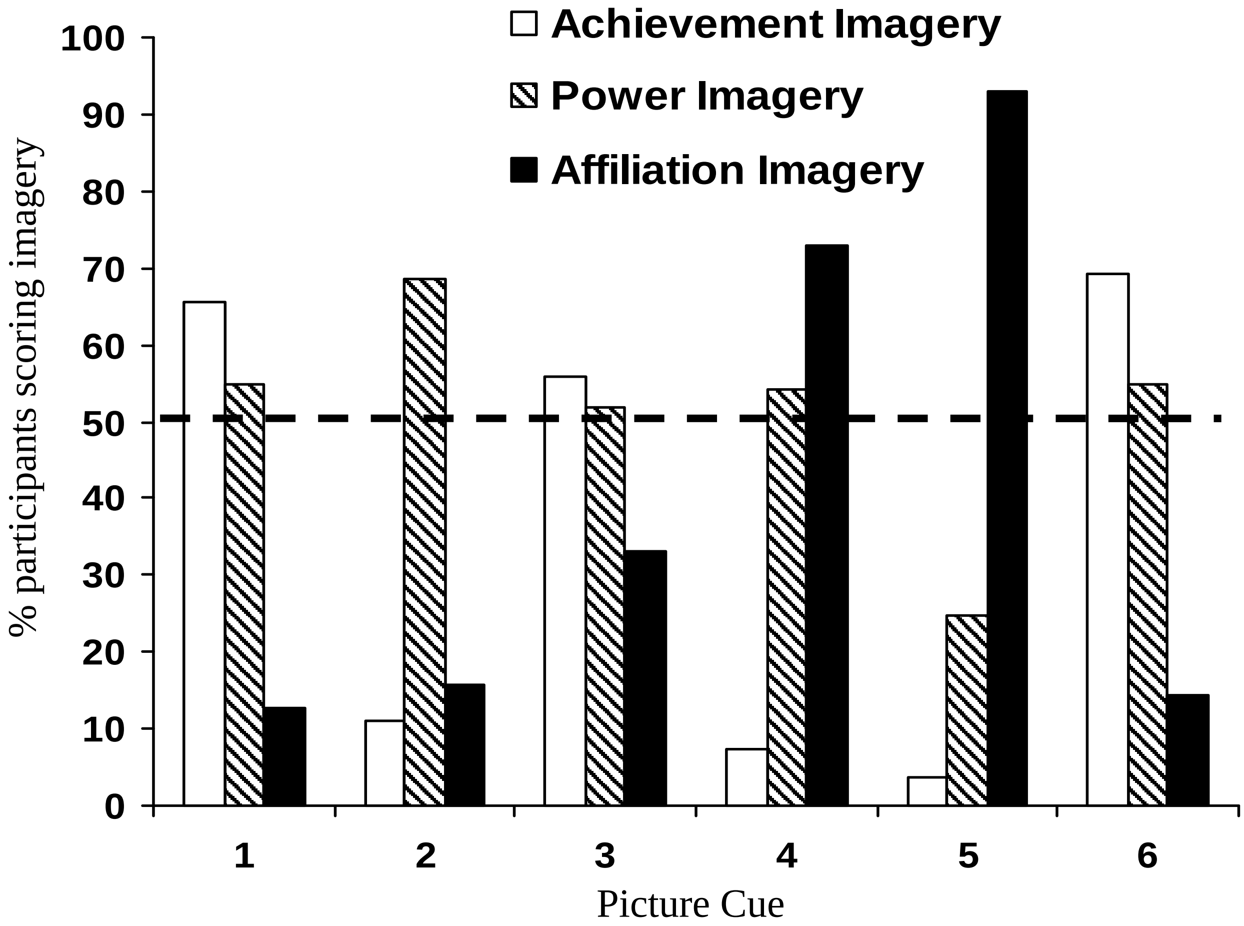

Figure 1. Percentage of participants scoring at least one motive image for each picture cue in a multi-motive picture set.

Note: N = 323. Data from Pang & Schultheiss (2005). Pictures used in this picture set are: 1 = Boxer; 2 = Ship Captain; 3 = Trapeze Artists; 4 = Nightclub; 5 = Bridge; 6 = Women in Laboratory.

Table 3. Means and Standard Deviations of Motive Scores by Picture Cue

	N Power		*N* Achievement		*N* Affiliation	
Picture	*M*	*SD*	*M*	*SD*	*M*	*SD*
Director[a]	0.36	0.60	0.57	0.74	0.04	0.20
Gymnast	.39	0.57	1.82	1.34	0.04	0.12
Mountain Climber	0.11	0.32	0.48	0.69	0.46	0.66
Soccer Duel	0.41	0.63	1.56	1.02	0.24	0.75
Man-at-Desk[a]	0.25	0.59	0.29	0.53	0.18	0.39
Pianist	0.46	0.67	0.69	0.70	0.37	0.60
Auto Mechanics	0.33	0.61	0.54	0.82	0.46	0.84
Workers[b]	0.33	0.64	0.30	0.57	0.41	0.60
Skaters	0.22	0.50	1.74	0.87	0.48	0.80
Women in laboratory[b]	0.57	0.86	0.69	0.72	0.31	0.58
Chemist	0.22	0.50	0.76	0.81	0.30	0.58

Note: N = 81. Underlined motive scores indicate that more than 50% of participants have responded with at least one instance of codeable motive imagery to the picture cue.

[a] From Heckhausen, 1963.

[b] From Smith, 1992.

All picture stimuli are available upon request from the author.

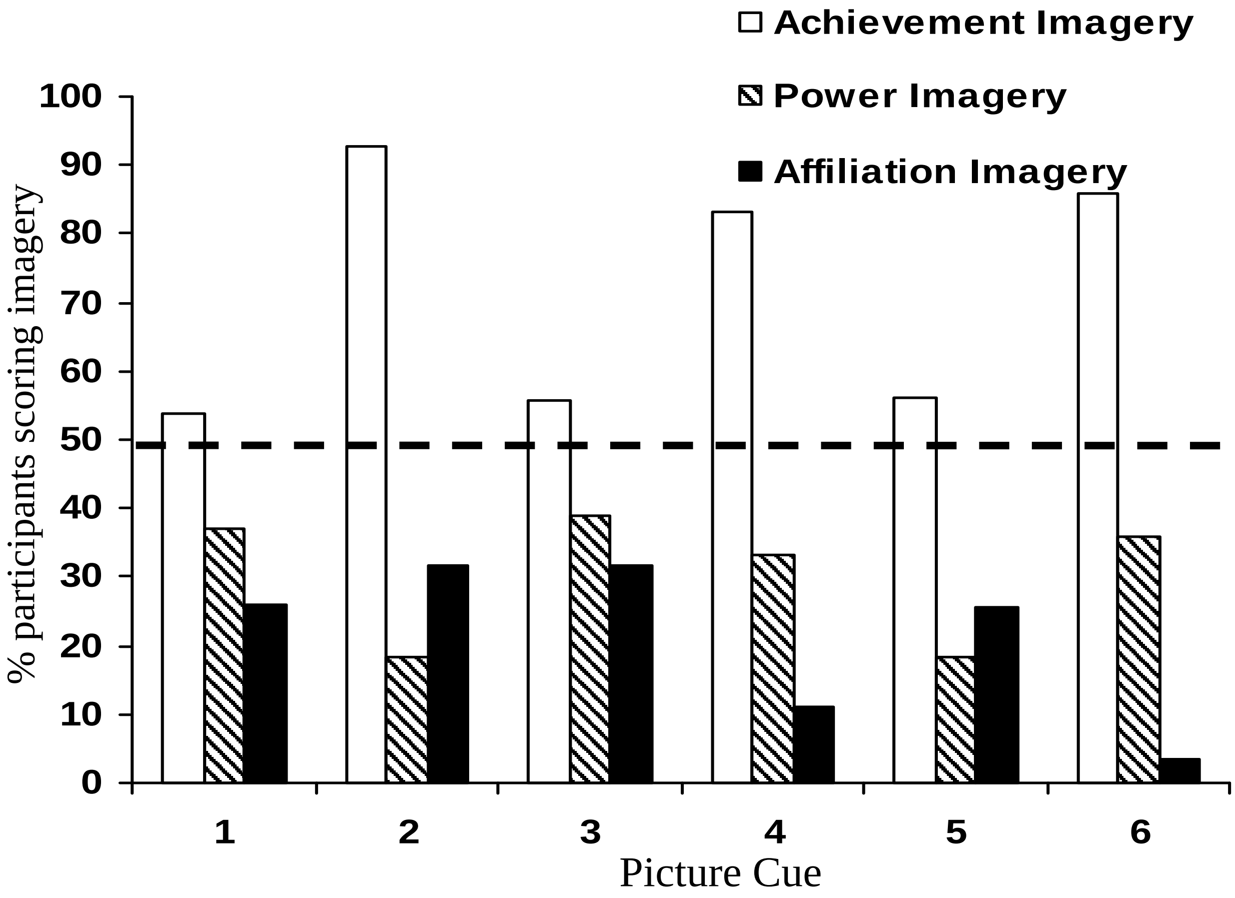

Figure 2. Percentage of participants scoring at least one motive image for each picture cue in a single-motive picture set.

Note: N = 81. Pictures used in this picture set are: 1 = Women in Laboratory; 2 = Skaters; 3 = Pianist; 4 = Soccer Duel; 5 = Chemist; 6 = Gymnast.