Motiv Emot
DOI 10.1007/s11031-012-9339-9

ORIGINAL PAPER

Set ambiguity: A key determinant of reliability and validity in the picture story exercise

Jonathan E. Ramsay · Joyce S. Pang

Abstract The picture story exercise (PSE), in which participants write imaginative stories in response to motivationally-arousing images, is the most commonly-used tool for the assessment of implicit motives. Despite decades of research into the qualities of effective individual picture cues, much less is known about the desirable properties of overall picture sets. The present research highlights a previously undocumented methodological consideration—set ambiguity—which has important implications for the reliability and validity of the PSE. In a four-part study of 74 undergraduates, motive scores derived from an ambiguous picture set comprising cues that vary in motivational focus displayed greater test–retest reliability, convergent validity, and predictive validity than those derived from an unambiguous picture set. Researchers are therefore advised to consider set ambiguity when selecting images for use in PSE research.

Keywords Motive assessment · Picture story exercise · Implicit motivation · Achievement · Methodology

Introduction

Many motivation researchers have claimed that significant aspects of human motivation are rooted in the unconscious mind (e.g. Murray 1938). These implicit motive systems drive behavior despite being inaccessible to conscious reflection; their nature and purpose frequently differing from the explanations we provide for our own actions – our explicit motives and goals (McClelland et al. 1989). Over the years a wealth of empirical evidence has supported this fundamental distinction between implicit and explicit motives, with each motivational class possessing distinct developmental origins (McClelland and Pilon 1983), incentives (McClelland et al. 1989), and behavioral outlets (McClelland 1987). Whereas explicit motives can be examined, reflected upon and verbalized, implicit motives initiate and shape behavior in more subtle ways which remain largely inaccessible to the individual in question.

An important implication of the implicit-explicit motive distinction relates to their tools of assessment. Whereas most personality constructs are assessed using self-report measures, implicit motives cannot be measured in this way since asking people to endorse self-referenced statements will necessarily arouse explicit motives. McClelland et al. (1953) sought to address this problem by adapting the Thematic Apperception Test (TAT; Morgan and Murray 1935) to their purpose of assessing the implicit achievement motive. The resulting instrument, known as the Picture Story Exercise or PSE (McClelland et al. 1989), has become the defining methodology for the assessment of implicit motives. The PSE involves sequential administration of picture cues showing individuals engaged in motive-relevant behaviors. Participants are shown each image for a brief period before being asked to write an imaginative story featuring the characters depicted. These stories are subsequently content-coded for the presence of motive-relevant imagery using objective coding systems (e.g. McClelland et al. 1953; Atkinson 1958; Heckhausen 1963; Winter 1994) derived through qualitative analysis of stories written in varying states of motive arousal.

J. E. Ramsay (✉) · J. S. Pang
Division of Psychology, School of Humanities and Social Sciences, Nanyang Technological University,
14 Nanyang Drive, Singapore 637332, Singapore
e-mail: jonathan1@e.ntu.edu.sg

Published online: 04 January 2013

Qualities of effective picture cues

While the majority of PSE research has focused on the refinement and development of these coding systems, some researchers have attempted to delineate the qualities that effective PSE picture cues should possess (c.f. Smith et al. 1992; Pang 2010a). The foremost of these criteria is motive pull, which is defined as the frequency with which the picture cue elicits imagery related to the motive in question. Specifically, a cue that causes the majority (i.e. more than 50 %) of participants to write narratives containing at least one instance of achievement-related imagery is deemed to have good cue strength, or high motive pull, for *n* Achievement (c.f. Schultheiss and Brunstein 2001). Adequate cue strength is essential since the presence of motive-relevant imagery in at least some (although perhaps not all) of the resultant stories is a necessary precondition for construct-valid motive assessment via the PSE. As such, researchers have consistently recommended using images possessing moderate to high cue strength for the motive under investigation (e.g. Haber and Alpert 1958; Pang 2010a; Smith et al. 1992; Schultheiss and Pang 2007).

A related and equally important criterion is cue ambiguity. While some picture cues tend to elicit imagery for a single motive, others exhibit cue ambiguity, which is the ability of an image to evoke multiple motives. A degree of cue ambiguity is considered desirable since unambiguous cues tend to elicit uniform stories with little variation in motive content. Consequently, motive scores derived from unambiguous images do not validly distinguish between high and low scorers, since all respondents respond in a similar fashion irrespective of their underlying motive disposition (Pang 2010a). Consequently, previous researchers recommend using cues that exhibit strong to moderate pull for the motive in question *and* weak pull for other motives (i.e. relatively low ambiguity), since this increases the likelihood that more respondents will express motive-relevant content (Smith et al. 1992). This need for balance between cue strength and ambiguity was confirmed by Murstein (1965), who found that ambiguous images with moderate pull possessed greater validity than images with either high or low pull.

Qualities of effective picture cue batteries

However, while the effectiveness of the PSE is dictated in part by the nature of the individual picture cues, it is also important to consider the relationships between those cues and the composition of the overall picture battery. Research has identified several important considerations when assembling picture cue batteries, including battery size and motive extensity, while others such as presentation order, have been found to have minimal impact on resulting motive scores (Pang and Schultheiss 2005).

For example, concerns regarding validity place upper and lower limits on the number of pictures in a PSE battery. Whereas increasing the number of items on a traditional self-report measure frequently improves its psychometric properties, writing more than eight imaginative stories has been shown to fatigue participants and compromise the validity of their resulting motive scores (Reitman and Atkinson 1958). On the other hand, Schultheiss and Pang (2007) observed an inverse relationship between motive score variance and picture battery size, with smaller batteries (i.e. those containing four images or fewer) giving rise to extremely positively skewed score distributions which are not amenable to regression analysis, even after square root or logarithmic transformation. Since these distributions approach normality with five or more pictures, the authors recommended using picture batteries of between five and eight images to balance these twin concerns of fatigue and statistical necessity.

Picture batteries can also vary in terms of motive extensity, in that they can contain images expressing a wide variety of motivationally-relevant situations and contexts (high extensity) or comparatively little variation (low extensity). For example, a hypothetical picture set could contain images depicting a range of different sporting scenarios. While this picture set might be effective in eliciting achievement-imagery from individuals with a keen interest in sports, it would almost certainly be less effective in eliciting such imagery from less athletic, more academically inclined individuals. Such a picture set would be said to exhibit low extensity, since it applies to only a limited range of motivationally-relevant situations, whereas a more diverse picture set featuring academic, sporting, and employment scenarios would be said to exhibit high extensity. Although systematic studies of extensity are lacking (Pang 2010a), researchers nonetheless recommend using picture sets that feature a wide range of motive-relevant scenarios (e.g. Schultheiss and Pang 2007; Smith et al. 1992).

Picture set ambiguity: An unknown quantity

Despite extensive empirical work on the question of picture cue ambiguity, it is currently unknown whether motive ambiguity is also desirable at the level of the picture set. While a given picture set may comprise images which satisfy ambiguity requirements individually, in that they exhibit strong to moderate pull for the target motive and lesser pull for one or more alternate motives, such a picture set would not possess *set ambiguity*, since all the constituent images pull preferentially for the same target motive. Thus, taken as a whole, picture sets tend to be unambiguous. Although researchers have previously suggested choosing images that pull preferentially for the target motive when creating picture sets (e.g. Schultheiss and Pang 2007; Smith et al. 1992), there are several reasons to suspect that a picture set that contains *only* picture cues with high pull and low ambiguity may be counterproductive.

Firstly, unambiguous picture sets may suffer from decreased validity for the same reasons that unambiguous picture cues do. More ambiguous picture sets including items which pull preferentially for alternate motives may give rise to more valid scores because of increased response variability, which in turn makes it easier to distinguish high and low target motive scorers. For example, a picture set which exclusively comprises images that pull primarily for *n* Achievement may produce similar scores for both highly and moderately achievement motivated individuals, whereas only the highly motivated individual is likely to project achievement motivation onto a picture set containing a number of images which pull preferentially for alternate motives. As such, it is possible that unambiguous picture sets would give rise to a contracted range of scores, borne of uniform responding to achievement-related content, that do not accurately reflect the true motive variability of the sample. Such scores would display reduced predictive validity, since the ability of an independent variable to predict an outcome is necessarily reduced as its variability is artificially compressed.

Secondly, Atkinson and Birch's (1970) dynamics of action (DOA) theory suggests that writing motivationally-relevant stories possesses *consummatory value*, in that expressing the underlying need is motivationally satisfying in and of itself, reducing the tendency for further expression. This theory was used to explain both the "sawtooth" pattern of rising and falling motivational strength over time and also the low internal consistency of PSE implicit motives measures (Atkinson 1992). According to DOA theory (see Atkinson 1992, pp 26–28), the tendency to write motive-relevant imagery (T) is a function of the instigating force (F; i.e. the underlying motive) and the consummatory force (C; the motivational satisfaction derived from expression). When responding to uniform, unambiguous picture sets, C remains consistently high, decreasing the tendency to write motive-relevant imagery and reducing the frequency of the aforementioned motivational wave. Over the course of a picture set, typically including between five and eight images, this may under-represent the intensity of the motive disposition, since measurement points will more likely fall in these amotivational troughs where consummatory force is high. More ambiguous picture sets may offer improvement, since the inclusion of images which arouse alternate motives enforces "breaks" during which target motive-related C can subside, increasing the tendency to write target motive-related imagery in response to subsequent pictures. As such, ambiguous picture sets may be less likely to underestimate the true values of implicit motives.

Thirdly, responses to funneled debriefing questions in previous PSE studies conducted by the present authors have suggested a degree of hypothesis awareness among participants responding to unambiguous picture sets. After completing an unambiguous achievement PSE some participants indicated that they believed that they were being asked to write stories about success and failure in achievement tasks. This is problematic, because if participants are altering their responses due to demand characteristics, then the PSE becomes more of a measure of explicit motivation, since participants are in effect endorsing self-referenced motivational statements.

There is also reason to suspect that unambiguous picture sets may also give rise to less reliable scores. This may appear to be somewhat counter-intuitive. Given our previous assertion that unambiguous picture sets may lead to reduced response variability, it might actually be expected that unambiguous picture sets give rise to scores that are more stable over time despite being less valid. However, researchers (e.g. Schultheiss and Pang 2007) have noted that PSE respondents frequently feel compelled to be original and change their stories when responding to pictures for a second time, thus reducing test–retest reliability. We foresee that this tendency will be exaggerated in individuals responding to successive administrations of an unambiguous picture set, since the similarity in content of the stories written at time one should heighten the need to introduce variety at time two, thus further reducing test–retest reliability.

It is worth noting that researchers such as McClelland employed ambiguous picture sets without directly addressing issues of set ambiguity in their work. McClelland, Clark, Roby, and Atkinson's (1949) seminal achievement motive arousal study, in which the categories of the McClelland et al. (1953) *n* Achievement scoring system were first delineated, made use of four images, of which at least one gave rise to affiliative themes: *father talking to son*, image 7BM of Morgan and Murray's (1935) TAT picture set, which frequently elicited themes of parental succor or pressure (Eron 1950, 1953). Smith et al. (1992, p 631–632) cite many such instances of implicit motivation researchers making use of more ambiguous picture sets in their research, lending credibility to the idea that their extensively validated content coding systems could have been constructed using data derived from ambiguous picture sets.

The present research

These concerns, coupled with the current lack of empirical work, speak of a need to directly investigate the importance of set ambiguity in PSE methodology. The present study seeks to address this need by comparing the validity and reliability of scores derived from an ambiguous picture set with those derived from an unambiguous picture set. The two picture batteries were compared in terms of three key test properties: test–retest reliability, convergent validity, and predictive validity.

Generally, we hypothesized that scores from the ambiguous picture set would exhibit greater test–retest reliability, convergent validity, and predictive validity than scores from the unambiguous picture set. More specifically, we hypothesized that motive scores derived from a novel ambiguous picture set would display greater 1 week stability coefficients (H1) and stronger associations with scores derived from a previously-validated and commonly used multi-motive picture set (H2), compared to scores derived from an unambiguous picture set. We also predicted that ambiguous picture set motive scores would better predict theoretically-relevant behaviors in two tasks—the Wisconsin Card Sorting Test (WCST; Grant and Berg 1948) and the Balloon Analogue Risk Task (BART; Lejuez et al. 2002)—than unambiguous picture set motive scores (H3-H6).

In this study we chose to investigate the issue of picture set ambiguity for the achievement motives of Hope of Success (HS) and Fear of Failure (FF). Both HS and FF have been found to predict choice of ambitious task goals, performance increases in difficult tasks, and higher muscle tone during mental activity, observations which are consistent with their hypothesized role as facets of achievement motivation (Heckhausen 1963, 1968, 1980). Differences have also been observed, with high HS scorers having a better memory for successful peers, while high FF scorers are more likely to remember unsuccessful peers (Pang et al. 2009). HS-motivated individuals have also been found to prefer moderately challenging goals (de Charms and Carpenter 1968), while students motivated primarily by FF perform worse under time pressure and take longer to complete their homework (Heckhausen 1980).

Given that good performance in the WCST is contingent on both speed and accuracy, we reasoned that the approach orientation of high HS scorers should manifest as a desire to maximize positive outcomes rather than minimize negative outcomes, leading to greater speed at the expense of accuracy. Such behavior would be consistent with Schultheiss and Brunstein's (2005) suggestion that HS-motivated individuals exhibit "greater tolerance for frustrations" (p 11), in that they are less concerned about the possibility of encountering momentary frustrations and temporary setbacks in their pursuit of good performance. In the WCST, this should manifest as a willingness to make a few mistakes as long as overall goal pursuit (a generally high level of speed and accuracy) is not compromised. As such, we hypothesized that HS scores from the ambiguous picture set would positively predict sorting speed (H3) and negatively predict sorting accuracy (H4), whereas unambiguous picture set HS scores would not.

In the case of the BART, a measure of risk-taking behavior, we hypothesized that ambiguous FF would negatively predict both risk-taking (H5) and task speed (H6), whereas unambiguous FF would not. Since the possibility of failure in the BART can only be mitigated by engaging in less risky behavior, it was predicted that ambiguous FF would negatively predict risk taking whereas unambiguous FF would not. Additionally, since performance in the BART is time-independent, we reasoned that scores from a valid FF measure should negatively predict task speed, since FF scorers are afforded the chance to approach the task in a more cautious and deliberative manner. Heckhausen (1980) has documented the tendency of FF-motivated individuals to spend longer on homework assignments to ensure good performance. Since performance in the BART is not judged in terms of speed, we reasoned that this task should allow FF-motivated individuals to indulge this natural tendency for caution and deliberation. This contrasts with the WCST, in which rapid responding is explicitly requested, denying FF-motivated individuals the opportunity to approach the manner which suits them best. We therefore predicted that FF scores from the ambiguous picture set would predict BART speed whereas unambiguous picture set FF set would not.

Method

Participants

All participants ($N = 80$) were undergraduate students from a large Singaporean university, who completed the assigned tasks in exchange for partial course credit. All study elements were administered in English, which is the university's teaching language as well as the vernacular in Singapore. The sample was 54.7 % female and 80.0 % ethnically Chinese with a mean age of 21.52 years ($SD = 1.83$). 4.0 % participants were Malay, 10.7 % were ethnically Indian, while the remaining 5.3 % described their ethnicity as "other". All participants were recruited from the university's introductory psychology class. Data for six participants were excluded, either in light of their failure to complete one or more of the experimental sessions or due to their generation of extremely brief PSE stories that preclude valid scoring for motive content. This resulted in a sample of 74 participants who contributed data to the final analysis.

Design and procedure

Participants were randomly assigned to one of two experimental groups: ambiguous ($n = 41$) or unambiguous ($n = 33$). Motive scores for the participants in the unambiguous group were derived from their responses to an unambiguous 8-picture set of images specifically pre-tested to pull for both HS and FF simultaneously, while motive scores for participants in the ambiguous group were derived from their responses to an ambiguous 8-picture set, comprising the six most effective of the pretested HS/FF images and two additional images which have previously been shown to pull predominantly for *n* Power or *n* Affiliation (Pang 2010b). See the materials section below for further details of the ambiguous and unambiguous picture sets.

Participants in both conditions completed a series of four experimental sessions, each one lasting around half an hour, over a period of 10 days. Session one consisted of a PSE administration of either the ambiguous or unambiguous picture set (depending on group allocation), while session two was a straightforward replication of session one that took place exactly 1 week later. Session three, which featured the validation tasks, took place on the day following session two. Finally, the fourth session, which involved the completion of another 8-picture PSE comprising previously-validated comparison images as well as explicit measures of achievement motivation, was administered 1 day after session three. Informed consent was obtained in all cases, and participants were fully debriefed after the final experimental session.

The entire experiment was administered individually and remotely. Using the web version of the popular experimentation engine Inquisit (Millisecond Software, Seattle, WA) in combination with the online survey utility Qualtrics (Qualtrics Labs Inc., Provo, UT), participants were asked to complete the various experimental tasks either from home or using the university's computer facilities. Participants were instructed to complete the tasks alone and to refrain from engaging in any other activities (e.g. internet browsing, listening to music) while the experiment was in progress. An experimenter was available during each administration to answer queries either by phone or email.

In each of the three PSE administrations, participants received the same standardized instructions adapted from Schultheiss and Pang (2007). Also in accordance with Schultheiss and Pang (2007), guiding questions which referenced these instructions remained visible in the top left hand corner of the screen throughout the experimental session. After pressing the space bar to acknowledge that they had read and understood the instructions, participants were shown each image for 10 s, after which the image disappeared and they were given 4 min to write an imaginative story. Once the 4 min time limit had been reached, participants were prompted to proceed to the next image. Image presentation order was randomized in both the ambiguous and unambiguous conditions.

All protocols were coded for HS and FF by two independent raters using the English translation of Heckhausen's (1963) measure (Schultheiss 2001). The raters, both of whom had previously reached a level of 85 % agreement with the manual's training materials, first coded 10 % of the dataset and established an inter-rater reliability of >85 % before proceeding to code the rest of the data. The final concordance rate for the two scorers across the entire dataset was 99.7 % after all initial coding disagreements were resolved by rater discussion. The few remaining disagreements were resolved by averaging counts across the two differing scores. The final scores for each story were then corrected for word count using a procedure recommended by Pang (2010b), whereby the total picture-specific motive scores for HS and FF are multiplied by 1,000 and then divided by word count, which restates the scores as the number of instances of motive-relevant imagery per 1,000 words. These word count corrected HS and FF scores were then averaged across the eight stories in each experimental session to provide mean HS and FF scores, which were entered into the subsequent analyses as indicators of hope of success and fear of failure.

Materials

Pretesting and selection of picture cues

Two rounds of stimulus pretesting were conducted in order to identify images suitable for the PSE assessment of HS and FF. Given that some researchers have raised concerns about the declining efficacy of older picture cues for implicit motive assessment (Pang 2010a), while others have noted the absence of picture sets specifically developed to pull for HS and FF rather *n* Achievement (J. Schüler, personal communication, February 10, 2011), we decided to develop a novel picture set specifically for the assessment of HS and FF in this study. These pretested images were to be used in both the ambiguous and the unambiguous picture sets, with the unambiguous picture set comprising only pretested HS/FF images, while the ambiguous picture set included a subset of these same images and two images previously shown to pull for an alternate motive.

Pretesting began with an initial cohort of 28 royalty-free candidate images which were purchased from an online vendor. Candidate images were chosen on the basis of perceived verisimilitude and the presence of one or more actors who appeared to be engaged in goal-directed behavior within an achievement-related context. Achievement-related contexts were identified from a comprehensive survey of the extant literature (c.f. McClelland et al. 1953) on PSE-measured achievement motives. We established that workplace, academic, and competitive sporting environments were the most widely-used. Additionally, images from a novel category of achievement-related contexts, referred to as creative endeavor, were also included so to compensate for the under-representation of creative pursuits (e.g. culinary arts, visual arts) in the achievement motive literature. One hundred pre-test participants subsequently judged the achievement-relatedness of this initial cohort by listing adjectives to describe each image. These adjectives were subsequently rated as either achievement-related or unrelated by two independent raters trained in the McClelland et al. (1953), Heckhausen (1963), and Winter (1994) systems, and the 18 images which solicited the greatest percentage of achievement-related adjectives were retained for further pretesting.

In the second pretesting round, the 18 retained images were assessed using full PSE methodology. Eighty-six pretest participants wrote PSE stories in response to a six-picture subset of the pretest images, and the resulting protocols were scored for HS and FF by two independent raters trained in the application of the Heckhausen (1963) coding system. The final eight images were selected on the basis of two criteria—motive pull and cue ambiguity—since many researchers recommend striking a balance between these competing concerns when selected new images for PSE motive assessment (e.g. Veroff et al. 1960; Winter 1989, cited in Smith et al. 1992). Firstly, any image which did not demonstrate significant motive pull (i.e. fewer than 50 % of candidates wrote narratives containing at least one instance of achievement-related imagery, see Schultheiss and Brunstein 2001) for either HS or FF was discarded. Secondly, the mean number of instances of both HS- and FF-related imagery per thousand words (across all participants) was calculated for each of the remaining images. These indices of HS and FF pull were then consolidated into a third variable measuring ambiguity by taking the larger of the two numbers and dividing by the smaller. As a result, highly ambiguous images had an ambiguity value approaching one while unambiguous images with much higher pull in one category than the other gave rise to a larger number. Images were then ranked smallest to largest in terms of their ambiguity value, with the eight smallest (and most ambiguous) being retained. Of the eight retained images, the most ambiguous image had an ambiguity value of 1.01 (indicating that it pulls equally for HS and FF, whereas the least ambiguous of the retained images had an ambiguity value of 2.35, indicating that it was more than twice as potent pulling for FF than HS. The remaining images had ambiguity values that fell between these two extremes.

In the present research, these eight most ambiguous HS/FF images (*basketball, blueprints, forehead, lecture theater, operating theater, skaters, squash, student*) formed the unambiguous picture set, since all these images had been pretested to pull for a combination of HS and FF. The ambiguous picture set on the other hand, comprised the six most ambiguous HS/FF pictures (*basketball, blueprints, lecture theater, operating theater, skaters, student*) and two images which have previously been shown to pull for an alternate motive (Pang 2010b): *couple by river* (*n* Affiliation) and *hooligan attack* (*n* Power), which replaced the two least ambiguous (i.e. least effective) HS/FF images *forehead* and *squash*. As such, the ambiguous and unambiguous picture sets differed in only in terms of two images. Reproductions of all images can be found in Appendix 1.

HS and FF scores derived from a valid picture cue battery should correlate with those derived from previously-validated images—i.e. the novel picture set should exhibit convergent validity. To this end, a multi-motive 8-picture set of previously-validated images was assembled for the purpose of motive profile comparison. The picture cues *women in lab* and *ship captain* were originally used by McClelland and associates (reproduced from Smith 1992), *director* and *man at desk* are pictures B and E from Heckhausen's original six-image assessment battery (Schultheiss 2001), while *soccer duel* has been widely used in research (e.g. Schultheiss and Rohde 2002). Of the remaining four images, *chemist* and *gymnast* were utilized by Pang et al. (2009), while *piano lesson* has been pretested for its ability to arouse achievement imagery (Pang 2010a). Reproductions of these images can be found in Appendix 2.

Validation tasks

Wisconsin card sorting test (WCST)

An adapted version of the WCST was administered during session 3 in order to assess and compare the predictive validity of the ambiguous and unambiguous picture sets. Originally developed as a measure of executive function, we reasoned that the challenging and dynamic nature of the WCST, in which subjects are required to constantly adapt their strategy in the face of changing card-sorting criteria, makes it suitable for the arousal of the achievement motive and for the expression of achievement relevant behaviors. This is in line with Brunstein and Schmitt's (2004) argument that tests of concentration are well-suited to the assessment of motivationally-relevant behaviors, since effective performance (operationalized both in terms of accuracy and speed) requires a great deal of mental effort.

The present adaptation of the WCST comprised three blocks, each consisting of a maximum of 128 trials. Participants were instructed to sort the cards appearing at the top of the screen into one of four piles, and were informed that they would receive immediate trial-by-trial feedback as to whether they had made the right or wrong choice. They were also advised that their performance would be judged both on the speed and the accuracy of their responses, and that they would receive a cash payment that reflected how well they performed. The card stimuli could be classified according to three independent criteria—color (red, blue, yellow, green), form (circle, star, cross, triangle), and number (one, two, three, four)—although participants were not given any information regarding which of these sorting criteria was being applied during each trial. Participants were required to deduce the nature of these changing expectations from the feedback provided and to modify their responses accordingly. The WCST was programmed such that the sorting criterion being applied (e.g. color) remained for only four trials before switching to the next criterion (e.g. form, and subsequently, number). This four-trial cyclical re-definition of sorting rules continued until the participant had correctly identified six sorting criteria or all of the 128 cards were sorted, at which point the block would terminate. The first block was envisaged as a practice block to familiarize the participants with the task, while the second and third blocks were experimental blocks which yielded actual monetary payment. After the second block, participants were provided with a feedback summary of their performance and invited to try and improve their scores in the third and final block. The highest scoring participant earned $4.94 SGD (Singapore Dollars) over the course of the three experimental blocks, while the lowest scoring participant earned $0.88 SGD. Since self-referenced feedback has previously been shown to be highly effective in arousing implicit achievement motivation (Pang 2010b), responses from the third block provided the accuracy (percentage of correct trials) and speed (mean trial latency) data that were to be predicted by HS and FF.

Two criterion variables were derived from the adapted WCST. The percentage of correct responses over the course of the third block provided a measure of accuracy, while the mean latency in milliseconds (time elapsed between presentation of the card to be categorized and decision by participant) provided a measure of speed.

Balloon analogue risk task (BART)

In order to further assess the predictive validity of the ambiguous and unambiguous picture sets, participants also completed a version of the BART. The BART assesses risk-taking behavior, and task scores have been found to be associated both with real-world risk-taking behaviors such as smoking (Lejuez et al. 2002) and with self-report measures of impulsivity and illicit drug use (Lejuez et al. 2003).

In the BART, participants are required to inflate a series of simulated balloons in order to earn money. The balloon is inflated by pressing an on-screen “pump” button, and money is accumulated with each successive pump until the balloon bursts, at which point the participant loses all the money gained on that trial. Participants can however choose to “bank” the money at any time and proceed to the next balloon in the series. Thus, with every pump the participant must balance the potential gain of accruing more money with the potential risk of bursting the balloon and losing everything they had gained on that particular trial. As in the adapted WCST, participants were paid real money in accordance with how they performed over the course of the task. The highest scoring participant received $10.90 SGD while the lowest scoring participant received $1.45 SGD.

Three dependent measures were derived from the BART. Following the example of the Lejuez et al. (2003), the first risk-taking measure was the average number of pumps on balloons that do not burst, while the second dependent variable was the total number of burst balloons across the ten trials. In both cases higher scores indicate greater risk-taking behavior. The third variable derived from BART was mean trial latency, which provided a measure of performance speed in a time-independent task (unlike the WCST, in which performance was explicitly related to speed).

Measures of explicit need for achievement

The Personality Research Form (PRF; Jackson 1974) has been used extensively by researchers to investigate the relationship between implicit and explicit motives in the big three domains of achievement, affiliation, and power (e.g. Koestner et al. 1988). Additionally, the Hope of Success and Fear of Failure scale (Schultheiss and Murray 2002), which was developed to provide an explicit measure of HS and FF which was content-matched to categories present in the Heckhausen (1963) coding system, has recently been employed by several researchers investigating implicit/explicit motive congruence (e.g. Thrash et al. 2007). Both measures were administered immediately after completion of the comparison PSE in session 4.

Results

Test–retest reliability

Simple bivariate correlations were conducted in order to compare the test–retest reliabilities of the ambiguous and unambiguous picture sets (see Tables 1 and 2). HS and FF scores derived from the session one PSE administration were correlated with those derived from the session two PSE administration in order to assess their stability over time.

Session one HS scores were found to be significantly correlated with session two HS scores in both the ambiguous and unambiguous conditions, indicating that both

Table 1 Means, standard deviations, and inter-correlations of implicit and explicit motives in ambiguous condition

	S1 HS	S1 FF	S2 HS	S2 FF	S4 HS	S4 FF	saHS	saFF	sanAch
S1 HS									
S1 FF	.054								
S2 HS	.465**	.125							
S2 FF	−.150	.328*	.245						
S4 HS	.364*	.106	.778**	.322*					
S4 FF	−.167	.391*	−.071	.372*	.105				
saHS	−.143	.043	−.014	−.089	−.067	.088			
saFF	−.091	−.013	.119	−.028	.246	−.113	.184		
sanAch	.049	.124	−.001	−.147	−.152	−.066	−.104	.283	
Mean	12.99	6.27	13.77	7.19	17.79	6.00	3.52	3.28	9.49
SD	10.97	5.47	9.64	7.12	12.89	5.99	0.34	0.28	2.67

S1 session 1, *S2* session 2, *S4* session 4, *saHS* explicit HS, *saFF* explicit FF, *sanAch* explicit nAch

** $p < .01$ (2-tailed), * $p < .05$ (2-tailed)

Table 2 Means, standard deviations, and inter-correlations of implicit and explicit motives in unambiguous condition

	S1 HS	S1 FF	S2 HS	S2 FF	S4 HS	S4 FF	saHS	saFF	sanAch
S1 HS									
S1 FF	.205								
S2 HS	.390*	−.111							
S2 FF	−.128	.154	−.121						
S4 HS	.378*	−.036	.722**	−.079					
S4 FF	−.247	.001	−.209	.058	−.019				
saHS	.107	.091	.012	−.257	−.103	−.270			
saFF	.049	.067	.167	.224	.091	.300	−.004		
sanAch	−.327	−.170	−.091	.060	−.074	−.005	−.134	−.296	
Mean	22.99	10.57	22.69	11.03	23.76	6.24	3.54	3.26	0.60
SD	16.73	4.98	12.80	6.07	14.35	5.25	0.26	0.25	0.16

S1 session 1, *S2* session 2, *S4* session 4, *saHS* explicit HS, *saFF* explicit FF, *sanAch* explicit nAch

** $p < .01$ (2-tailed), * $p < .05$ (2-tailed)

ambiguous and unambiguous picture sets exhibit satisfactory test–retest reliability. However in the case of FF, a significant correlation between session one and session two motive scores was observed only in the ambiguous condition, and not in the unambiguous condition.

These results offer support for H1, suggesting that more ambiguous picture sets exhibit better test-reliability than less ambiguous picture sets. While the observed 1 week stability coefficients are only moderate, given that Schultheiss and Pang (2007) reported an average value of $r = .60$ in their meta-analysis of PSE studies, the present results suggest that test–retest reliability drops markedly when using unambiguous picture sets, particularly in the case of FF.

Convergent validity

In the order to compare the convergent validity of the ambiguous and unambiguous picture sets, motive scores from the session two PSE administration were correlated with motive scores from the session four PSE administration comprising previously-validated picture cues (see Tables 1 and 2). Echoing the test–retest reliability results above, the ambiguous picture set was found to exhibit greater convergent validity than the unambiguous picture set. Both ambiguous session two HS scores and FF scores were found to be significantly correlated with their session four counterparts. However, while session two HS scores in the unambiguous condition were significantly correlated with session four HS scores, the unambiguous session two and session four FF scores were uncorrelated. These results offer support for H2.

Predictive validity

WCST

Several hierarchical multiple linear regressions were conducted in order to assess the predictive validity of ambiguous and unambiguous HS and FF scores with respect to speed and accuracy in the WCST. In the first such regression analysis, post-feedback speed (mean response latency in ms for the third experimental block) was predicted by practice block speed and the average of the HS scores from session one and session two. This analysis allowed for the predictive power of HS with respect to speed in the post-feedback trials to be assessed while controlling for individual differences in task ability, as indicated by their performance on the practice trials. Practice trial latency (i.e. speed) was included as a predictor in the first step while the HS motive score was added in the second step. Practice trial latency was found to be a significant predictor of post-feedback trial latency ($\beta = .58$, $t(39) = 4.44$, $p < .001$), while ambiguous HS was also a significant predictor of post-feedback trial latency ($\beta = -.37$, $t(38) = -3.11$, $p < .01$). Identical regression analyses revealed that unambiguous HS ($\beta = -.06$, $t(30) = -.36$, $p = .72$) was not a significant predictor of post-feedback trial latency. H3 was therefore supported.

In the second round of regression analyses, post-feedback accuracy (percentage of correct responses in the third experimental block) was predicted by practice block accuracy and HS. Once again, this analytic approach allowed us to assess the predictive power of HS while controlling for baseline task ability, this time in terms of sorting accuracy, by including practice trial performance as the sole predictor in the first step before adding HS in the second step. Practice trial accuracy was found to be a significant predictor of post-feedback trial accuracy ($\beta = .37$, $t(39) = 2.46$, $p < .05$), while ambiguous HS was also a significant predictor of post-feedback trial accuracy ($\beta = -.36$, $t(38) = -2.54$, $p < .05$). Identical regression analyses revealed that unambiguous HS ($\beta = .12$, $t(30) = .76$, $p = .45$) did not significantly predict post-feedback trial accuracy, meaning that H4 was also supported.

BART

Linear regression analyses were conducted to assess the predictive validity of FF scores derived from the ambiguous and unambiguous picture sets in the context of the BART. Specifically, average session one and session two FF scores were used to predict two indices of risk-taking behavior—the total number of burst balloons and the average number of pumps on non-bursting trials—as well as task speed, operationalized as mean trial latency.

Neither of the risk-taking dependent variables were significantly predicted by ambiguous or unambiguous FF. Contrary to our prediction in H5, these results imply that fear of failure is unrelated to risk-taking behavior as measured by the BART. This may suggest that general risk-taking behaviors are unsuitable for use as outcome variables in FF predictive validity comparisons. Mean trial latency however was significantly predicted by ambiguous FF ($\beta = .32$, $t(36) = 2.03$, $p = .05$), while unambiguous FF ($\beta = -.03$, $t(27) = -.13$, $p = .90$) was a non-significant predictor. H6 was therefore supported by the data.

General discussion

Taken together, the above results suggest that set ambiguity plays an important role in determining the reliability and validity of the picture story exercise. Motive scores derived from an ambiguous picture set exceeded those derived from an unambiguous picture set in three key indicators of psychometric performance: test–retest reliability, convergent validity, and predictive validity. Although a number of important questions still need to be addressed (discussed in more detail below), these preliminary findings illustrate the importance of taking a holistic approach to picture cue selection, by attending not only to the properties of individual cues but also to the composition of the overall picture set. Critically, they suggest that all PSE picture sets, whether constructed for the purpose of single-motive or multi-motive assessment, should be varied in terms of the individual picture cues' motivational focus. While implicit motive researchers have long espoused the benefits of cue variability when studying multiple motives (e.g. Pang 2010a), the prevailing wisdom has been that single-motive assessment is best conducted using uniform picture sets consisting entirely of images which pull preferentially for the motive under investigation (e.g. Smith et al. 1992). The present research suggests that this is not the case.

Inclusion of two cues which pull preferentially for an alternative motive alongside a majority of cues which pull for the target motive afforded substantial improvement in reliability and validity. Motive scores derived from the ambiguous picture set were more stable over time, more highly-correlated with scores from previously-validated images, and most importantly, more predictive of theoretically-relevant behaviors than scores derived from the unambiguous picture set. Specifically, ambiguous picture set HS predicted the tendency to respond faster but less accurately in a time-dependent task (the WCST), demonstrating the HS motivated individual's previously-documented "greater tolerance for frustrations" (Schultheiss and Brunstein 2005, p 11) when pursuing an achievement goal. Ambiguous picture set FF on the other hand, predicted the tendency for individuals to work more slowly in a time-independent task (the BART), an observation which rings true given Heckhausen's (1980) observation that FF-motivated individuals work more slowly and cautiously on achievement tasks when given the opportunity to do. Scores derived from the unambiguous picture set did not significantly predict any of these behaviors. Given the improvements in reliability and validity of the HS and FF scores of the ambiguous picture set relative to the unambiguous picture set, we therefore recommend that researchers construct ambiguous picture sets when conducting PSE research.

This is not to say that set ambiguity is the only or even the primary consideration when constructing a PSE battery. Research has repeatedly demonstrated the importance of selecting picture cues which possess significant pull for the motive in question, and the need to consider the cue strength of the various component images is well-documented. Rather, the present results speak of a need to balance the competing concerns of pull and ambiguity, not only at the level of the individual picture, but also at the level of the picture battery. Commonly used multi-motive picture sets have been found to pull too weakly for valid assessment of a single target motive (Ramsay 2011), while the present research suggests that unambiguous picture sets that focus solely on the target motive also suffer from decreased validity. Just as individual cues should pull moderately so that resulting score variance is neither too high nor too low, so too should picture sets exhibit moderate ambiguity if validity is to be ensured.

Further empirical work should be conducted in order to replicate and extend these preliminary findings. Several areas warrant particular attention. Firstly, researchers should replicate the above findings in context of other content coding systems with varying motivational foci, such as McClelland et al. (1953), McAdams (1980), Winter (1973), and Winter (1994). While it seems likely that the principle of set ambiguity should apply equally to these alternate systems and motives, empirical confirmation of these suspicions is required. Secondly, research is needed to identify the optimal ratio of primary to alternate motive pull images for single-motive assessment batteries. While the present research demonstrates the advantage of including two alternate pull cues in an eight picture set, the possibility remains that including more alternate pull cues—i.e. increasing set ambiguity still further—offers even greater improvements in reliability and validity. A study in which ambiguity is varied incrementally, conducted in a manner similar to Schultheiss and Pang's (2007) investigation of picture set size, would shed light on this issue. Furthermore, we cannot currently be certain that including alternate motive pull images is a strict requirement for ensuring reliability and validity in single-motive PSE assessment, since including a number of neutral pictures with low pull for *all* motives may also suffice. A picture set comprising both target motive pull images and neutral, low pull images would also exhibit set ambiguity, and future research should attempt to establish whether neutral images can serve the same purpose as the alternate motive pull images used here. Finally, further work is needed if we are to explain exactly why set ambiguity improves the psychometric properties of the PSE. Although we tentatively suggested in our introduction that unambiguous picture sets might exert undesirable effects due to demand characteristics, no evidence has been gathered that directly supports this contention. No participants indicated their awareness of the experimental hypothesis in their responses to funneled debriefing questions in the present study, and unambiguous HS and FF were not found to be significantly correlated with explicit HS and FF respectively, although correlations between average HS and FF and their explicit counterparts were more positive in the unambiguous condition. Future research should explore this and other possible explanations more fully.

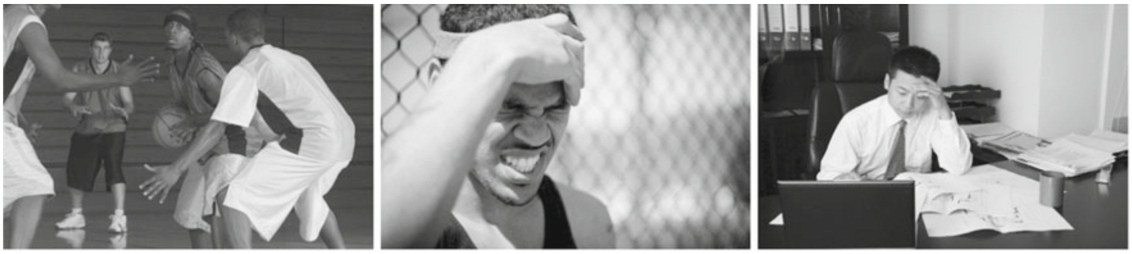
Limitations of the present research should also be noted. One of our hypotheses with regard to the enhanced predictive validity of the ambiguous picture set was not borne out by the data: the hypothesized relationship between FF and risk-taking in the BART. Since this predictive relationship was found to be non-significant in both the ambiguous and the unambiguous conditions, the possibility remains that this behavioral outcome is not related to FF in the manner we suggested, meaning that it is unsuitable for comparing the predictive validity of different picture sets. Without further experimentation, it is impossible to know whether this null result speaks against the importance of set ambiguity, or whether it is merely a result of sub-optimal task selection. While a clearer picture of Heckhausens's FF has begun to emerge, e.g. Pang et al.'s (2009) findings regarding memory for successful versus unsuccessful peers, more work is needed to identify behavioral correlates of fear of failure. The small sample size represents a further limitation of the present research, and the possibility remains that effects were not detected due to a lack of statistical power. Future replications and extensions should attempt to offer improvement in this regard, replicating these findings using other behavioral correlates of both HS and FF as outcome variables.

Conclusion

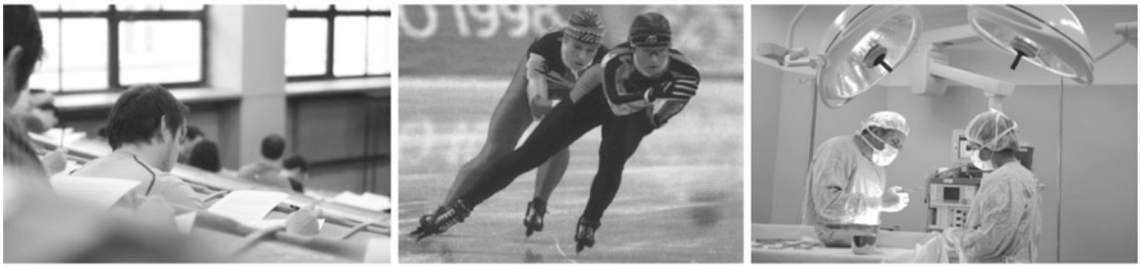
Set ambiguity, a previously unstudied attribute, is an important determinant of reliability and validity in the picture story exercise. In the present study, HS and FF motive scores derived from an ambiguous picture set containing two non-achievement images were found to exhibit greater test–retest reliability, convergent validity, and predictive validity than those derived from an unambiguous achievement picture set. Researchers are therefore advised to use or construct ambiguous picture sets when assessing implicit motives using the PSE.

Appendix 1

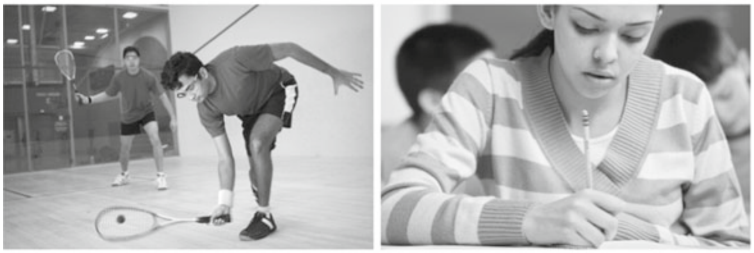
L-R: Basketball, Forehead, Blueprints



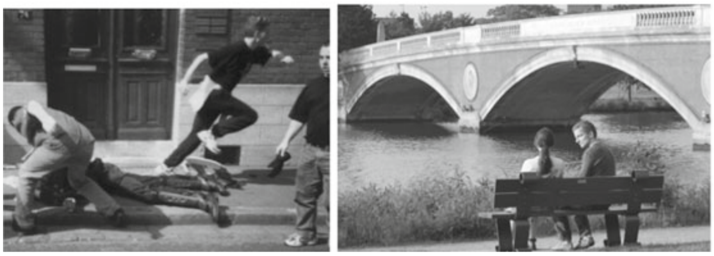
L-R: Lecture Theater, Skaters, Operating Theater



L-R: Squash, Student

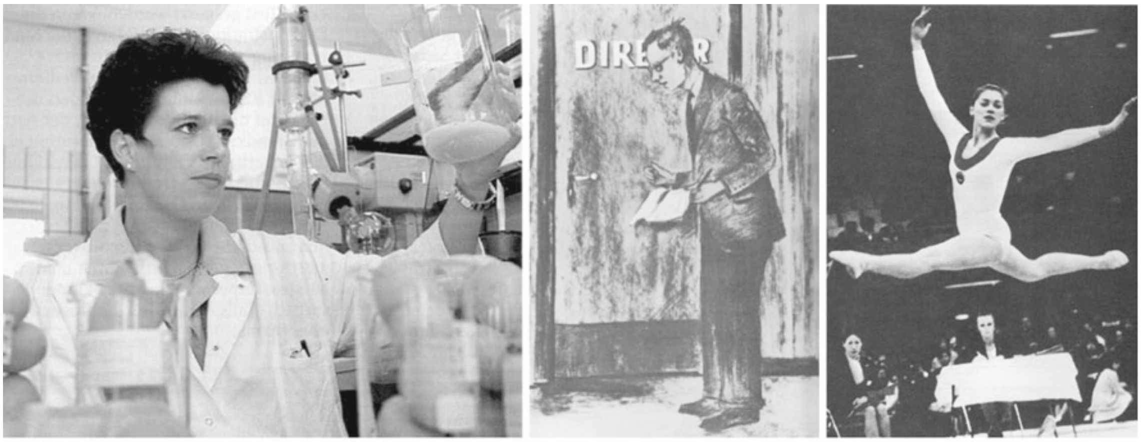


L-R: Hooligan Attack, Couple by River

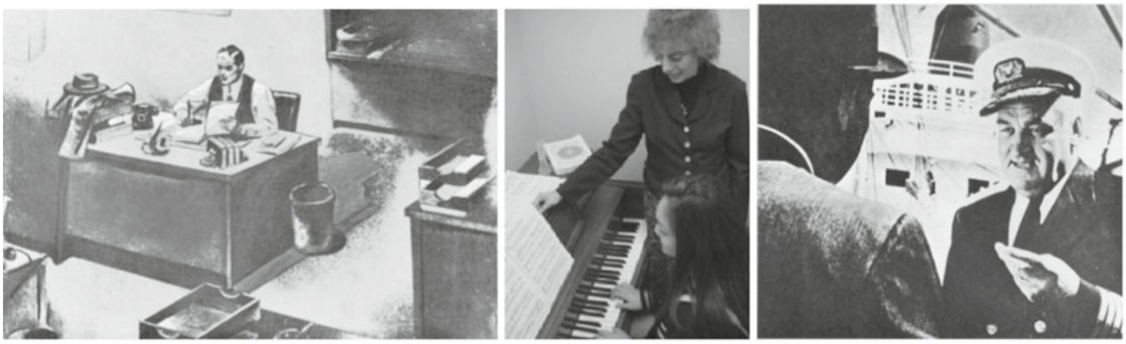


Appendix 2

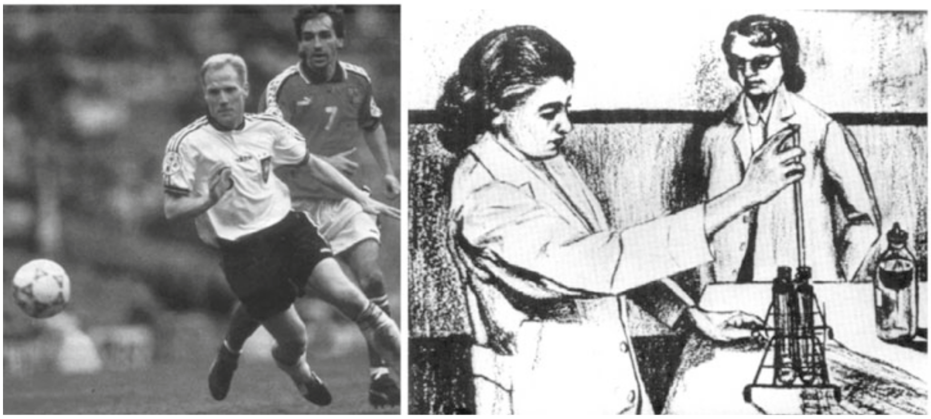
L-R: Chemist, Director, Gymnast



L-R: Man at Desk, Piano Lesson, Ship Captain



L-R: Soccer Duel, Women in Lab



References

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

Atkinson, J. W. (1992). Motivational determinants of thematic apperception. In C. P. Smith (Ed.), *Motivation and personality: handbook of thematic content analysis* (pp. 21–48). New York, NY: Cambridge University Press.

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

Brunstein, J. C., & Schmitt, C. H. (2004). Assessing individual differences in achievement motivation with the Implicit Association Test. *Journal of Research in Personality, 38*(6), 536–555. doi:10.1016/j.jrp.2004.01.003.

de Charms, R., & Carpenter, V. (1968). Measuring motivation in culturally disadvantaged children. In H. Klausmeirer & G.

O'Hearn (Eds.), *Research and development toward the improvement of education*. Madison, WI: Educational Research Services.

Eron, L. D. (1950). A normative study of the Thematic Apperception Test. *Psychological Monographs: General and Applied, 64*(9), 1–48. doi:10.1037/h0093627.

Eron, L. D. (1953). Responses of women to the Thematic Apperception Test. *Journal of Consulting Psychology, 17*(4), 269–282. doi:10.1037/h006282.

Grant, D. A., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology, 38*(4), 404. doi:10.1037/h0059831.

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, NY: Van Nostrand.

Heckhausen, H. (1963). *Hoffnung und furcht in der leistungsmotivation*. Meisenheim, Germany: Anton Hain.

Heckhausen, H. (1968). Achievement motive research: Current problems and some contributions towards a general theory of motivation. In W. J. Arnold (Ed.), *Nebraska symposium on motivation* (Vol. 16). Lincoln, NE: University of Nebraska Press.

Heckhausen, H. (1980). *Motivation und handeln*. New York, NY: Springer.

Jackson, D. (1974). *Personality research form manual*. Goshen, NY: Research Psychologists Press.

Koestner, R., Weinberger, D. R., McClelland, D. C., & Healy, J. (1988). *How motives and values interact with task and social incentives to affect performance*. Unpublished manuscript, Department of Psychology, Boston University. Boston.

Lejuez, C., Aklin, W. M., Jones, H. A., Richards, J. B., Strong, D. R., Kahler, C. W., et al. (2003). The balloon analogue risk task (BART) differentiates smokers and nonsmokers. *Experimental and Clinical Psychopharmacology, 11*(1), 26. doi:10.1037/1064-1297.11.1.26.

Lejuez, C., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., et al. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied, 8*(2), 75. doi:10.1037/1076-898X.8.2.75.

McAdams, D. P. (1980). A thematic coding system for the intimacy motive. *Journal of Research in Personality, 14*(4), 413–432.

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

McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). *The achievement motive*. New York, NY: 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. doi:10.1037/0033-295X.96.4.690.

McClelland, D. C., & Pilon, D. A. (1983). Sources of adult motives in patterns of parent behavior in early childhood. *Journal of Personality and Social Psychology, 44*(3), 564. doi:10.1037/0022-3514.44.3.564.

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

Murray, H. A. (1938). *Explorations in personality: A clinical and experimental study of fifty men of college age*. New York, NY: Oxford University Press.

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.

Pang, J. S. (2010a). Content coding methods in implicit motive assessment: Standards of measurement and best practices for the Picture Story Exercise. In O. Schultheiss & J. Brunstein (Eds.), *Implicit motives* (Vol. 1, pp. 119–151). New York, NY: Oxford University Press.

Pang, J. S. (2010b). The achievement motive: A review of theory and assessment of *n* achievement, hope of success, and fear of failure. In O. Schultheiss & J. Brunstein (Eds.), *Implicit motives* (Vol. 1, pp. 30–71). New York, NY: Oxford University Press.

Pang, J. S., & Schultheiss, O. C. (2005). Assessing implicit motives in US college students: Effects of picture type and position, gender and ethnicity, and cross-cultural comparisons. *Journal of Personality Assessment, 85*(3), 280–294.

Pang, J. S., Villacorta, M. A., Chin, Y. S., & Morrison, F. J. (2009). Achievement motivation in the social context: Implicit and explicit hope of success and fear of failure predict memory for and liking of successful and unsuccessful peers. *Journal of Research in Personality, 43*(6), 1040–1052.

Ramsay, J. E. (2011). *Refining the picture story exercise: Towards a better understanding of hope, fear and the achievement motive*. Unpublished manuscript, Nanyang Technological University, Division of Psychology, Singapore.

Reitman, W. R., & Atkinson, J. W. (1958). Some methodological problems in the use of thematic apperceptive measures of human motives. In J. W. Atkinson (Ed.), *Motives in fantasy, action, and society* (pp. 664–683). Princeton, NJ: Van Nostrand.

Schultheiss, O. (2001). *Manual for the assessment of hope of success and fear of failure (English translation of Heckhausen's need Achievement measure)*. Unpublished manuscript, University of Michigan. Ann Arbor, MI.

Schultheiss, O., & Brunstein, J. (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. doi:10.1207/S15327752JPA7701_05.

Schultheiss, O. C., & Brunstein, J. C. (2005). An implicit motive perspective on competence. In A. J. Elliot & C. Dweck (Eds.), *Handbook of competence and motivation* (pp. 31–51). New York: Guilford.

Schultheiss, O., & Murray, T. (2002). *Hope of success/fear of failure questionnaire*. Unpublished manuscript, University of Michigan. Ann Arbor, MI.

Schultheiss, O., & Pang, J. (2007). Measuring implicit motives. In R. Robins, R. Fraley, & R. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 322–344). New York: The Guildford Press.

Schultheiss, O. C., & Rohde, W. (2002). Implicit power motivation predicts men's testosterone changes and implicit learning in a contest situation. *Hormones and Behavior, 41*(2), 195–202. doi:10.1006/hbeh.2001.1745.

Smith, C. P. (1992). *Motivation and personality: Handbook of thematic content analysis*. New York, NY: Cambridge University Press.

Smith, C. P., Feld, S., & Franz, C. (1992). Methodological considerations: Steps in research employing content analysis systems. In C. P. Smith (Ed.), *Motivation and personality: Handbook of thematic content analysis* (pp. 515–536). New York, NY: Cambridge University Press.

Thrash, T. M., Elliot, A. J., & Schultheiss, O. C. (2007). Methodological and dispositional predictors of congruence between implicit and explicit need for achievement. *Personality and Social Psychology Bulletin, 33*(7), 961–974. doi:10.1177/0146167207301018.

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*(12), 1.

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

Winter, D. G. (1989). *Technical note on measuring motivation*. Unpublished manuscript, University of Michigan, Department of Psychology. Ann Arbor, MI.

Winter, D.G. (1994). *Manual for scoring motive imagery in running text*. Unpublished manuscript, University of Michigan, Department of Psychology. Ann Arbor, MI.