Journal of Research in Personality 42 (2008) 1560–1571

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

The reliability of a Picture Story Exercise measure of implicit motives: Estimates of internal consistency, retest reliability, and ipsative stability

Oliver C. Schultheiss [a,*], Scott H. Liening [b], Daniel Schad [c]

[a] Friedrich-Alexander University, Department of Psychology and Sport Sciences, Kochstrasse 4, 91054 Erlangen, Germany
[b] The University of Texas at Austin, Department of Psychology, 1 University Station A8000, Austin, Texas 78712-0187, USA
[c] Institute for Psychology, Karl-Liebknecht-Str. 24/25, 14476 Potsdam OT Golm, Germany

ARTICLE INFO

Article history:
Available online 29 July 2008

Keywords:
Test reliability
Picture story exercise
Implicit motives
Need for power
Need for achievement
Need for affiliation
Content coding
Computer-aided assessment
Stability
Test theory
Longitudinal assessment
Learning theory
Verbal fluency
Personality theory
Thematic apperception

ABSTRACT

Ninety participants completed an 8-picture picture story exercise (PSE) on two occasions spaced 2 weeks apart. Writing condition (handwrite versus type) was varied experimentally at each occasion. Stories were scored using Winter's [Winter, D. G. (1994). *Manual for scoring motive imagery in running text* (4th ed.). Department of Psychology, University of Michigan, Ann Arbor, unpublished manuscript] integrated coding system for the assessment of motive imagery. Typed stories were longer by a third than handwritten stories. Word count-corrected motive scores from typed stories shared stable variance with and did not significantly differ from motive scores obtained from handwritten stories. Retest stability for word count-corrected power, achievement, and affiliation scores was .39, .37. and .61, respectively. Across all three motive measures, inter-scorer reliability was high (.70–.86), internal consistency as estimated by Cronbach's α was low (−.02 to .43), and motive scores showed substantial ipsative stability as assessed by averaged intraindividual correlations of picture profiles across testing occasions (.21–.40). These findings are consistent with Mischel and Shoda's [Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*, 246–268] conceptualization of personality dispositions as stable *if…then* contingencies between situational cues and behavioral responses.

1. Introduction

Research on implicit motives, nonconscious dispositions to experience certain types of incentives as pleasurable, has made considerable strides in recent years, leading to new insights into their biological basis (e.g., McClelland, 1989; Schultheiss, 2007; Schultheiss, in press), and their effects on cognitive processes (e.g., Woike, Lavezzary, & Barsky, 2001), affective consequences (e.g., Fodor, Wick, & Hartsen, 2006), and socio-historical manifestations (e.g., Winter, 1993). Yet, our understanding of the Picture Story Exercise (PSE), the chief instrument for the assessment of implicit motives for over 50 years, appears to be lagging behind. Only very recently researchers have attempted to systematically describe frequently used PSE picture cues in terms of their ability to elicit specific types of scorable motive imagery (Blankenship et al., 2006; Langan-Fox & Grant, 2006; Pang & Schultheiss, 2005; Schultheiss & Brunstein, 2001). And to this day, the question of if and why the PSE can have validity despite its low internal consistency remains a point of contention between advocates and critics of the measure.

* Corresponding author. Fax: +49 9131 8526770.
E-mail address: oliver.schultheiss@psy.phil.uni-erlangen.de (O.C. Schultheiss).

0092-6566/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jrp.2008.07.008

In their pioneering work on the achievement motive, McClelland, Atkinson, Clark, and Lowell (1953) were the first to point out that the PSE, which was derived from Murray's (1943) Thematic Apperception Test, has low reliability in the sense that scores obtained in stories that were written in response to one picture did not correlate very well with scores derived from stories written in response to other pictures. This finding emerged despite the fact that independent coders could typically agree on the presence or absence of motive imagery in a given story in 85% of the cases or more. Subsequently, Entwisle (1972) blasted PSE motive measures, arguing that a measure with low internal consistency is unlikely to predict anything validly according to the laws of classical test theory (cf. Nunnally, 1967). Similar points were raised by Fineman (1977) and Lilienfeld, Wood, and Garb (2000). Notably, however, PSE motive measures have substantial retest reliability, as demonstrated in a recent meta-analysis (Schultheiss & Pang, 2007). Across studies, the average correlation for 1-week retest intervals was .60, for one month it was .52, and after 1 year it was still .37. So how can the PSE be valid, have high inter-scorer and substantial retest reliability and yet have low internal consistency at the same time?

1.1. A classic account of low internal consistency: Dynamics of action theory

The first and, until now, only constructive attempt to solve this riddle was provided by Atkinson (1981) who argued that low consistency of motive imagery scores across pictures is a function of the cyclical arousal and satiation of motivational needs through the story-writing process. Originally, this argument was based on two lines of evidence. First, in a study described in McClelland et al.'s (1953) original book on the achievement motive, participants produced a saw-tooth pattern of motive scores across pictures cueing achievement. Atkinson (1950; see McClelland, 1980) interpreted this finding as evidence that in the presence of a stable environment, represented by a series of pictures all related to achievement, motivation oscillates between arousal (high motive imagery) and satiation (low motive imagery). After Atkinson and Birch (1970) had formalized and elaborated these assumptions into the mathematical models that constituted their dynamics of action (DOA) theory, Atkinson and colleagues used the theory to run simulations of the ebb and flow of motivation in an environment with constant incentive cues (e.g., Atkinson, Bongort, & Price, 1977). These studies demonstrated that the interplay between a stable motive disposition in the writer and stable incentive cues represented in PSE pictures could give rise to variable motive expression from one story to the next. The simulation studies thus represented the second line of evidence in favor of the PSE picking up a cyclical motivational process that, by its very nature, could not produce constant or consistent response from one moment to the next. Note, however, that these studies were computer simulations, not actual empirical studies that would have independently validated the assumptions of DOA.

Reuman's (1982) aimed to provide this empirical support for the DOA theory. He predicted from the dynamics of action theory that higher score variability across PSE pictures should signify that the motive could be assessed more precisely, because several arousal-satisfaction cycles could be observed within one PSE. In contrast, low score variability was assumed to make an exact diagnosis of motive strength more difficult, because the actual length of one arousal-satisfaction cycle exceeded the length of the test. This led Reuman to predict that on a PSE designed to assess achievement motivation, high intraindividual score variability and the resulting low internal consistency should reflect high oscillation and thus be associated with better test validity, whereas low intraindividual score variance and the resulting high internal consistency should reflect low oscillation and thus be associated with low test validity. Empirical data from a study with 61 male students were in agreement with these predictions. In a subsample of individuals with high ipsative (i.e., within-subjects) variability of achievement motive scores across 4 picture stories (Cronbach's $\alpha = -1.23$), the achievement motive correlated at .62 with a criterion (attempted solutions on arithmetic problems). In contrast, in a subsample of individuals with low ipsative score variability (Cronbach's $\alpha = .42$), the achievement motive correlated only at .20 with the criterion.

1.2. Problems with the DOA account of low consistency

Although the basic tenet of DOA—that observable motivated behavior can change dynamically as a function of stable dispositional needs and environmental factors—is consistent with biopsychological and ethological accounts of motivation (cf. Ball & Balthazart, 2008) and other advanced measurement models in personality psychology (e.g., Nesselroade & Boker, 1994), we contend that it is not suitable to explain the low internal consistency of PSE motive measures.

There are two basic assumptions behind the interpretation of both the saw-tooth effect observed by Atkinson (1950) and Reuman's (1982) findings that are particularly problematic. First, the PSE pictures used in these authors' studies were assumed to represent a constant environment, that is, to have *comparable incentive value* for a given motive and a given person.[1] Second, DOA theory assumes that motive expression in response to one picture cue leads to *consummation of the motivational need* and motive expression will therefore become less likely on subsequent pictures.

Both Atkinson (1950) and Reuman (1982) considered the comparable-incentive-value assumption as met because the pictures they used generated similar average achievement motivation scores. However, it is problematic to equate similar motive scores at the sample level with similar incentive value at the individual level. Let us illustrate this point with an example: just because vanilla ice cream and chocolate ice cream are sold in similar quantities in grocery stores does not indi-

[1] As a reviewer pointed out, individual PSE cues can have varying pull for respondents from different cultural or educational background (see Hofer & Chasiotis, 2004).

cate that every customer likes both flavors equally and buys equal quantities of each; only that across customers stable individual differences cancel out, thus yielding similar average sales. The same could hold for motive scores: the PSE pictures Women in Laboratory and Trapeze Artists both yield high achievement motive scores on average (Pang & Schultheiss, 2005). But person A may write stories saturated with achievement imagery about the former picture, but not about the latter, because she has learned to associate situations resembling the one depicted in Women in Laboratory with the expression and satisfaction of her need for achievement, but not situations resembling the one depicted in Trapeze Artists. Person B's learning history may have taken a different turn, making him more likely to express achievement in situations similar to Trapeze Artists, but not in situations like the one shown in Women in Laboratory (see Brownell & Goss, 1957; Goss & Brownell, 1957, for related arguments). For this reason, it would be a mistake to attribute either person's inconsistent responses to these picture cues a priori as evidence for dynamic changes between motive arousal and satisfaction.

The assumption of a consummatory effect of motive expression in PSE story writing is problematic on conceptual and empirical grounds. The expression of a motivational need does not necessarily lead to the reduction of that need, as older drive-reduction accounts of motivation would have it. Modern incentive theories of motivation argue, based on ample empirical evidence, that incentive consummation can also lead to an *increase* in motivation, particularly if incentive consummation is not profound (like eating a full meal) but light (like eating a salted peanut; cf. Berridge, 2001). The axiomatic assumption that motive expression in PSE stories has consummatory value and leads to a form of drive reduction is therefore problematic to begin with.

In line with these arguments, there is no direct evidence for a satiation effect of motive expression on the PSE. Tuerlinckx, De Boeck, and Lens (2002) tested various models of thematic apperception for their fit with PSE motive scores obtained from a very large sample of participants. Among them was a DOA model in which the probability of expressed motive imagery in a given picture depended on whether it was already expressed on the previous picture. This model did not fit the data well, because there was no evidence of dynamic effects of the expression of motivational imagery in response to a given picture on the expression of imagery in response to subsequent pictures. In contrast, models that conceived of the expression of motive imagery in response to a picture as a joint function of the person's motivational need and the specific incentive value of the picture fit the data better and more parsimoniously.

1.3. An alternative account of low internal consistency: Cognitive-affective system theory

So if DOA theory is not suited to explain the conundrum of the PSE's low internal consistency, substantial retest stability, and high criterion validity, is there a better alternative? We believe that Mischel and Shoda's (1995) cognitive-affective system theory (CAST) of personality provides an excellent foundation for understanding the characteristics of PSE motive measures. According to CAST, people stably construe some situations, but not others, as opportunities to express a specific disposition. They thus manifest their personality in stable profiles of *if…then* contingencies of situational cues and behavioral responses. In support of this prediction, Shoda, Mischel, and Wright (1994) found that children who attended a summer camp did not show substantial consistency of aggressive behavior between different types of situations (e.g., if approached by peers, if reprimanded by teachers). However, they showed substantial stability in their behavior *within a given situation* from one occasion to the next. Thus, one child was more likely to aggress whenever he was approached by peers, but rarely showed signs of aggression if reprimanded by a teacher. Another child showed the opposite pattern. And although both children thus expressed their personality in stable situation-behavior contingencies across different occasions, neither child's personality was consistently expressed in the same way across different situations.

In the present research, we proceed on the assumption that stable situation-behavior contingencies as conceived of in CAST also drive individuals' story-writing responses to PSE picture cues. As sketched out in our previous critique of the assumptions of DOA theory, we propose that intraindividual variability in motive imagery produced in response to PSE picture cues reveals which types of situations a person has learned to perceive as ripe with motivational incentives and which not. If this assumption is true, then consistency of motive imagery scores should not be expected from one picture cue to the next, but for a given picture cue from one testing occasion to the next. Note that stable intraindividual picture-response profiles would not only be consistent with (and indeed almost require) low PSE reliability as estimated through internal consistency coefficients, but also accommodate the observation that motive imagery scores aggregated across all PSE picture stories show substantial retest stability. In this case, retest stability reflects, in part, stability in the overall number of different situations (i.e., picture cues) that a person interprets as opportunities for motive expression.

1.4. Overview of present study

To test the hypothesis that stable *if…then* contingencies between-picture cues and expression of motive imagery in story writing is behind the PSE's low internal consistency and high validity, we conducted a study in which we tested participants twice with the same 8-picture PSE, with testing occasions spaced two weeks apart. As evidence for stable *if…then* contingencies, we expected, from the first to the second testing occasion, significant and positive ipsative correlations of participants' power, achievement, and affiliation motive scores above and beyond sample-level differences in motive scores between pictures. Based on earlier findings by Shoda et al. (1994) on the ipsative stability of personality dispositions, we expected the size of the ipsative correlations to be in the medium range (i.e., around .30; cf. Cohen, 1992). We also examined other frequently used measures of PSE reliability: inter-scorer reliability, which in line with many previous studies we ex-

pected to be high; sum-score retest reliability, which based on Schultheiss and Pang's (2007) meta-analysis we expected to be in the range between .50 and .60; and internal consistency, which based on numerous previous studies we expected to be in the .00–.40 range (Cronbach's α). In addition, because previous studies suggest that the property of a given PSE picture cue to elicit specific types of motive imagery is remarkably stable from one study to the next (cf. Langan-Fox & Grant, 2006; Pang & Schultheiss, 2005; Schultheiss & Brunstein, 2001), we also took this opportunity to assess at the sample level the stability of between-picture differences in motive imagery (e.g., whether a picture that elicits high levels of power imagery across all participants at the first occasion will also do so at the second occasion).

Finally, to examine a question with considerable pragmatic relevance to motive researchers, we explored effects of writing PSE stories by hand versus typing them directly into the computer. Findings from a pioneering study by Blankenship and Zoota (1998) suggested that power motivation scores obtained with a traditional handwritten PSE do not differ from scores obtained through stories written on the computer keyboard. However, because writing condition was varied between-subjects in that study, it remained unclear whether stories handwritten or typed by the *same* participants would correlate. Substantial correlation across writing conditions would suggest that in a given participant motivational needs are reliably expressed in PSE stories, regardless of the mode of story writing. A low correlation, on the other hand, would suggest that switching from handwriting to typing somehow alters motive expression, potentially rendering motive scores obtained with the new method less valid (cf. Lundy, 1988). To examine this issue, we varied writing condition orthogonally across testing occasions. This yielded two groups in which writing condition was unchanged from the first to the second occasion (handwrite–handwrite, type–type) and two in which the writing condition changed (handwrite–type, type–handwrite). If the mode of story writing influences motive expression, retest correlations in the unchanged conditions should be different from those in the changed conditions. Conversely, if the mode of story writing does not bias motive expression, retest correlations should be similar between conditions.

2. Method

2.1. Participants

One-hundred and six students enrolled at the University of Michigan, Ann Arbor, participated in a study with two data collection sessions spaced 14 days apart. Participants signed up in response to fliers advertising the research as a study on "attention and performance" posted in multiple locations on campus. Of the initial pool of participants, four did not show up for the second data collection session and 12 were administered an incorrect PSE version due to a programming error. After these participants were dropped from the sample, the final data set available for all analyses was based on 90 participants (63 women, 24 men, 3 did not indicate their gender) with a mean age of 20 years. Fifty-seven percent of participants in the final sample self-identified as Caucasian, 29% as Asian, 6% as African-American, 3% as Pacific Islander; the remainder belonged to other or mixed ethnic groups.

2.2. Design and procedure

The study's design was a 2×2 factorial, with the first factor determining whether participants handwrote or typed their PSE stories at the first testing occasion (T1) and the second determining whether they handwrote or typed their PSE stories at the second occasion (T2). Participants were randomly assigned to the four conditions resulting from the 2×2 factorial design. Testing sessions (T1 and T2) were spaced exactly 14 days apart and for each participant time of day of the testing session was matched on both occasions. At both sessions, participants completed a test battery starting with an assessment of their implicit motives with the PSE, followed by various cognitive performance tasks and the collection of saliva samples for later hormone analyses. Findings for the cognitive tasks and hormone measurements will be reported elsewhere. All tasks and instructions were presented via personal computers and interaction with the experimenter was therefore limited to the initial greeting and getting participants set up at the data collection computers.

2.3. PSE administration

Participants worked on an 8-picture PSE following standard instructions for computer administration described in Schultheiss and Pang (2007). The PSE was programmed in Inquisit 2.0 (Millisecond Software, Seattle, WA). The following pictures were used in this study: Women in Laboratory, Ship Captain, Nightclub Scene, Couple by River, Trapeze Artists, Girlfriends in Café with Male Approaching, Bicycle Race, and Boxer. These pictures have been used in previous research on implicit motives and their cue properties and original source are described in Schultheiss and Pang (2007). Picture order was randomized by the software for each participant at each testing session. Each picture was shown on a black background for 10 s and then was replaced by a screen with writing instructions. Participants in the handwriting condition were instructed to write their stories on eight numbered writing sheets lying next to the computer keyboard. These sheets also contained in the upper left-hand corner the standard guiding questions for the collection of PSE stories (cf. Schultheiss & Pang, 2007). Participants in the typing condition were instructed to type their stories directly into a window on the screen, with the guiding questions appearing above the writing window. After 4 min had elapsed, a text appeared in the lower half of the screen

instructing participants to finish the story and move on to the next picture, along with instructions to hit "CTRL + Enter" when they were ready to proceed.

When participants took the PSE at the second session, the following paragraph was added to the instructions:

> "You may remember seeing 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 you told last time."

This paragraph is based on instructions used by Lundy (1985) and Winter and Stewart (1977) and was added to ensure that participants would not feel pressed to invent completely different stories this time, an effect that can artificially lower the retest stability of PSE motive scores (cf. Lundy, 1985; Winter & Stewart, 1977).

2.4. PSE scoring

After the end of data collection, protocol length of handwritten stories was determined by manual word count, a highly reliable method (cf. Schultheiss & Rohde, 2002). Protocol length of typed stories was determined through a utility programmed in Matlab 7.0 (MathWorks, Natick, MA). Stories were later coded for motivational imagery by two trained scorer using Winter's (1994) *Manual for Scoring Motive Imagery in Running Text*. According to the manual, imagery indicative of a need for power (or *n* Power) is scored whenever a story character shows a concern with having impact on others through strong, forceful actions, controlling, influencing, helping, impressing or eliciting emotions in others. *n* Achievement imagery is scored whenever a character shows a concern with a standard of excellence, as indicated by positive evaluations of goals and performances, winning or competing with others, disappointment about failure, or unique accomplishments. *n* Affiliation imagery is scored whenever a story character shows a concern with establishing, maintaining or restoring friendly relations, as expressed by positive feelings toward others, sadness about separation, affiliative activities, or friendly, nurturing acts.

The scorers had previously exceeded 85% inter-scorer agreement on calibration materials prescored by an expert that are contained in the manual. The scorers independently scored all handwritten and typed PSEs for motive imagery separately for each assessment, in random participant order, and without knowledge of a given participant's motive score at the other testing session. Scores on handwritten PSEs were later entered into data spreadsheets on the computer. Scores were assigned to typed PSE stories by entering them directly into the text documents. These were then extracted by a utility programmed in Matlab 7.0 that automatically wrote each participant's motive scores into a data file. Scores for each motive were averaged per picture across scorers for all further analyses and aggregations.

3. Results

3.1. Effects of condition on PSE scores

Writing condition (handwrite versus type) had significant effects on PSE protocol length both at T1 ($M = 677$ vs. $M = 962$ words, $t(88) = 5.57$, $p < .000005$) and at T2 ($M = 706$ vs. $M = 960$ words, $t(88) = 5.14$, $p < .000005$); PSE stories written in the type condition contained about 36% more words than those written in the handwrite condition. This effect was also significant within-subjects in the two conditions in which participants switched writing condition from Time 1 to Time 2, whereas little change in protocol length occurred in groups which were assigned to the same writing condition at both assessments (cf. Table 1). Protocol length was positively correlated across assessments both for

Table 1
Effects of experimental conditions (Time 1, Handwrite vs. Type; Time 2, Handwrite vs. Type) on mean (SD) raw motive scores and word count at Time 1 (T1) and Time 2 (T2) and on retest stability (Pearson correlations) of motive scores (residualized scores) and word count

		Handwrite T2			Type T2		
		T1	T2	r_{T1T2}	T1	T2	r_{T1T2}
Handwrite T1	*n* Power	4.38 (2.44)	5.23 (3.03)	.27	4.94 (2.38)**	6.44 (2.89)	.57***
	n Achievement	6.44 (2.16)	6.13 (2.98)	.14	5.96 (2.10)*	7.13 (3.21)	.56***
	n Affiliation	7.25 (2.67)	6.71 (3.82)	.59***	7.21 (2.81)	8.46 (4.13)	.62***
	Word count	631 (113)*	673 (125)	.68***	723 (140)***	987 (203)	.57***
			n = 24			*n* = 24	
Type T1	*n* Power	6.48 (4.27)	5.07 (2.46)	.35	5.83 (3.08)	6.17 (3.74)	.27
	n Achievement	7.21 (2.53)***	5.88 (1.86)	.42	6.98 (2.37)	7.05 (3.40)	.34
	n Affiliation	8.48 (2.47)	7.86 (3.18)	.62***	8.61 (3.88)*	7.26 (3.71)	.62***
	Word count	938 (276)***	744 (144)	.35	984 (369)	930 (389)	.93***
			n = 21			*n* = 21	

Note. Asterisks between T1 and T2 means indicate that means are significantly different (*t* test).
*$p < .05$, **$p < .01$, ***$p < .005$.

the entire sample (Table 2.) and for individual conditions (Table 1) indicating that participants who wrote longer stories at T1 also wrote longer stories at T2. The correlation was strongest in the two conditions in which participants wrote stories under the same condition and weaker in the two conditions in which they switched writing condition from T1 to T2.

Total raw motive scores generally followed the pattern observed for word count, reflecting their partial dependence on overall protocol length (cf. Table 1). But did motive scores vary as a function of writing condition once condition effects on PSE protocol length were taken into account? To answer this question, we residualized the motive score at T1 for word count at T1 and the motive score at T2 for word count at T2 and converted the residuals to *z* scores (cf. Schultheiss & Pang, 2007). Residualized motive scores did not significantly differ between conditions at T1 or at T2 (MANOVA $Fs(3,86) < 1$, $ps > .10$). Thus, despite the fact that participants who typed their stories imbued their stories with more motive images because they wrote more overall, they produced the same amount of motive imagery as participants who handwrote their stories once differences in protocol length were taken into account.

3.2. Inter-scorer reliability

Scoring agreement between both scorers ranged from acceptable (.70) to good (.87) for all motives and at both assessment times (cf. Table 2).

3.3. Retest reliability

Two-week retest correlations were highly significant and substantial for all three motives (residualized scores); the retest correlation for *n* Affiliation was higher than that of the other two motives (cf. Table 2 and Fig. 1). Of the three retest correlations, only the ones for *n* Affiliation and *n* Achievement differed significantly, $z = -2.03$, $p < .05$. But did the between-group fluctuations in motive retest correlations reported in Table 1 reflect a systematic effect of writing condition on retest reliability above and beyond differences in word count? To address this issue, we ran hierarchical regression analyses with word count-corrected motive score at T2 as dependent variable and word count-corrected motive score at T1, writing conditions at T1 and T2, and their multiplicative interaction terms as predictors. In all cases, the motive × condition (T1), the motive × condition (T2), and the motive × condition (T1) × condition (T2) effects were nonsignificant ($p > .10$), indicating that retest stability was not systematically influenced by writing condition.

Table 2
Reliability estimates for motive scores and word count

	Reliability					
	Inter-scorer[a]		Retest[b]	Internal consistency[c]		Ipsative stability[d]
	T1	T2		T1	T2	
n Power	.86	.78	.39	.38	.43	.21
n Achievement	.70	.77	.37	−.02	.26	.24
n Affiliation	.81	.87	.61	.20	.43	.40
Word count	—	—	.61	.97	.93	.14

[a] Pearson correlations for raw motive imagery scores.
[b] Pearson correlations for total motive scores (residualized for total word count).
[c] Cronbach's α for picture motive scores (residualized for picture word count and converted to *z* scores).
[d] Within-participant Pearson correlations for picture motive scores (residualized for picture word count and converted to *z* scores) and word count scores (converted to *z* scores), averaged across participants.

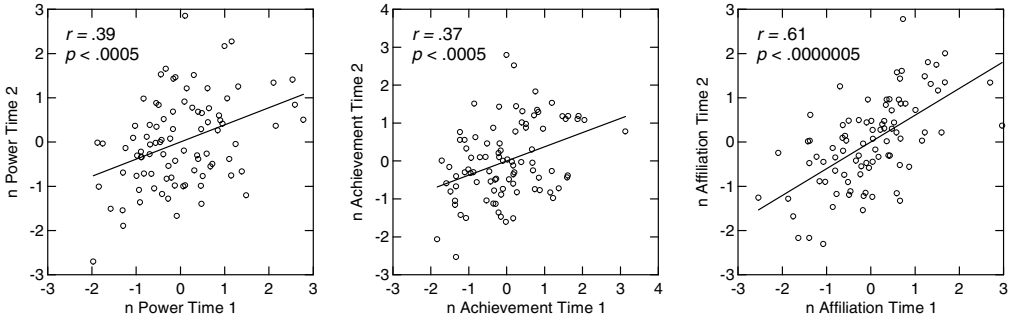

Fig. 1. Plots of motive scores, residualized for word count and converted to *z* scores, with fitted linear regression lines and Pearson correlation coefficients.

3.4. Internal consistency

Internal consistency, as estimated by Cronbach's α, was excellent for the eight picture word count scores at both assessments (cf. Table 2). To calculate Cronbach's α for picture motive imagery scores, we removed the effect of each picture's word count from each picture's motive score and converted the residuals to *z* scores. This transformation also removed the effect of a picture's overall cue strength for a given motive. Internal consistency for the resulting motive scores across all eight pictures was low at both assessments and in one case even negative (cf. Table 2). This indicates that for a given participant, motive scores on one picture were not particularly predictive of motive scores on other pictures after controlling for the effect of word count and overall picture cue strength.

3.5. Sample-level profile stability

To examine to what extent profiles of *average* motive scores for the eight pictures we employed in our PSE were stable from one testing session to the next, we ran three correlation analyses of the mean motive scores for each picture ($N = 8$). As shown in Fig. 2 the PSE profiles for each motive were highly similar from one assessment to the next. For instance, across the entire sample, Ship Captain produced the highest, and Couple by River the lowest, amount of power imagery at both assessments; Bicycle Race pulled for the most achievement imagery whereas Couple by River had very low achievement imagery levels at both times; and Girlfriends in Café was among best cues for eliciting affiliation imagery and Boxer was among the

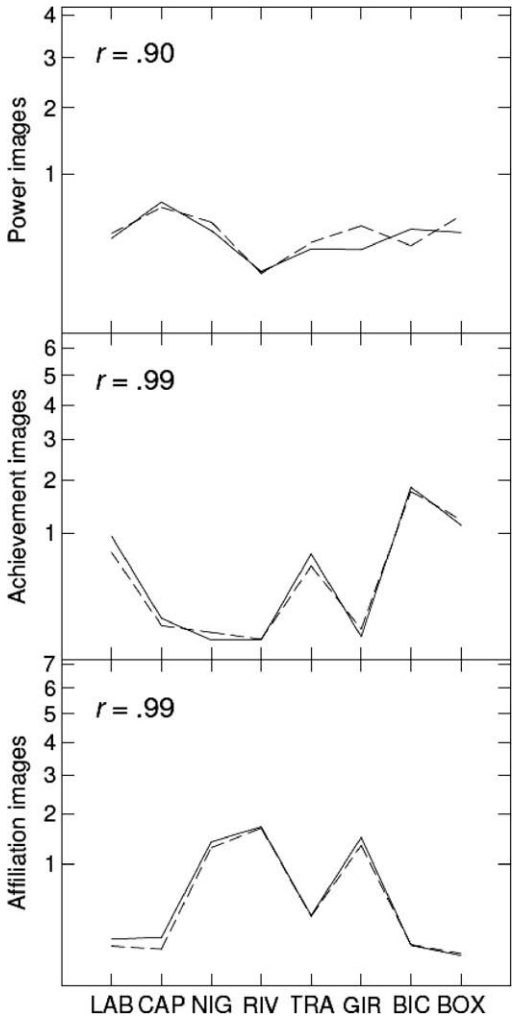

Fig. 2. Number of scored images (log-transformed), broken down by motive, picture cue, and assessment (T1, solid line; T2, striped line). Pearson correlations were calculated for the profile correlation between T1 and T2, based on mean scores per picture ($N = 8$ pictures; all $ps < .01$). LAB: Women in laboratory, CAP: Ship captain, NIG: Nightclub scene, RIV: Couple by river, TRA, Trapeze Artists; GIR: girlfriends in café with male approaching; BIC, bicycle race, BOX: boxer.

worst each time. As a consequence, profile correlations at the sample level were all very high and significant for all three motives, *p*s < .01. We also examined the profile stability for the average number of words participants wrote in response to all pictures and found it to be somewhat lower ($r = .72$, $p < .05$) than the profile correlations we had obtained for motive imagery.

3.6. Ipsative stability

To determine ipsative stability, that is, the extent to which the profile of scores for a given motive across pictures cues remained stable for each participant across assessments, we adopted the following procedure (cf. Shoda et al., 1994): (1) We residualized motive scores for each picture for the picture's word count and converted the residuals to *z* scores. The resulting motive scores thus indicate how much motive imagery a given participant injected into a given story *above and beyond* what would be expected given the length of that participant's story and the average amount of imagery a given picture elicited in participants. (2) Next, we calculated for each participant separately the ipsative correlation of her or his motive scores at T1 and T2. Because motive scores were derived from eight pictures, the correlation was based on eight measurements. (3) To determine the average ipsative correlation across all participants for each motive, we finally subjected each participant's ipsative-stability coefficient to an *r*-to-*z* transformation, averaged the *z* values across participants, and converted the averaged *z* value back to *r*. To determine the ipsative stability of word count scores, we first converted word count scores to *z* scores to remove average between-cue differences in story length and then followed steps 2 and 3 of the procedure. The results are presented in Table 2.

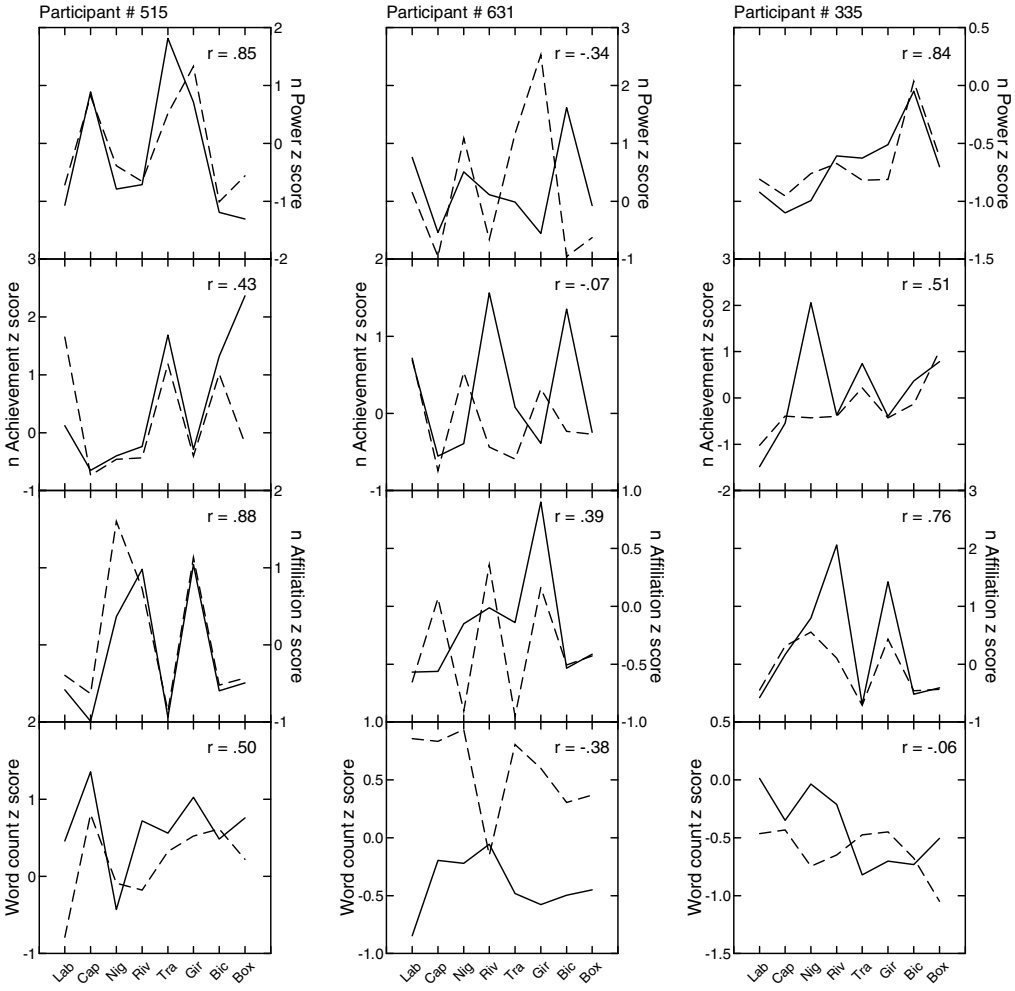

Fig. 3. Illustrative intraindividual profiles of *n* Power, *n* Achievement, and *n* Affiliation (residualized for word count and converted to *z* scores for each picture) and word count (converted to *z* scores for each picture) across eight picture cues (T1, solid line; T2, striped line). Ipsative correlation coefficients were calculated to the similarity of profiles between T1 and T2, based on each individual's scores per picture ($N = 8$ pictures). LAB, Women in laboratory; CAP, ship captain, NIG, nightclub scene; RIV, couple by river, TRA, Trapeze Artists, GIR, girlfriends in café with male approaching; BIC, bicycle race; BOX, boxer.

In all cases, ipsative stability was positive, significant ($ts > 3.97$, $ps < .01$, based on one-sample t tests for r-to-z transformed ipsative correlation coefficients), and accounted for 4–16% of the variance in motive scores. The ipsative correlation for n Affiliation was significantly higher than those for n Achievement and n Power ($ts > 2.00$, $ps < .05$, based on paired-sample t tests for r-to-z transformed ipsative correlation coefficients), whereas the ipsative correlations for n Power and n Achievement were not significantly different. To illustrate the meaning of the ipsative-stability finding, we plotted motive score and word count profiles for three representative participants from our sample (Fig. 3). Participant 515 (left panel of Fig. 3) wrote stories saturated with power imagery to the picture cues Ship Captain, Trapeze Artists, and Girls in Café both at T1 and at T2, whereas he wrote stories with less-than-expected power imagery about Women in Laboratory and Boxer at both assessments. The profiles of his n Power scores were highly similar at T1 and T2, as indicated by an ipsative correlation coefficient of .85. At both assessments, the same participant also wrote stories with a high level of achievement imagery in response to Trapeze Artists and Bicycle Race and a low level of achievement imagery in response to Ship Captain, Nightclub Scene, and Girls in Café, yielding an ipsative stability of .43 for n Achievement. Finally participant 515 showed considerable stability ($r = .88$) in his affiliation motive scores, with high levels observed at both testing occasions for the pictures Nightclub Scene, Couple by River, and Girls in Café, and very low levels for the pictures Ship Captain and Trapeze Artists.

In contrast, participant 631 illustrates the other end of the ipsative-stability spectrum. This participant had low and in two cases even negative ipsative-stability coefficients for motive scores, and closer inspection of the middle panel of Fig. 3 reveals why. Although the participant's scores for n Power appeared to be stable for some pictures (e.g., Women in Laboratory, Ship Captain, and Nightclub Scene), they fluctuated wildly for the rest of the pictures, yielding high scores for a given picture at one testing occasion and low scores at the next (e.g., Girls in Cafe). Similar discrepancies between scores at T1 and at T2 are apparent for the other two motives, although n Affiliation scores show moderate stability across assessments.

Note that because we had regressed word count out of motive scores, these between-subject differences in ipsative motive stability cannot be attributed to differences in ipsative word count stability, that is, to how much a participant wrote about a given story. The data for participant 335 illustrate this independence. This participant had high positive ipsative-stability coefficients for all three motives despite a negative stability coefficient for the word count variable. In other words: although participant 335's stories about the same picture cues differed in length across testing occasions, the relative amount of motive imagery she imbued them with remained stable.

4. Discussion

Our findings support our hypothesis that low PSE motive score consistency reflects the presence of stable *if…then* contingencies revealed in the way participants respond to PSE picture cues. As predicted, participants showed substantial stability from one testing occasion to the next in the way they imbued stories written in response to specific pictures with motivational imagery. As illustrated in Fig. 3 some participants had impressively large ipsative correlation coefficients for one, two or even all three motives, while for others ipsative stability was lower. But across all participants, these differences in size averaged out to yield ipsative correlation coefficients that were positive, significant, and in the expected .20–.40 medium effect size range for all three motives we examined (see Cohen, 1992; Shoda et al., 1994). This finding cannot be explained on the basis of differences in word count between stories or cue characteristics of the pictures we used, because the influence of these factors was statistically removed from motive scores before we examined ipsative stability. Thus, participants showed stable differential responses, as reflected in motive imagery, to picture cues *above and beyond* stable differences in how many words they wrote about each picture cue or whether a given picture cue happened to have high or low pull for a given type of motive imagery at the sample level. The finding is also unlikely to be explained by serial position effects of picture cues because picture sequence was randomized between participants and testing occasions.

Our findings are therefore in agreement with a central tenet of CAST (Mischel & Shoda, 1995), namely, that stable personality dispositions do not manifest themselves by individuals construing every situation in the same way (e.g., all situations are equally good opportunities for affiliating with others), but in the way in which they stably construe specific situations (e.g., being in a quiet place with another person is always a good opportunity for affiliative contact, but being stopped by a traffic cop is not). Of course, some people perceive more situations as suitable outlets for their motivational needs, leading them to respond with high amounts of motive imagery to several different picture cues, whereas others imbue only one picture story with lots of imagery, because other picture cues are not perceived as suitable opportunities for motive expression. One consequence of these interindividual differences is that the former group of PSE writers will have a higher overall amount of motive imagery, because more stories are contributing imagery to the total score, than the latter group. Hence, the total amount of imagery harvested from a PSE can remain stable from one testing occasion to the next, because the person who responded to more pictures with motive imagery the first time will also do so at the second time. Ipsative stability thus contributes to sum-score stability.

We think that besides replicating the observed ipsative-stability effect in other samples and with a wider range of retest intervals, a closer examination of the factors that give rise to low or high ipsative stability in motive scores merits further investigation. For instance, Mischel and Shoda (1995) found that individuals who viewed themselves as having a consistent personality also had higher ipsative stability in situation-specific conscientiousness (although they did not show greater trans-situational consistency in this trait) than individuals who did not view themselves as consistent. Does self-attributed

consistency also influence the magnitude of ipsative motive score stability? Research on motive stability suggests that major life events or stressors can lower the retest stability of PSE motive scores (cf. Koestner, Franz, & Hellman, 1991, cited in Smith, 1992). Is this the result of a reorganization in the way that particular situations—as approximated by PSE picture cues—elicit motivational responses in individuals experiencing these events and stressors? We think that these questions are particularly worthwhile to explore in future studies.

Our findings also breathe new life into the old concept of motive *extensity*. McClelland et al. (1953) argued that the strength of a motive is not only determined by how well a person has learned to associate a given situation with need satisfaction and hence how intensely she or he expresses the motive in response to the situation. According to these authors, motive strength is also a function of the number (i.e., extensity) of different situational cues the person has learned to view as an opportunity for need satisfaction. It remains an open question to what extent motive extensity expressed in the PSE corresponds to motive extensity in real-world behavior. Does a person who responds to many types of PSE picture cues with achievement imagery also construe more situations outside the testing laboratory as relevant for achievement than a person who responds only to few PSE cues? And how similar would the real-world situations be to the ones depicted in the PSE? Schultheiss and Pang (2007) recently reported anecdotal evidence indicating that motive scores from PSE pictures that are thematically linked to a subsequent experimental task appear to have better predictive validity for performance on that task than pictures with equally high pull for the motive in question, but lacking any similarity between their cue content and the experimental task. If systematic research on this issue indeed revealed that people do not only respond similarly to the same picture cue on different testing occasions but also to actual situations infused with incentives similar to those represented in that picture, the laws of learning theory would likely provide a better account of what is happening when a person takes the PSE than classical test theory or DOA theory.

Retest reliability of motive sum scores were substantial and significant for all three motives examined, although only the retest coefficient for *n* Affiliation was of a magnitude we had expected based on Schultheiss and Pang's (2007) meta-analysis; coefficients for *n* Power and *n* Achievement were below the predicted range of .50–.60. This finding was also paralleled by our observation of higher ipsative stability for *n* Affiliation than for *n* Power or *n* Achievement. One reason for the lower-than-expected retest results for *n* Power and *n* Achievement may be that the PSE we used appeared to be particularly well-suited for eliciting high amounts of affiliation imagery and had less pull for power and achievement imagery. Also, many of the studies on PSE retest reliability incorporated into Schultheiss and Pang's (2007) meta-analysis used picture sets designed to assess a single motive well. In contrast, we used a PSE with picture cues aimed at the assessment of all three major motives. This difference may also account for the overall lower retest reliability of our PSE.

Like in numerous previous studies, inter-scorer reliability was high, indicating that two independent scorers could readily agree on the presence or absence of motive imagery in a given PSE story. Moreover, like in previous studies, internal consistency of motive scores across picture cues was low for all three motives and at both testing occasions, once more underscoring the basic point that how a person responds motivationally to one PSE picture says little about how the person responds to other pictures. In this context, it is noteworthy to observe that the length of PSE stories, that is, the matrix in which motive imagery is scored, was extremely consistent from one picture to the next. However, probably due to the high level of picture-to-picture consistency, PSE length showed considerably less ipsative stability for a given picture from one testing occasion to the next than, for instance, *n* Affiliation scores after the influence of word count had been removed from them. Both *n* Affiliation and word count yielded the same retest correlation coefficient of .61, which in the case of *n* Affiliation was more a function of substantial ipsative stability than internal consistency and in the case of word count was primarily due to picture-to-picture consistency in story length. These results therefore illustrate that a measure's retest reliability can be the result of varying admixtures of internal consistency and ipsative stability.

Results from our study replicate the stable differences in the pull various pictures exert on motivational imagery observed in previous research (cf. Langan-Fox & Grant, 2006; Pang & Schultheiss, 2005; Schultheiss & Brunstein, 2001). For instance, in our study, like in those earlier studies, Women in Laboratory elicited substantial amounts of achievement imagery, Couple by River elicited high amounts of affiliation imagery, and Ship Captain was particularly likely to yield power motive imagery. Moreover, our findings indicate that such parallels between earlier studies and ours are due to precisely replicable pull characteristics of individual picture cues. At the sample level, average motive scores for the 8 pictures we employed were highly correlated from one testing occasion to the next, despite the randomized order in which pictures were presented at both occasions and further randomization during scoring. Combined with earlier findings, these results therefore underscore that each PSE picture cue has specific motivational pull properties that are highly stable and replicable within and across sample populations (see Langan-Fox & Grant, 2006; Pang & Schultheiss, 2005; Schultheiss & Brunstein, 2001).

Our comparison of typed and handwritten PSE stories showed that research participants produce substantially more material when they can type their stories on a computer keyboard than when they write their stories by hand. But once variations in story length due to writing condition were statistically controlled, writing condition did not have a significant effect on the amount of motive imagery participants injected into their stories. And most importantly, we found no evidence that retest stability of motive scores differed between participants who had produced PSE stories under the same writing condition at T1 and T2 and participants who had been tested under changing writing conditions. Although the similarity of stability coefficients between writing conditions does not constitute proof positive for similar validity of motive scores based on handwritten versus typed PSE stories, our results indicate that the same stable portion of variance in PSE motive scores is captured by handwritten and by typed stories.

This is good news for researchers, because it means that for the PSE both picture presentation and recording of written responses can be handled by a computer. This provides a couple of advantages over the traditional paper-and-pencil administration format. First, because all instructions can be given by the computer in a highly standardized manner and testees are completing the PSE on their own, experimenter effects, which can profoundly affect the amount and validity of PSE motive scores (cf. Klinger, 1967; Lundy, 1988; McClelland et al., 1953), are minimized. Second, typed PSE stories are longer by more than a third and therefore also provide more scoreable material. Although this does not, as we have shown, alter the amount of imagery per given amount of text, the concomitant increase in raw motive score means and variances is advantageous from a diagnostic point of view, because it provides more differentiated scores and normal score distributions (cf. Schultheiss & Pang, 2007, for a discussion of this issue). Third, transcription of handwritten stories on a word processor—an extremely time-consuming task—is no longer necessary. Fourth, computer administration of the PSE dramatically reduces the amount of time invested in subsequent story scoring and sources of random error associated with it. For instance, determination of word count, which used to take up considerable time and be a monotonous and thus error-prone task for PSE scorers, can be handled automatically by computer software, cutting hours of work down to mere seconds. Also, PSE scoring can be done by inserting score signatures into text documents on the computer. Specialized software can then extract scores and transcribe them into a data file. This, too, saves time and error by eliminating the manual entering of motive scores from PSE stories into a data file. In our estimate, all of these savings add up to reducing by a half or more the time it takes to go from PSE stories to scores that can be statistically analyzed. A final advantage of collecting typed PSE stories is, of course, that they can be readily archived on a computer storage medium without taking up any shelf space in the laboratory and that they can be easily made available to others as electronic files.

To conclude, the present research not only replicates the frequent observation that PSE motive measures can yield stable overall scores in the absence of acceptable internal consistency of scores across picture stories, it also provides an explanation for this conundrum by revealing that PSE motive scores have substantial ipsative stability, that is, that individuals respond similarly to the same PSE picture cue across testing occasions. We have argued that this finding is consistent with Mischel and Shoda's (1995) conceptualization of personality dispositions as sets of stable propensities to respond only to some situations, but not all situations equally, in a specific way (*if. . .then* contingencies). Our findings also suggest that motive scores obtained from PSE stories written on the computer share stable variance with and are comparable to motive scores obtained from handwritten PSE stories. Because the computer administration of the PSE carries significant advantages for the collection, processing, and scoring of PSE stories, we recommend this assessment approach for future research on implicit motives.

Acknowledgement

All PSE picture cues and Inquisit and Matlab scripts used in this research are available upon request from the first author. Parts of this research were supported by a National Science Foundation grant (BCS-0444301).

References

Atkinson, J. W. (1950). *Studies in projective measurement of achievement motivation*. Unpublished doctoral dissertation, University of Michigan, Ann Arbor, MI.
Atkinson, J. W. (1981). Studying personality in the context of an advanced motivational psychology. *American Psychologist, 36*, 117–128.
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.
Ball, G. F., & Balthazart, J. (2008). How useful is the appetitive and consummatory distinction for our understanding of the neuroendocrine control of sexual behavior? *Hormones and Behavior, 53*, 307–311.
Berridge, K. C. (2001). Reward learning: Reinforcement, incentives and expectations. In D. L. Medin (Ed.). *The psychology of learning and motivation* (Vol. 40, pp. 223–278). New York, NY: Academic Press.
Blankenship, V., Vega, C. M., Ramos, E., Romero, K., Warren, K., Keenan, K., et al (2006). Using the multifaceted Rasch model to improve the TAT/PSE measure of need for achievement. *Journal of Personality Assessment, 86*(1), 100–114.
Blankenship, V., & Zoota, A. L. (1998). Comparing power imagery in TATs written by hand or on the computer. *Behavior Research Methods, Instruments, & Computers, 30*, 441–448.
Brownell, M. H., & Goss, A. E. (1957). Stimulus–response analysis of inferences from projective test behavior. *Journal of Personality, 25*, 525–538.
Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.
Entwisle, D. R. (1972). To dispel fantasies about fantasy-based measures of achievement motivation. *Psychological Bulletin, 77*, 377–391.
Fineman, S. (1977). The achievement motive construct and its measurement: Where are we now? *British Journal of Psychology, 68*, 1–22.
Fodor, E. M., Wick, D. P., & Hartsen, K. M. (2006). The power motive and affective response to assertiveness. *Journal of Research in Personality, 40*, 598–610.
Goss, A. E., & Brownell, M. H. (1957). Stimulus–response concepts and principles applied to projective test behavior. *Journal of Personality, 25*, 505–523.
Hofer, J., & Chasiotis, A. (2004). Methodological considerations of applying a TAT-type picture-story test in cross-cultural research: A comparison of German and Zambian adolescents. *Journal of Cross-Cultural Psychology, 35*(2), 224–241.
Inquisit 2.0.50401 [Computer software]. (2005). Seattle, WA: Millisecond Software LLC.
Klinger, E. (1967). Modeling effects on achievement imagery. *Journal of Personality and Social Psychology, 7*, 49–62.
Koestner, R., Franz, C. E., & Hellman, C. (1991). *Life changes and the reliability of TAT motive assessment*, unpublished manuscript.
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. E., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1*, 27–66.
Lundy, A. (1985). The reliability of the Thematic Apperception Test. *Journal of Personality Assessment, 49*, 141–145.
Lundy, A. (1988). Instructional set and Thematic Apperception Test validity. *Journal of Personality Assessment, 52*, 309–320.
MATLAB 7.0 [Computer software]. (2007). Natick, MA: The MathWorks, Inc.

McClelland, D. C. (1980). Motive dispositions. The merits of operant and respondent measures. In L. Wheeler (Ed.). *Review of personality and social psychology* (Vol. 1, pp. 10–41). Beverly Hills, CA: Sage.
McClelland, D. C. (1989). Motivational factors in health and disease. *American Psychologist, 44*, 675–683.
McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). *The achievement motive*. New York: Appleton-Century-Crofts.
Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*, 246–268.
Murray, H. A. (1943). *Thematic Apperception Test*. Cambridge, MA: Harvard University Press.
Nesselroade, J. R., & Boker, S. M. (1994). Assessing constancy and change. In T. F. Heatherton & J. L. Weinberger (Eds.), *Can personality change?* (pp 121–147). Washington, DC: American Psychological Association.
Nunnally, J. C. (1967). *Psychometric theory*. New York, 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*(3), 280–294.
Reuman, D. A. (1982). Ipsative behavioral variability and the quality of thematic apperceptive measurement of the achievement motive. *Journal of Personality and Social Psychology, 43*(5), 1098–1110.
Schultheiss, O. C. (2007). A biobehavioral model of implicit power motivation arousal, reward and frustration. In E. Harmon-Jones & P. Winkielman (Eds.), *Social neuroscience: Integrating biological and psychological explanations of social behavior* (pp. 176–196). New York: Guilford.
Schultheiss, O. C. (in press). Implicit motives. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed.). New York: Guilford..
Schultheiss, O. C., & Brunstein, J. C. (2001). Assessing 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., & Pang, J. S. (2007). Measuring implicit motives. In R. W. Robins, R. C. Fraley, & R. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology* (pp. 322–344). New York: Guilford.
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*, 195–202.
Shoda, Y., Mischel, W., & Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: Incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology, 67*(4), 674–687.
Smith, C. P. (1992). Reliability issues. In C. P. Smith (Ed.), *Motivation and personality: Handbook of thematic content analysis* (pp. 126–139). New York: Cambridge University Press.
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*, 448–461.
Winter, D. G. (1993). Power, affiliation, and war: Three tests of a motivational model. *Journal of Personality and Social Psychology, 65*, 532–545.
Winter, D. G. (1994). *Manual for scoring motive imagery in running text* (4th ed.). Department of Psychology, University of Michigan, Ann Arbor, unpublished manuscript.
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.
Woike, B., Lavezzary, E., & Barsky, J. (2001). The influence of implicit motives on memory processes. *Journal of Personality and Social Psychology, 81*, 935–945.