Assessment of Multiple Implicit Self-Concept Dimensions Using the Extrinsic Affective Simon Task (EAST)

SARAH TEIGE, KONRAD SCHNABEL,
RAINER BANSE and JENS B. ASENDORPF*

Humboldt University, Berlin

Abstract

This study explored the psychometric properties of the Extrinsic Affective Simon Task (EAST; De Houwer, 2003a) as adapted for the measurement of the implicit self-concept of personality. The EAST was adapted to allow the simultaneous assessment of the three traits shyness, anxiousness, and angriness. In order to test the EAST's psychometric properties, 100 participants completed a trait EAST, Implicit Association Tests (IATs), and direct self-ratings. The EAST showed low internal consistencies and correlated neither with the IATs nor with the direct measures. The main problem of the EAST, namely its low reliability, is discussed, and general conclusions regarding the indirect assessment of the personality self-concept by EASTs are derived. Copyright © 2004 John Wiley & Sons, Ltd.

How can we best assess personality traits? Although verbal self-reports are widely used, they suffer from two main problems: limited cognitive accessibility and differential self-presentation tendencies such as social desirability. In order to overcome these problems, there is flourishing research on alternative methodologies, particularly reaction-based indirect measures of individual associative structures such as the *Implicit Association Tests* (IATs) by Greenwald, McGhee, and Schwartz (1998).

The present study aims at contributing to the latter approach. Extending prior research by Asendorpf, Banse, and Mücke (2002) and Schnabel (unpublished dissertation) on the IAT, it explores the usefulness of a new class of indirect measures that has been recently developed by De Houwer (2003a): the *Extrinsic Affective Simon Tasks* (EASTs). Using the three traits shyness, anxiousness, and angriness, it was tested whether it is possible to design an EAST that (i) reliably assesses individual differences in the implicit self-concept of these traits and (ii) shows convergent and discriminant validity.

*Correspondence to: Jens B. Asendorpf, Institut für Psychologie, Humboldt-Universität zu Berlin, Rudower Chaussee 18, 12489 Berlin, Germany. E-mail: asen@rz.hu-berlin.de

For this purpose, we will first review differences between the explicit and implicit self-concept of personality and their related direct and indirect measures. Second, previous indirect measures, particularly the Implicit Association Tests (IATs), and their main shortcomings, are outlined. Third, the EAST as a promising indirect measure is presented in detail. Fourth, five hypotheses aiming at replicating earlier findings for IATs and evaluating the EAST's psychometric properties are generated and tested.

EXPLICIT VERSUS IMPLICIT PERSONALITY SELF-CONCEPT

In line with a general definition of the self-concept by Greenwald et al. (2002), we define the self-concept of personality as an 'associative network containing all associations of the concept of self with attribute concepts describing one's personality, thus attributes that describe individual, relatively stable, non-pathological characteristics of the person' (Asendorpf et al., 2002, p. 381). The self-concept of personality is assumed to be represented both *explicitly* and *implicitly*. These two kinds of mental representation are thought to be based on two modes of information processing, as suggested by the Reflective Impulsive System Model (RIS model) by Strack and Deutsch (in press): the *reflective* system, which processes information more consciously and in a more controlled way, and the *impulsive* system, which processes information more unconsciously, automatically, and intuitively (for further dual-process theories, see also Chaiken & Trope, 1999; Epstein, Pacini, Denes-Raj, & Heier, 1996; Greenwald et al., 2002; Hogarth, 2001; Lieberman, 2000; Wilson, Lindsey, & Schooler, 2000). Accordingly, *explicit representations* of the personality self-concept are based on the propositional structures of the *reflective system*, whereas *implicit representations* of the personality self-concept are based on the associative structures of the *impulsive system*. This distinction of explicit and implicit mental representations at the construct level is related to two different approaches of measurement at the empirical level. *Direct measures* (such as verbal self-report) assess intentionally given information about the self and are thus particularly suited to assess the explicit personality self-concept. In contrast, *indirect measures* (such as projective tests, priming, reaction-based measures) aim at assessing information about the self that is *not* intentionally given and are thus suited to assess the implicit personality self-concept. As there is some evidence that explicit and implicit representations are not identical (Fazio & Olson, 2003; Schwarz & Bohner, 2001; Strack & Deutsch, in press), the assessment of the self-concept of personality should concentrate on both the assessment of explicit representations by direct measures *and* the assessment of implicit representations by indirect measures.

RELIABLE ASSESSMENT BY INDIRECT MEASURES

In search of suitable indirect measures, personality psychologists could benefit from the flourishing social cognition research on the indirect assessment of implicit attitudes and stereotypes (Fazio & Olson, 2003). In this field, especially semantic or evaluative priming methods have been applied to the assessment of implicit attitudes (see e.g. Fazio, Jackson, Dunton, & Williams, 1995), stereotypes (see e.g. Blair & Banaji, 1996; Wittenbrink, Judd, & Park, 1997), self-esteem (see e.g. Bosson, Swann, & Pennebaker, 2000), and self and significant others (see e.g. Banse, 1999, 2001). However, the effect sizes are typically small, and the reliability of individual difference measures, which are based on priming, is

typically low (see e.g. Banse, 1999, 2001; Bosson et al., 2000) or at best moderate (see e.g. Cunningham, Preacher, & Banaji, 2001; Kawakami & Dovidio, 2001). Because such unreliable measures fail to detect interindividual differences, their applications to the assessment of implicit traits at the individual level are highly problematical (Greenwald & Banaji, 1995). Thus, a methodology is needed that routinely shows internal consistencies and retest reliabilities between 0.70 and 0.90 as it is the case with direct measures assessing explicit traits.

IMPLICIT ASSOCIATION TESTS (IATs)

Recently, Greenwald et al. (1998) developed a promising class of indirect measurement tools, the *Implicit Association Tests* (IATs), that meet these requirements to a large extent. Standard IATs assess the strength of associations between a bipolar target concept (such as Blacks versus Whites) and a bipolar attribute concept (such as Pleasant versus Unpleasant) by comparing the response times for two differently combined discrimination tasks. Participants are instructed to categorize stimuli that represent the four concepts (e.g. Black versus White names and pleasant versus unpleasant words). For categorization, participants use two response keys, each assigned to two of the four concepts. The IATs' basic assumption is that if two concepts are highly associated, the categorization will be easier (i.e. faster) when the two associated concepts share the same response (*compatible* task, De Houwer, 2003b) than when they require different responses (*incompatible* task, De Houwer, 2003b). For example, individuals with implicit prejudices against Blacks should react faster when Black names and unpleasant attributes are assigned to the same response key (Black and Unpleasant) in comparison with the reversed configuration (Black and Pleasant).

The effects produced by IATs are typically much larger than priming effects, and several studies have shown that IATs assess individual differences in implicit attitudes and self-esteem with internal consistencies (Cronbach's α) around 0.80 (Banse, Seise, & Zerbes, 2001; Bosson et al., 2000; Cunningham et al., 2001; Greenwald & Farnham, 2000; Greenwald & Nosek, 2001). Even IATs for the assessment of implicit personality traits show satisfactory internal consistencies that approach 0.80 (Asendorpf et al., 2002; Egloff & Schmukle, 2002; Schnabel, unpublished dissertation). However, the IATs' test–retest correlations, which typically range between 0.60 and 0.70 at best (see Egloff & Schmukle, 2002), are not fully satisfactory, although much higher than those reported for priming methods.

Problems of the IATs
Despite their encouraging reliability, several aspects of the IATs have recently been criticized (for an overview, see Greenwald & Nosek, 2001). Much of this criticism revolves around the fact that IAT scores are based on a comparison of performance in the two separate compatible and incompatible tasks (De Houwer, 2003a).

As a first consequence, IAT measures are limited to the assessment of the *relative strengths of associations* between bipolar concepts (see e.g. Nosek, Greenwald, & Banaji, unpublished manuscript). For instance, a shyness IAT assesses the association strength between the target concept of interest ('Me') and the attribute concept of interest ('Shy') only relative to the association strength between the contrary target concept ('Others') and the same attribute concept ('Shy'). Consequently, the implicit concept of the self cannot be

assessed separately, but just in comparison with the concept of others. This is critical because one is usually more interested in simple association strengths than in relative association strengths.

A second problem of IAT measures is also a consequence of the different pairing of concepts within the two combined tasks and concerns the IATs' limitation to the assessment of *one bipolar target concept per test*. In contrast to direct measures that have the same format, IATs cannot assess multiple traits, attitude objects, or stereotypes simultaneously. Thus, IATs for each trait have to be completed separately. Even more worrisome than this inefficiency are recent findings by Schnabel (unpublished dissertation): the position of consecutive anxiousness and angriness IATs affected the correlations between the IATs and the direct measures, and therefore appear as a critical limitation to the applicability of IATs. Schnabel's findings might be explained by transfer effects resulting from participants' strategy use.

This strategy use represents the third problem of the IATs, namely *task recoding problems* (De Houwer, 2003a). There is some evidence that participants recode the IATs' task instructions and use strategies that may facilitate the two separate discrimination tasks. As long as interindividual differences in IAT effects are unaffected by this strategy use, recoding is unproblematic for IATs in differential applications (e.g. as it seems to be the case for shift in response criteria, Brendl, Markman, & Messner, 2001, or recoding with regard to a figure–ground schema, Rothermund & Wentura, 2001). Mierke and Klauer (2003) showed an effect of interindividual differences in task-switching costs on IAT scores but this effect disappeared when a new IAT outcome measure was used (the D measure proposed by Greenwald, Nosek, and Banaji, 2003). If such differential recoding occurs, it is highly critical for the indirect assessment of interindividual differences in implicit traits, and should be controlled.

Alternatives to the IATs

More recently, further promising reaction based indirect measures for the assessment of the implicit personality self-concept have been developed such as the *Single Target IATs* (STIATs; for further detail, see Wigboldus, 2003) or *Implicit Association Procedures* (IAPs; for further detail, see Schnabel, unpublished dissertation). These indirect measures show encouraging results, as they not only replicate IAT findings but also overcome some of the IATs' above mentioned problems. However, the *Extrinsic Affective Simon Tasks* (EASTs) recently developed by De Houwer (2003a) might have the highest potential so far for overcoming the IATs' problems in assessing implicit traits. Like most of the indirect measures, the EAST methodology was originally developed for the assessment of attitudes and self-esteem. In the following, such an original EAST assessing self-esteem is exemplarily described in more detail, and its potential in comparison with the IAT methodology is discussed. Subsequently, some modifications regarding the EAST's adaptation to the assessment of implicit personality traits are briefly presented.

EXTRINSIC AFFECTIVE SIMON TASKS (EASTs)

In the EASTs, attribute stimuli (e.g. good versus bad), presented as white words, and target stimuli (e.g. flowers versus insects, me versus others), presented as coloured words, appear one-by-one on a computer screen. Depending on the word's colour, participants are instructed to discriminate stimuli by pressing a left or a right response key according to

different features: white words (attributes) have to be discriminated according to their *semantics* (such as their valence, e.g. good versus bad, or their categorical meaning, e.g. shy versus nonshy), whereas coloured words (targets) have to be discriminated according to their *colour* (blue versus green) regardless of their semantics.

Standard EASTs (e.g. for the assessment of self-esteem) consist of three phases, two for practicing and one for testing. In the first phase, participants practice semantic discrimination, here according to valence. Therefore, only white attribute words of the *good* and *bad* category are presented, and participants are instructed to press a left key in response to pleasant words and a right key in response to unpleasant words. It is assumed that after many valence discriminations, the response keys become *extrinsically* related to (i.e. associated with) semantics, although they are *intrinsically* unrelated to semantics, i.e. key pressing per se has no semantic relation. Accordingly, the left key becomes (extrinsically) associated with positive valence, while the right key becomes (extrinsically) associated with negative valence, leading to a left 'positive key' and a right 'negative key'. In the second phase, participants practice colour discrimination. Therefore, only coloured target words of the category *me* versus *others* are presented, each word appearing in both blue and green colour. Participants are instructed to press the left 'positive key' in response to blue coloured words and the right 'negative key' in response to green coloured words (or vice versa). In the third phase, the critical EAST phase, the two tasks are combined: both white attribute and coloured target words are presented in a fixed random order, and participants are instructed to discriminate white words according to their semantics (here, valence) and coloured words according to their colour by pressing the 'positive key' versus the 'negative key' (e.g. blue colour–positive key versus green colour–negative key).

The EASTs' basic assumption is that participants automatically associate the target stimuli (e.g. me) with semantics (e.g. good), although they are instructed to concentrate exclusively on colour for discrimination. Thus, their performance on responding to target stimuli by pressing the semantically associated (here, positive or negative) key is determined by both the feature *colour* that is relevant for discrimination and the feature *semantics* (here, valence) that is irrelevant for discrimination. Given that the target stimuli vary in their associated valence, this produces evaluatively *compatible* trials, in which the target stimulus and the correct response are associated with the same valence, and evaluatively *incompatible* trials, in which the target stimulus and the correct response are associated with a different valence. For example, individuals with a high self-esteem should react faster when, due to their word colour, *me* stimuli have to be assigned to the 'positive key' (e.g. the blue colour–positive key needs to be pressed because the word 'me' is presented in blue colour) in comparison with the reversed configuration (e.g. 'me' has to be attached to the green colour–negative key because it is presented in green colour).

Altogether, the main difference to the IAT approach is that the EASTs measure associations within a single task (i.e. performances on compatible *trials* can be compared with performances on incompatible *trials*), whereas IATs measure associations within two different tasks (i.e. performances on the compatible *task* (i.e. test block) are compared to performances on the incompatible *task* (i.e. test block)). Hence, EASTs are not subject to the main source of the IATs' problems, namely the two separate compatible and incompatible tasks.

Applying the EAST to the assessment of implicit traits

Like the IATs, the EASTs have been already successfully applied as a measure of attitudes toward multiple target stimuli (e.g. flower, insect) and as a measure to assess self- and

other-esteem (De Houwer, 2003a). However, unlike the IATs, the EASTs have not yet been adapted as a measure of interindividual differences in implicit personality traits, despite their potential for assessing multiple traits with one test. The present study therefore aimed to fill this gap by adapting the EAST to the assessment of implicit traits. In view of the reliability problem, from which most indirect measures suffer, special attention was paid to the EAST's psychometric properties. To profit from its advantage of the simultaneous assessment of multiple concepts, the EAST was adapted to provide an indirect measure of the three traits shyness, anxiousness, and angriness (for reasons for their selection, see next section). In the following, the term 'EAST' is used in the singular although referring to the shyness, anxiousness, and angriness EAST scores, because a single test is used to assess the three traits simultaneously.

Obviously, such a multiple trait EAST has to be designed differently from the standard EAST developed for the assessment of self-esteem. Three important modifications have to be made. First, just as in the adaptation of the self-esteem IAT to self-concept IATs by Asendorpf et al. (2002), attributes are trait descriptive, not valence descriptive. Second, attributes indicating different traits are presented in the same test. This requires, third, that the discrimination tasks of the target concept and attribute concepts have to be reversed. In the original EAST, the attribute concepts are presented as white words, which have to be discriminated according to semantics, thereby leading to semantically associated response keys. If the discrimination tasks in the present multiple trait EAST were not reversed, the response keys would become extrinsically associated with shyness, anxiousness, and angriness at the same time, thereby leading to a 'shy, anxious, angry key' and a 'nonshy, secure, self-controlled key'. As a result, the concepts would interfere with each other.

To avoid this problem, the attribute stimuli were presented in colour and therefore had to be discriminated according to their colour (*blue* or *green*). The target stimuli, on the other hand, were presented in white colour and therefore had to be discriminated according to their semantics, i.e. categorical meaning. Over many trials, response keys should become associated with *others* and *me*, thereby leading to an 'others key' and a 'me key'. Note, however, that this necessary inversion rules out the possibility of assessing simple associations: like the IATs, the present EAST assesses the association strength between the target concept of interest ('me') and the attribute concept ('shy') only relative to the association strength between the opposite target concept ('others') and the same attribute concept ('shy').

To prevent confusion about the terms 'compatible trial' and 'incompatible trial' in the context of assessing personality traits, the following remarks may be helpful. Consistent with De Houwer (2003b), compatibility refers to a *structural analysis* of the indirect measure and an *a priori* definition. Thus, usage of the terms (in)compatibility is *not* a function of the participant. For instance, in a shyness EAST, a trial is 'compatible' if *shy* stimuli have to be attached to the 'me key' and *nonshy* stimuli have to be attached to the 'others key'. A trial is 'incompatible' if *nonshy* stimuli have to be attached to the 'me key' and *shy* stimuli have to be attached to the 'others key'.

SUITABILITY OF SHYNESS, ANXIOUSNESS, AND ANGRINESS

The traits shyness, anxiousness, and angriness were selected for two main reasons. First, shyness, anxiousness, and angriness have been measured both directly and indirectly in

recent studies (Asendorpf et al., 2002; Schnabel, unpublished dissertation). Therefore, data of the present trait EAST can be compared to IAT data with respect to internal consistency as well as correlations with direct and indirect measures. Second, these three traits allow clear predictions of discriminant validity. Shyness and anxiousness are similar, but not identical, concepts at the construct level, loading negatively on extraversion and positively on neuroticism. Angriness, on the other hand, has positive factor loadings on extraversion and negative loadings on agreeableness, thereby being orthogonal to shyness and anxiousness in the Big Five space. Thus, direct and indirect shyness and anxiousness measures should be highly correlated, but should not correlate with direct or indirect angriness measures.

DESIGN AND HYPOTHESES

Reliability as well as convergent and discriminant validity of the EAST can only be evaluated if the suitability of the present data is secured. As the present study is an extension of the studies by Asendorpf et al. (2002) and Schnabel (unpublished dissertation), it should first and foremost replicate their findings.

To correctly evaluate the convergent validity of the EAST, its correlation with another indirect measure, e.g. an IAT, is needed. Therefore, in the present study, a shyness and an angriness IAT were included. An anxiousness IAT was omitted to keep the experimental session at a reasonable length. This does not pose a problem for the purposes of the present study, as the convergent validity of the EAST can be sufficiently evaluated through its correlations with the shyness and angriness IATs. Because the participants had to complete one EAST and two IATs, an EAST retest was not included in the design in order to avoid overtesting; therefore, it was not possible to evaluate the EAST's retest reliability.

Replications of earlier findings

Hypothesis 1 (replication of the IAT effects and results of the direct measures)
With regard to the means of the IATs and the direct measures, the present study should be comparable to the studies by Asendorpf et al. (2002) and Schnabel (unpublished dissertation). In particular, negative means are expected for both IATs.

Hypothesis 2 (replication of the psychometric properties of the IATs)
The internal consistencies of the IATs should be comparable to the studies by Asendorpf et al. (2002) and Schnabel (unpublished dissertation), where they approached 0.80. The present study should replicate the findings by Asendorpf et al. (2002) concerning the correlations between the shyness IAT and its corresponding direct measures (approximately 0.40) as well as its non-correlations with non-corresponding measures. For the angriness IAT comparable results are expected.

Psychometric properties of the EAST

Hypothesis 3 (reliability of the EAST)
As reliability has been a serious problem with most of the prior indirect measures, but is a necessary prerequisite for each measure assessing interindividual differences, the present EAST has to meet the requirement of satisfactory internal consistency. Thus, the internal consistencies of the three EAST scores are expected to approach an α of 0.80, similarly to

that found with the IATs for the same traits (see Asendorpf et al., 2002; Schnabel, unpublished dissertation).

Hypothesis 4 (convergent validity of the EAST)

Correlations between different indirect measures for the same trait are typically lower than correlations between different direct measures for the same trait. The latter regularly approach at least 0.70, whereas the former approach 0.50 at best (Cunningham et al., 2001; Schnabel, unpublished dissertation). This inferiority of indirect measures is caused by two general problems: low reliability and high method variance. Despite their satisfactory internal consistencies, the IAT and the IAP for shyness correlated only 0.50 (Schnabel, unpublished dissertation). However, IATs and IAPs are structurally more similar, i.e. share more method variance, than IATs and EASTs (for structural analyses of indirect measures, see De Houwer, 2003b). Therefore, correlations between the EAST scores and their corresponding IATs are expected to be 0.50 at best. Correlations between the EAST scores and their corresponding direct measures should be similar to the IATs' correlations with their corresponding direct measures (see Asendorpf et al., 2002). That is, the shyness EAST score should correlate with the direct shyness measures close to 0.40. A comparable strength of correlations is expected between the anxiousness EAST score and the direct anxiousness measures as well as between the angriness EAST score and the direct angriness measures.

Hypothesis 5 (discriminant validity of the EAST)

According to the similarity of shyness and anxiousness at the construct level, the shyness EAST score is expected to show positive correlations with direct and indirect anxiousness measures, and the anxiousness EAST score is expected to show positive correlations with direct and indirect shyness measures. If both concepts were assessed by the same method (e.g. EAST), they should correlate even higher, because there is no specific method variance that could lower the correlation (see Campbell & Fiske, 1959). However, the shyness and anxiousness EAST scores should not correlate with the direct or indirect angriness measures because of the dissimilarity between angriness and shyness/anxiousness at the construct level.

METHOD

This section starts with a description of the sample. Second, the study's overall procedure is briefly presented. Third, each measure is described in more detail in chronological order of the procedures' appearance within the study.

Participants

Participants were 100 non-psychology university students who were native speakers of German, right handed, and not colour blind. Most participants were approached on the campus of Humboldt University, Berlin, and were asked to participate; some were recruited through postings in the main university building; other participants were acquaintances of the experimenters.

   To motivate students to participate and to foster honest judgments on their self-reports minimizing socially desirable responding, participants were told that the study was on 'self-perception'. Furthermore, they were informed that they would receive either 6 € (approximately 6 US $) or a cinema ticket for their participation in the 50 minute study.

One participant left before completion of the study's last test and was therefore excluded from all analyses. This led to a final sample of 99 participants (50 male, 49 female; age $M = 24.0$ years, range 20–31 years).

Assessments and measures

Overall procedure

All participants (i) rated themselves on bipolar personality-describing items, (ii) completed an EAST assessing shyness, anxiousness, and angriness simultaneously, (iii) judged themselves on other questionnaire scales, (iv) completed a shyness IAT and an angriness IAT, (v) reported personal data (sex, age, dominant hand, academic subject, and length of time spent at university), (vi) were interviewed about the indirect procedures, and (vii) were completely debriefed. All tests were presented on a computer, and data from the indirect procedures were recorded using ERTS software (Beringer, 1994).

The shyness and angriness items were identical for the two indirect procedures (EAST and IATs) and were—in addition to the anxiousness items of the EAST—directly self-rated in step (i). The order of the shyness IAT and the angriness IAT in step (iv) was counterbalanced. The assignment to the two conditions was balanced for sex and alternated between successive participants. Finally, the participants were thanked for their participation, asked for permission to analyse their data (all agreed), and received either money or the cinema ticket. Three months after completion of the data collection, participants were sent a letter explaining the procedures and general findings of the study.

Bipolar adjective pairs

Participants rated 30 bipolar adjective pairs on a seven-point scale. The adjective pairs were introduced with a trait-instruction (i.e. 'How well do these two opposite adjectives describe your personality in general?') and presented one by one on the computer screen in a random order that was fixed for all participants to minimize interindividual variance due to order effects. Five adjective pairs were selected for assessing *shyness*, *anxiousness*, *angriness*, *agreeableness*, *conscientiousness*, and *intellect*, respectively. The adjective pairs of the last three traits served as distractors and corresponded to the first five of the ten agreeableness-, conscientiousness-, and intellect-descriptive bipolar adjective pairs by Asendorpf et al. (2002) used in their study. The 10 shyness-descriptive adjectives were the same as in the shyness IAT by Asendorpf et al. (2002). The 10 anxiousness- and angriness-descriptive adjectives were the same as in Schnabel's study (unpublished dissertation). Note that the shyness- and anxiousness-descriptive adjectives had an overlap of three identical adjectives.

Extrinsic Affective Simon Task (EAST)

The EAST was designed on the basis of two criteria: high similarity to the original EAST by De Houwer (2003a) and high comparability to the IAT used in the studies by Asendorpf et al. (2002) and Schnabel (unpublished dissertation). Target stimuli were presented as white words and therefore had to be discriminated according to their categorical meaning (*others* versus *me*). Attribute stimuli (*shy* versus *nonshy*, *anxious* versus *secure*, and *angry* versus *self-controlled*) were presented as coloured words and therefore had to be discriminated according to their colour (*blue* versus *green*). In all phases of the EAST, participants were instructed to respond as quickly and accurately as possible.

Table 1 shows the EAST's task sequence and stimuli. In a first practice block, participants discriminated the white stimuli of the category *others* versus *me* by pressing the letter A on the left-hand-side of the keyboard or the number 5 of the right-hand-side numeric keypad in response to the word's categorical meaning, thus defining an 'others key' and a 'me key'. In a second practice block, they discriminated the coloured stimuli of the categories *shy* versus *nonshy*, *anxious* versus *secure*, and *angry* versus *self-controlled* by pressing the same left 'others key' or right 'me key' in response to the word's colour. The tasks were combined in the two identical test blocks, each of which started with two warm-up trials (one white and one coloured word).

Following De Houwer (2003a), all analyses were based on the 110 coloured trials of the test blocks excluding the first two trials and trials with incorrect responses. EAST scores for shyness, anxiousness, and angriness were each calculated as the difference between the mean log-transformed reaction time in the incompatible trials (i.e. due to the stimuli's word colour, participants had to press the 'me key' in response to *nonshy*, *secure*, or *self-controlled* stimuli and the 'others key' in response to *shy*, *anxious*, or *angry* stimuli) and the mean log-transformed reaction time in the compatible trials (i.e. due to the stimuli's word colour, participants had to press the 'me key' in response to *shy*, *anxious*, or *angry* stimuli and the 'others key' in response to *nonshy*, *secure*, or *self-controlled* stimuli), respectively. Thus, positive EAST scores indicate faster and stronger associations between *me* and *shy* (or *anxious*, *angry*, respectively) than between *others* and *shy* (or *anxious*, *angry*, respectively).

Both the 10 target stimuli, presented during the white trials, and the 30 attribute stimuli for shyness, anxiousness, and angriness, presented during the coloured trials, can be found in Table 1. The attribute stimuli were identical with the corresponding bipolar adjectives in the direct self-ratings. Each stimulus was presented at the center of the screen until the participant pressed the 'others key' or the 'me key'. The time between the onset of a stimulus and the first key press was registered as well as whether a correct or incorrect response was given. As in the IAT by Asendorpf et al. (2002), the interstimulus interval was 250 ms; after an incorrect response the word *FEHLER* (German for error) immediately replaced the stimulus for 300 ms, resulting in a 550 ms interstimulus interval.

In order to strengthen the effect that the response keys become extrinsically associated with *others* versus *me*, they were labelled 'others key' and 'me key' in the instructions. For the same reason, the white target words were presented as often as the coloured attribute words in the two test blocks (i.e. 54 times each). In each test block, each attribute word was presented once in blue and once in green colour. In accordance with De Houwer (2003a), the hues of the blue and green colours could be discriminated well but were quite similar. Note that the three identical stimuli for shyness and anxiousness were not presented twice and counted for both the shyness and anxiousness EAST score. Following De Houwer (2003a), within the two identical test blocks of 110 trials, the order of the target and attribute stimuli was randomized with the restriction that the same word was not presented on two or more consecutive trials and that the required response was not the same on four or more consecutive trials. Again, all participants received the same random order. Participants needed approximately 8 minutes to complete the total 279 trials of the EAST, which is comparable to the IAT by Asendorpf et al. (2002) (280 trials).

Following De Houwer (2003a), response latencies below 300 ms were recoded as 300 ms, and latencies above 3000 ms were recoded as 3000 ms. These raw latencies were used only for reporting means and standard deviations, whereas all other analyses were based on log-transformed latencies to correct for the skewed latency distributions.

Table 1. Extrinsic Affective Simon Task (EAST) for shyness, anxiousness, and angriness: task sequence and stimuli

Block	No. of trials	Task	Response key assignment	
			Left key	Right key
1	30[a]	Target's semantic discrimination	Others	Me
2	29[b]	Attribute's colour discrimination	Blue	Green
3	110[c]	First combined task	Others, Blue	Me, Green
4	110[c]	Second combined task	Others, Blue	Me, Green

Stimuli

Target		Attribute					
		Shyness		Anxiousness		Angriness	
Me	Others	Shy	Nonshy	Anxious	Secure	Angry	Self-Controlled
I	They	Inhibited	Uninhibited	Anxious	Secure	Angry	Self-controlled
Self	Them	Insecure	Secure	Fearful	Daring	Hot-tempered	Thoughtful
My	Your	Timid	Daring	Insecure	Self-confident	Under-controlled	Self-disciplined
Me	You	Reticent	Candid	Worried	Unconcerned	Hot-headed	Adaptable
Own	Other	Reserved	Open	Overly cautious	Carefree	Irritable	Calm

The attributes 'secure', 'daring', and 'insecure' were both shyness- and anxiousness-descriptive adjectives and therefore were not presented twice. The original German stimuli can be obtained from the authors.

[a] 3×10 (targets) = 30 trials.

[b] 2 (warm-up) + 7 (shyness) + 7 (anxiousness) + 3 (shyness and anxiousness) + 10 (angriness) = 29 trials.

[c] 2 (warm-up) + 54 (targets) + 7 (blue shyness) + 7 (blue anxiousness) + 3 (blue shyness and anxiousness) + 10 (blue angriness) + 7 (green shyness) + 7 (green anxiousness) + 3 (green shyness and anxiousness) + 10 (green angriness) = 110 trials.

Questionnaire scales

In order to replicate the validation of the bipolar adjectives by Asendorpf et al. (2002) and Schnabel (unpublished dissertation), the *shyness scale* by Asendorpf et al. (2002), a shortened version[1] of the trait form of the *State Trait Anxiety Inventory, STAI* (Laux, Glanzmann, Schaffner, & Spielberger, 1981; English version by Spielberger, Gorsuch, & Lushene, 1970), and the trait form of the *State Trait Anger Expression Inventory, STAXI* (Schwenkmezger, Hodapp, & Spielberger, 1992; English version by Spielberger, 1988) were used for assessing enduring symptoms of shyness, anxiousness, and angriness, respectively. Participants rated the five shyness-descriptive items,[2] 10 STAI items and 20 STAXI items that were introduced with a trait instruction and were presented one-by-one in a fixed random order on a four-point-scale (1 = almost never, 4 = almost always).

Implicit Association Tests (IATs)

Participants completed two IATs consecutively, a shyness IAT that was identical with the IAT by Asendorpf et al. (2002) and an angriness IAT that was identical with Schnabel's (unpublished dissertation) angriness IAT. Target and attribute stimuli were the same as in the EAST and can be found in Table 1. The task sequence is shown in Table 2. First, participants discriminated *me* versus *others*, then *shy* versus *nonshy* (or *angry* versus *self-controlled*, respectively). In the initial combined task they discriminated *me* and *shy* from *others* and *nonshy* (or *me* and *angry* from *others* and *self-controlled*, respectively). Then, they discriminated *others* versus *me*, and finally, in the reversed combined task, *others* and *shy* from *me* and *nonshy* (or *others* and *angry* from *me* and *self-controlled*, respectively). Hence, both IATs were the same with the exception of the attribute dimension (*shy* versus

Table 2. Implicit Association Tests (IATs) for shyness and angriness: task sequence

Block	No. of trials	Task	Response key assignment	
			Left key	Right key
1	40	Target discrimination	Me	Others
2	40	Attribute discrimination	Shy (Angry)	Nonshy (Self-controlled)
3	80	Initial combined task	Me, shy (Angry)	Others, nonshy (Self-controlled)
4	40	Reversed target discrimination	Others	Me
5	80	Reversed combined task	Others, shy (Angry)	Me, nonshy (Self-controlled)

The procedures of the shyness and the angriness IATs were identical. Words in parentheses refer to the attributes used within the angriness IAT. The original German stimuli can be obtained from the authors.

[1]As shyness and anxiousness are similar concepts at the construct level, direct ratings on shyness-descriptive items might activate the anxiousness concept, and direct ratings on anxiousness-descriptive items might activate the shyness concept, thereby influencing consecutive indirect measures (see Strack & Deutsch, in press). Thus, to avoid activating the shyness and anxiousness concept more strongly than the angriness concept, the total number of shyness and anxiousness items should not exceed the number of angriness items. Therefore, the STAI, originally containing 20 items, was shortened by selecting the 10 items with the highest item–total correlation (above 0.57) based on Schnabel's (unpublished dissertation) data, which nonetheless led to a satisfactory internal consistency ($\alpha = 0.87$).
[2]In the study by Asendorpf et al. (2002), participants responded to the shyness scale on a five-point-scale format (1 = not at all true for me, 5 = completely true for me). In order to randomize the items of the shyness scale, the STAI, and the STAXI, the format of the shyness scale was adapted to the four-point-scale format of the STAI and the STAXI.

nonshy within the shyness IAT, *angry* versus *self-controlled* within the angriness IAT). To allow replication of the position effect that was found for the anxiousness and the angriness IATs (Schnabel, unpublished dissertation), the order of the shyness and the angriness IATs was counterbalanced across participants.

IAT scores were calculated as the difference between the mean log-transformed reaction time in the reversed combined task fifth block) and the mean log-transformed reaction time in the initial combined task (third block). Thus, high IAT scores indicate faster associations between *me* and *shy* (or *angry*, respectively) than between *others* and *nonshy* (or *self-controlled*, respectively).

In order to maximize the comparability of the present study with the studies by Asendorpf et al. (2002), IAT scores were based on the by Asendorpf et al. (2002) procedure. The first two responses in the combined tasks were not analysed, response latencies below 300 ms were recoded as 300 ms, and latencies above 3000 ms were recoded as 3000 ms. These raw latencies were used only for reporting means and standard deviations; all other analyses were based on log-transformed latencies to correct for the skewed latency distributions. To compare the present IAT data with Schnabel's (unpublished dissertation) IAT data, analyses of the latter were also based on the by Asendorpf et al. (2002) procedure. Note that in accordance with Schnabel's (unpublished dissertation) findings, IAT analyses based on the improved IAT scoring procedure recently suggested by Greenwald et al. (2003) revealed only minimal changes in the results compared with the by Asendorpf et al. (2002) procedure (differences in correlations below 0.07).

Debriefing

Finally, participants were thoroughly debriefed about the true purpose of the study. They were informed that the study was on shyness, anxiousness, and angriness and aimed at validating new computer based measures for these traits. In order to keep the true purpose of the study undisclosed for subsequent participants, participants were asked to keep the information about the study confidential until data collection was completed, at which point they would receive a letter about the general findings.

RESULTS

This section starts with the descriptive statistics for the indirect and direct measures. Second, reliabilities of the IATs and correlations between the IATs and the direct measures are presented in order to ensure the suitability of the present data for evaluating the EAST's psychometric properties by replicating the findings by Asendorpf et al. (2002) and Schnabel (unpublished dissertation). Finally, the reliability of the EAST and its correlations with both its corresponding and non-corresponding direct and indirect measures are shown.

Descriptive statistics of the main variables

EAST

For the EAST, the individual incorrect response rates for the 108 analysed responses to coloured trials in the two test blocks were similar to those regularly reported for indirect measures such as IAT or IAP, $M = 4.14\%$, $SD = 3.48\%$. Following De Houwer (2003a), two participants were excluded from all analyses because their error rates (19 and 18%

Table 3.   Descriptive statistics for the main variables

Variable	M	SD	Range
Shyness EAST score	62 ms	148 ms	−356–477
Anxiousness EAST score	−84 ms	137 ms	−464–233
Angriness EAST score	−57 ms	151 ms	−460–329
Shyness IAT	−57 ms	157 ms	−484–321
Angriness IAT	−176 ms	140 ms	−550–238
Bipolar shyness self-rating (1–7)	3.56	1.06	1.80–6.00
Bipolar anxiousness self-rating (1–7)	3.77	0.89	1.60–6.00
Bipolar angriness self-rating (1–7)	3.37	0.94	1.60–6.40
Shyness scale (1–4)	2.01	0.65	1.00–4.00
STAI (1–4)	1.91	0.54	1.00–3.50
STAXI (1–4)	1.92	0.42	1.00–3.30

$N = 97$; one participant less for both IATs. M, SD, and range refer to raw scores in the case of the EAST and IAT latencies.
EAST = Extrinsic Affective Simon Task.
IAT = Implicit Association Test.
STAI = State–Trait Anxiety Inventory (trait form).
STAXI = State–Trait Anger Expression Inventory (trait form).

errors, respectively) were more than three standard deviations higher than the group mean. Note, however, that excluding their data did not change the results. The following analyses refer to 97 participants (49 male, 48 female; age $M = 24.0$ years, range 20–31 years).[3]

Participants' performance during coloured trials did not differ between the left-hand 'others key' ($M = 919$ ms, SD $= 216$ ms) and the right-hand 'me key' ($M = 916$ ms, SD $= 215$ ms), $t(96) = 0.30$, n.s. Thus, participants were not faster in giving responses associated with *others* compared with responses associated with *me*. Table 3 shows the descriptive statistics for the main variables. The SDs for the EAST scores were high, and the distributions of the three EAST scores based on log-transformed latencies were not different from a normal distribution, $Z < 1$.

IATs

For the shyness IAT, the individual incorrect response rates for the 156 analysed responses in the two combined tasks were slightly lower than those reported by by Asendorpf et al. (2002), $M = 3.96\%$, SD $= 2.50\%$, $t(232) = 2.06$, $p < 0.05$. Error rates of the angriness IAT were similar to those reported by Schnabel (unpublished dissertation), $M = 4.09\%$, SD $= 2.59\%$. Inspection of the error distributions indicated one clear outlier (a participant with 22% errors in the shyness IAT and 24% errors in the angriness IAT). All other error rates were below 12 and 13%, respectively. Therefore, the IAT data of this participant were excluded from all analyses.

As Table 3 shows, both mean IAT scores were below zero (again, for possible reasons, see Asendorpf et al., 2002) and were similar to those reported by Asendorpf et al. (2002)

[3]De Houwer (2003a) suggested calculating EAST scores based on both the differences between reaction times and the differences between error rates. Because in the present study error rates and variances of the trait-describing coloured words were uniformly low ($M < 5.44\%$, SD $< 5.14\%$), EAST scores based on error rates were regarded as less suitable for analyses and are therefore not considered in the following. Note, however, that analyses that included these error based EAST scores led to the same conclusions regarding the psychometric properties of the EAST as analyses that did not include them: the internal consistency α was extremely low (for the shyness EAST score, −0.01, for the anxiousness EAST score, 0.11, and for the angriness EAST score, 0.20) and therefore not acceptable. Furthermore, EAST scores based on error rates correlated neither with their corresponding IATs and direct measures nor with their corresponding EAST scores based on reaction times.

and Schnabel (unpublished dissertation), $|t| < 0.65$, thus confirming Hypothesis 1. SDs for the IAT scores were high, indicating a good discrimination between participants. The distributions of both IAT scores based on log-transformed latencies were not different from a normal distribution, $Z < 1$.

The IAT scores were significantly less positive than the EAST scores for shyness, $t(95) = 6.07$, $p < 0.001$, and angriness, $t(95) = 7.04$, $p < 0.001$. The reason for this difference may be that the two mixed blocks of the IATs were not balanced within participants and thus confounded with an order effect. If the participants were faster in the reversed combined block than in the initial combined block due to a better handling of the IAT task, this would decrease the reaction time in the reversed block and therefore the IAT score. It should be noted, however, that mean scores are irrelevant for the present analyses of interindividual differences.

Direct measures
Consistent with Hypothesis 1, the means of the bipolar shyness self-ratings were not different from those of the study by Asendorpf et al. (2002), $t < 1$. Likewise, the means of the bipolar anxiousness self-ratings as well as the means of the STAI and STAXI were not different from those of Schnabel's (unpublished dissertation) study, $|t| < 1.5$. However, participants in the present study judged themselves as angrier on the bipolar angriness self-ratings than those of Schnabel's (unpublished dissertation) study, $t(195) = 3.16$, $p < 0.01$, $d = 0.45$. However, as neither the angriness IAT nor the STAXI indicated that Schnabel's (unpublished dissertation) and the present sample differed with regard to angriness, this finding might be due to chance. The internal consistencies of all direct measures were satisfactory, $\alpha > 0.74$, as indicated by Table 4, showing internal consistencies and intercorrelations of the main variables.

All in all, Hypothesis 1 was confirmed, as—with one exception—all means of the IATs and the direct measures of the study by Asendorpf et al. (2002) and Schnabel (unpublished dissertation) could be replicated. Therefore, it can be assumed that the sample of the present study was not differently selected from the samples of Asendorpf et al. (2002) with regard to shyness and Schnabel (unpublished dissertation) with regard to anxiousness and angriness.

Replications of the IATs' psychometric properties

To test whether the present data replicate findings by Asendorpf et al. (2002) and Schnabel (unpublished dissertation), first, internal consistencies of the IATs are analysed. Then, their convergent and discriminant validities are examined.

Reliability of the IATs
Following Asendorpf et al. (2002), for both IATs, Cronbach's α was determined using four IAT scores. The scores were calculated using the differences between the combined blocks for trials 3–20, 21–40, 41–60, and 61–80, respectively. As expected from Hypothesis 2, Cronbach's α for the shyness IAT was acceptable ($\alpha = 0.77$) and nearly the same as in the study by Asendorpf et al. (2002). The internal consistency α for the angriness IAT was nearly the same as in Schnabel's (unpublished dissertation) study and also acceptable ($\alpha = 0.76$).

Convergent and discriminant validity of the IATs
Consistent with Hypothesis 2, both IATs were positively correlated with their corresponding direct measures (see Table 4). Correlations between the shyness IAT and

Table 4. Reliabilities and intercorrelations of the main variables

	1	2	3	4	5	6	7	8	9	10	11
1. Shyness EAST score	*0.19*										
2. Anxiousness EAST score	0.14	*0.24*									
3. Angriness EAST score	0.03	0.12	*0.19*								
4. Shyness IAT	0.01	−0.19	−0.15	*0.77*							
5. Angriness IAT	0.02	0.04	−0.04	0.13	*0.76*						
6. Bipolar Shyness self-rating	−0.07	−0.06	−0.12	0.30**	−0.05	*0.85*					
7. Bipolar Anxiousness self-rating	0.04	−0.04	−0.11	0.11	0.01	0.69***	*0.74*				
8. Bipolar Angriness self-rating	−0.07	−0.10	0.04	−0.03	0.39***	0.02	0.16	*0.76*			
9. Shyness scale	0.01	0.11	−0.04	0.29**	−0.07	0.81***	0.55***	−0.07	*0.86*		
10. STAI	0.07	−0.01	−0.16	0.25*	0.00	0.54***	0.61***	0.08	0.55***	*0.87*	
11. STAXI	−0.11	−0.01	−0.01	−0.07	0.34***	0.13	0.34***	0.57***	0.11	0.26**	*0.74*

$N = 97$; one participant less for both IATs, one degree of freedom less for both IATs; internal consistencies (Cronbach's α) are printed in *italics* along the diagonal. EAST = Extrinsic Affective Simon Task; IAT = Implicit Association Test; STAI = State–Trait Anxiety Inventory (trait form); STAXI = State–Trait Anger Expression Inventory (trait form).
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

both the bipolar shyness self-rating ($r = 0.30$) and the shyness scale ($r = 0.29$) were comparable to those in the study, by Asendorpf et al. (2002), $z < 1.21$. Contrary to Schnabel's (unpublished dissertation) findings, the present study demonstrated convergent validity of the angriness IAT, which correlated with the direct angriness measures at expected levels. Its correlation with the bipolar angriness self-rating tended to be higher than in Schnabel's (unpublished dissertation) study (0.39 versus 0.13), $z = 1.94$, $p < 0.10$ (two tailed), and its correlation with the STAXI was significantly higher (0.34 versus 0.04), $z = 2.16$, $p < 0.05$ (two tailed). The shyness IAT showed discriminant validity, as it correlated neither with direct angriness measures nor with the angriness IAT, but correlated with the STAI ($r = 0.25$), which is consistent with the similarity of shyness and anxiousness at the construct level. The angriness IAT correlated neither with direct shyness nor with direct anxiousness measures. Also, in line with Schnabel (unpublished dissertation), correlations between the shyness and angriness IATs tended to be higher when the shyness IAT was completed first, 0.34 versus -0.01, $z = -1.73$, $p < 0.10$ (two tailed). All in all, Hypothesis 2 was confirmed, as both IATs showed satisfactory psychometric properties.

Altogether, in the present study, findings by Asendorpf et al. (2002) and Schnabel (unpublished dissertation) were by and large replicated. Hence, data can be considered suitable to evaluate the psychometric properties of the multiple trait EAST.

Psychometric properties of the EAST

Reliability of the EAST

Following De Houwer (2003a) who reported split half reliabilities, for each the shyness, anxiousness, and angriness EAST score, Cronbach's α was determined using two EAST scores. The scores were separately calculated for the two test blocks using the differences between the 10 incompatible shyness, anxiousness, and angriness trials and the 10 compatible shyness, anxiousness, and angriness trials, respectively. As Table 4 indicates, the internal consistencies for the shyness, the anxiousness, and the angriness EAST scores were disappointingly low ($\alpha < 0.24$). Thus, the EAST clearly failed to meet the requirement of a satisfactory reliability, which is demanded from a measure aiming at assessing interindividual differences. Accordingly, Hypothesis 3, which expected the internal consistencies of the shyness, anxiousness, and angriness EAST scores to reach approximately 0.80, had to be rejected.

Convergent validity of the EAST

As shown in Table 4, all correlations of the three EAST scores were not significant. Contrary to Hypothesis 4, the shyness EAST score correlated neither with the shyness IAT nor with the bipolar shyness self-ratings nor with the shyness scale, $|r| < 0.07$. The anxiousness EAST score correlated neither with the bipolar anxiousness self-ratings nor with the STAI, $|r| < 0.04$. The angriness EAST score correlated neither with the angriness IAT nor with the bipolar angriness self-ratings nor with the STAXI, $|r| < 0.04$. Thus, the EAST clearly missed the expected correlations of 0.50 with its corresponding IATs, $z > 3.68$, $p < 0.001$ (two tailed). Correlations between the EAST and its corresponding direct measures were expected to be highly similar to the IATs' correlations with their corresponding direct measures, but were uniformly significantly lower, $z > 1.97$, $p < 0.05$ (two tailed). These very low convergent correlations were also found when we computed them separately for the first and second halves of the EAST (for both the first and the second half, $|r| < 0.10$).

Discriminant validity of the EAST

Consistent with Hypothesis 5 concerning the EAST's discriminant validity, the shyness EAST score correlated neither with the indirect angriness measures (angriness IAT, angriness EAST score) nor with the direct angriness measures (bipolar angriness self-rating, STAXI). However, the shyness EAST score correlated neither with the direct anxiousness measures, $|r| < 0.07$, nor with the anxiousness EAST score, $r = 0.14$. The same pattern was found for the anxiousness EAST score: consistent with Hypothesis 5, it correlated neither with the indirect angriness measures (angriness IAT, angriness EAST score) nor with the direct angriness measures (bipolar angriness self-rating, STAXI). However, inconsistent with Hypothesis 5, the anxiousness EAST score was not correlated with the direct shyness measures, $|r| < 0.11$, and tended to correlate even negatively with the shyness IAT, $r = -0.19$, $p < 0.10$, which clearly contradicts the assumption of shyness and anxiousness being similar at the construct level. In line with Hypothesis 5, the angriness EAST score correlated neither with the shyness IAT nor with the direct shyness and anxiousness measures (bipolar shyness self-rating, shyness scale, bipolar anxiousness self-rating, STAI). All in all, non-significant correlations between the EAST and its non-corresponding indirect and direct measures were expected, but non-significant correlations between the shyness EAST score and direct and indirect anxiousness measures as well as between the anxiousness EAST score and direct and indirect shyness measures were not expected. Therefore, Hypothesis 5 concerning the discriminant validity of the EAST had to be rejected.

Taken together, all hypotheses concerning the EAST's psychometric properties had to be rejected. As data revealed suitability for such an evaluation, it has to be concluded that the present EAST for the simultaneous assessment of shyness, anxiousness, and angriness is neither reliable nor convergently or discriminantly valid.

DISCUSSION

This study tested five hypotheses with the main purpose of evaluating the psychometric properties of a newly developed indirect procedure, the Extrinsic Affective Simon Task (EAST), which was applied to the assessment of interindividual differences in multiple traits. The following section starts with a discussion of the results, separately for each hypothesis. Then, the main problem is discussed in detail. Finally, general conclusions regarding the indirect assessment of the personality self-concept by EASTs are formulated.

Replications of earlier findings

All hypotheses concerning the replication of recent findings by Asendorpf et al. (2002) and Schnabel (unpublished dissertation) were by and large confirmed. As expected by Hypothesis 1, the means of the Implicit Association Tests (IATs) and the direct measures could be replicated. Thus, the present sample was not differently selected from the sample by Asendorpf et al. (2002) with regard to shyness and Schnabel's (unpublished dissertation) samples with regard to anxiousness and angriness. Hypothesis 2 was fully confirmed, as both the shyness and angriness IATs showed satisfactory psychometric properties, replicating findings of earlier studies (Asendorpf et al., 2002; regarding reliability, Schnabel, unpublished dissertation). All in all, these results revealed the suitability of the present data for evaluating the psychometric properties of the EAST.

Psychometric properties of the EAST

All hypotheses concerning the psychometric properties of the EAST had to be rejected. Reliability analyses showed that, contrary to Hypothesis 3, the internal consistencies of the shyness, anxiousness, and angriness EAST scores were low, $\alpha < 0.24$, and thereby far lower than the expected α of 0.80. Importantly, the internal consistencies of the three EAST scores were not only much lower than the Cronbachs' α usually reached by IATs (and Implicit Association Procedures (IAPs) as well), but were not even equal to other indirect procedures such as priming methods and projective tests, which reach moderate reliabilities (see e.g. Cunningham et al., 2001; Kawakami & Dovidio, 2001). Thus, as the EAST clearly missed common standards of reliability, it seems to be unsuitable for the assessment of interindividual differences in implicit traits.

As reliability is a necessary prerequisite for validity, the findings of the EAST's validity are not surprising in view of its low internal consistency: In contrast to Hypothesis 4 concerning the EAST's convergent validity, the shyness, anxiousness, and angriness EAST scores showed no significant correlations with their corresponding indirect and direct measures. Hypothesis 5 concerning the discriminant validity of the EAST had to be rejected as well. Despite the similarity of shyness and anxiousness at the construct level, the shyness EAST score did not correlate with direct and indirect anxiousness measures, and the anxiousness EAST score did not correlate with direct and indirect shyness measures. This lack of discriminant validity is even more remarkable as the stimulus material of bipolar adjectives for the shyness and anxiousness measures overlapped for three identical items, which actually should increase correlations. With regard to these analyses, the zero correlations between the shyness, anxiousness, and angriness EAST scores and their non-corresponding indirect and direct measures cannot be interpreted as indices of discriminant validity.

Undoubtedly, analysing behavioral correlates in the sense of a double dissociation strategy for showing an indirect measure's convergent and discriminant validity would have been more convincing (e.g. for the problem of spurious discriminant validities see Asendorpf et al., 2002). However, interpretations of the present correlational analyses are so clear that such behavioral correlates can be expected to lead to the same conclusions.

Given the successful replication of findings by Asendorpf et al. (2002) and Schnabel (unpublished dissertation) in this study, it is unlikely that the findings concerning the EAST's psychometric properties are caused by problems of the present sample or data. Thus, it has to be concluded that the present multiple trait EAST is neither a reliable nor a convergently or discriminantly valid measure. Because reliability is a necessary requirement for validity, the central question for further reflection on the outcome of this study is the low reliability of the EAST scores.

Low reliability

Importantly, low reliability has not only been a problem of the present EAST. Even the standard EASTs (De Houwer, 2003a) showed a mean split half reliability of 0.25, which is equivalent to an internal consistency of approximately 0.40 (the self-esteem EAST reached the highest split half reliability of 0.48, approximately $\alpha = 0.65$). Although reliabilities of the standard EASTs are clearly higher than for the present EAST, they are not satisfactory. Hence, reliability appears to be a serious problem for the EASTs *in general*. There are four arguments that might explain this general reliability problem: (i) *insufficient representativeness* of the stimuli, (ii) *low interindividual variance* (e.g. participants do not differ in

their attitudes toward specific target objects), (iii) *procedural elements* varying between participants, and (iv) a *low aggregation level* of the EAST scores.

The first aspect refers to analyses of the mechanisms underlying IATs versus priming procedures (Fazio & Olson, 2003), which can be partly transferred to EASTs. Several studies showed that IATs more likely operate at the level of the category labels, whereas priming measures operate at the level of the exemplars, that is, stimuli of these category labels (e.g. De Houwer, 2001; Fazio & Olson, 2003; Neumann, Totzke, Popp, & Fernandez, unpublished dissertation; but see Mitchell, Nosek, & Banaji, 2003). Thus, IATs seem to assess associations to category labels, whereas priming measures assess automatically activated responses to the individual exemplars of these category labels. For EASTs, the mechanisms can be assumed to be similar to priming procedures, because EAST scores are more likely influenced by the nature of the selected stimuli than by the categories (see Experiment 3 of De Houwer, 2002). As a result of this higher sensitivity to the individual stimuli, the stimuli's representativeness of the concept plays a greater role for the measure's reliability and validity (see e.g. Livingston & Brewer, 2002; Lord, Lepper, & Mackie, 1984). Low representativeness, which can be based on bad stimulus selection or the heterogeneity of the concepts, leads to low item–total correlations, therefore lowering the internal consistency. It also lowers validity, as the EAST score for a particular concept more likely reflects the mean of the attitudes toward the specific exemplars selected to represent the concept than the attitude toward the concept itself. Hence, the EASTs' low reliability and validity might be caused by an insufficient representativeness of the EASTs' stimulus material. Therefore, further EAST studies should carefully select the stimuli according to their representativeness of the category of interest (for some methods, see Nosek et al., unpublished manuscript). Note, however, that in the present study the EAST's stimulus material, i.e., the bipolar adjectives for shyness, anxiousness, and angriness, seemed to be representative at least at an explicit level: item–total correlations of all bipolar adjective pairs were satisfactory, $r_{it} > 0.47$, with one exception for anxiousness, $r_{it} = 0.27$, leading to acceptable internal consistencies, $\alpha > 0.74$. Moreover, the EAST's stimulus material was validated at an explicit level. Finally, a bad stimulus selection should have at least partially affected the shyness and angriness IATs, but these IATs were both reliable and valid.

De Houwer (2003a) himself suggested the other three explanations for the low reliability (ii–iv). Although these arguments might be true for the standard EASTs in De Houwer's (2003a) studies, none of them explain the low internal consistency of the present EAST. Contrary to (ii), participants differed from each other with respect to shyness, anxiousness, and angriness, as data analyses revealed. Thus, in the present study the traits showed interindividual variance. Contrary to (iii), procedural elements of the EAST such as task assignments and trial order were kept constant across participants. Thus, as all participants responded to the identical task, the measurement of interindividual differences was optimized. Consistent with (iv), the aggregation level of the EAST scores was low because the EAST was designed very similarly to the IATs. Therefore, the EAST assessed three traits simultaneously with about the same number of trials as an IAT assessing just one trait. That led to different aggregation levels of the EAST scores versus IAT scores: each of the shyness, anxiousness, and angriness EAST scores were calculated by the mean of 20 difference scores, while each IAT score was calculated by the mean of 78 difference scores.

Despite its low aggregation level, the EAST should show a satisfactory internal consistency, as internal consistencies of both IATs tended to be acceptable, even if the aggregation level of their scores was adjusted to the EAST's aggregation level. Analysing

the data with IAT scores that were calculated based on difference scores for the first 20 trials of both combined blocks revealed only minimal changes in the results compared with the original IAT scores: the internal consistencies were just slightly lower (0.67 for both the shyness and angriness IATs), and differences in correlations were below 0.10. Thus, there is evidence that in the present study the EAST's low internal consistency is not mainly a result of the low aggregation level. Besides, as completing an EAST requires high concentration, the total number of its trials is limited: if an EAST assessing three traits needed three times as many trials as an IAT assessing one trait, the EAST would become very long. The longer it takes to complete the EAST, the more likely it promotes boredom and fatigue (Fazio & Olson, 2003), which might distort results. Therefore, the requirements for multiple trait EASTs are particularly strong.

As De Houwer's (2003a) arguments fail to explain the low reliability of the present EAST, is the conclusion justified that the EAST is an unreliable measure in general and therefore unsuitable for differential applications? Probably not, because (i) the standard EAST showed clearly higher although not satisfactory reliabilities in comparison with the present trait EAST, and (ii) recently Ellwart, Becker, and Rinck (manuscript submitted for publication) successfully applied an EAST version that discriminated well between spider fearful and non-fearful participants, $d = 0.77$. Their EAST version differed from the standard and the present EAST, as it was preceded by the priming of a spider concept (for further details, see Ellwart et al., manuscript submitted for publication). This spider primed EAST showed encouraging psychometric properties: the split half reliability of 0.75 (Ellwart, personal communication, August, 2003) was as convincing as the convergent validity (correlation with a spider questionnaire of 0.31). Thus, the EAST *can* be a more reliable measure.

Therefore, it seems to be a promising approach to derive reasons for the particularly low internal consistency of the present EAST from a comparison between the standard EAST (De Houwer, 2003a) or primed EAST (Ellwart et al., 2003) and the present multiple trait EAST. As one of the reviewers of an earlier version of this article noted, such a comparison yields at least three differences: (i) it is assumed that trait-descriptive attributes such as 'shy' automatically activate the meaning of 'describes me' or 'describes others' if *me* and *others* are made salient in the task (in the original self-esteem EAST, automatic activation is assumed for valence that is made salient in the task); (ii) resting on assumption (i), it is assumed that using (non)self words such as 'I' or 'you' facilitates or inhibits responses to (non)self-related attributes such as 'shy' even if these attributes are categorized for an arbitrary feature such as colour (the particular variant of the Simon effect in this case); (iii) this particular Simon effect also applies to neutral responses such that they acquire the meaning of '(non)self'. Although these assumptions seem to be reasonable ones (see, e.g., Bargh, 1994), violation of one of the three can render the present EAST effect uninterpretable and can be a reason for its low reliability.

Below we discuss these and additional potentially critical points of the present EAST procedure that concern (i) *reading* of the coloured words, (ii) *automatically activated associations* in response to the coloured words, (iii) the *interstimulus interval*, and (iv) the *reversed discrimination tasks*.

Reading of the coloured words

The supposed necessity of the word's reading is based on the assumption that only *read* coloured words can produce automatically activated associations, which the EASTs aim at assessing. That is, despite the instruction to concentrate only on colour when

discriminating coloured words, participants are assumed to read the coloured words. However, in the present study, half of the participants stated in the post-experimental interview that they in fact had used the strategy to concentrate only on colour to facilitate the EAST's completion. This might indicate that participants were so successful in concentrating on colour that they did not even read the coloured words. However, there is no reason why participants in the present study should be able to prevent reading, whereas participants in the studies by De Houwer's (2003a) and Ellwart et al. (manuscript submitted for publication) were evidently not. Indeed, there is strong evidence for the participants' reading of the coloured words.

Early on, several studies showed that words are read faster than colours are named (e.g. Cattell, 1886; Fraisse, 1969). This effect also underlies the well known Stroop effect (for an overview of the Stroop effect, see MacLeod, 1991) and is expected to be due to a higher automaticity of reading in comparison with colour naming. Therefore, it can be assumed that even concentrating on the words' colour does not prevent the reading of the words. This assumption is confirmed by analyses of the present data: the EAST scores showed strong effects (for the shyness EAST score, $d = 0.90$; for the anxiousness EAST score, $d = 1.47$; for the angriness EAST score, $d = 0.92$). Furthermore, data analyses revealed an emotional Stroop effect for the EAST: performance on discriminating the coloured trait-describing words differed in dependence on their poles (for an overview of the emotional Stroop task, see Williams, Mathews, & MacLeod, 1996). That is, participants were faster in giving responses to the *Nonshy, Secure,* and *Angry* stimuli in comparison with the *Shy, Anxious,* and *Self-Controlled* stimuli (for shyness, $d = 0.88$; for anxiousness, $d = 0.44$; for angriness, $d = 0.88$).[4]

Automatically activated associations in response to the coloured words
Because failure to read the coloured words does not seem to be a problem, a second process assumed to underlie the EASTs' functioning seems to fail: the automatic activation of the interesting associations. Importantly, the discrimination dimension underlying the associations differed between the compared EASTs: in contrast to both the standard EASTs and the primed EASTs, in the present EAST, participants discriminated the white words not according to valence, but according to categorical meaning. However, the functioning of the EAST procedure might depend on valence as a necessary discrimination dimension. The standard EAST evidently assesses automatic activations of an *evaluation* (valence) associated with the coloured stimuli that lead to a processing advantage for evaluatively compatible trials in comparison with evaluatively incompatible trials (De Houwer, 2003a). That is, as associations with valence are assumed to be very strong (Bargh, 1994; Greenwald, Draine, & Abrams, 1996), participants within the standard or primed EAST automatically associate the read coloured words with a positive versus negative valence. Consequently, depending on compatible or incompatible trials, participants give faster or slower key responses, which is expressed in the EAST scores.

However, associations of an automatically activated *categorical meaning* such as the concept of self or the concept of others in response to the coloured trait stimuli might be so

---

[4]Concerning angriness, the results seem to contradict the assumption of the emotional Stroop task that discriminating stimuli with a negative valence takes longer than discriminating stimuli with a positive valence (see e.g. De Houwer, 2003b; Pratto & John, 1991, but see Wentura, Rothermund, & Bak, 2000). Although angriness was assumed to be negatively evaluated, it was discriminated faster. However, this result might be due to an even more negative valence of *Self-Controlled*, the bipolar counterpart of *Angry*. That is, although the valence of both traits might be negative, the valence of *Self-Controlled* could be even more negative than that of *Angry*.

weak that they do not lead to the expected processing advantage for semantically compatible trials in comparison with semantically incompatible trials. That is, these associations might be less obtrusive, and therefore participants might have weaker or even no automatic associations with the *me* versus *others* meaning in response to the read coloured words. As a result, the multiple trait EAST shows unsatisfactory reliability and validity.

Because valence as discrimination dimension cannot be integrated into a trait EAST, other solutions are needed to overcome this problem. One promising approach might be to force participants to process the coloured words of the trait EAST semantically (Egloff & Schmukle, personal communication, May 2003), because reading the coloured words does not necessarily imply their semantic processing. However, exactly that might be the prerequisite for the automatic activation of the interesting associations between the coloured (trait) words and their categorical meaning. Therefore, a non-evaluative discrimination feature other than colour should be selected that forces semantic processing. Egloff and Schmukle (personal communication, May 2003) considered the discrimination feature *grammatical category* (e.g. noun versus adjective) as suitable (for the usage of this feature in the Affective Simon Task, see also De Houwer & Eelen, 1998).

Interstimulus interval
Another difference between the standard and primed EASTs versus the present trait EAST concerns the interstimulus interval. In the present EAST, it was of the same duration as in the IATs (250 ms), because the indirect measures' completion should take similarly long for the reason of high comparability. As a result, the interstimulus interval was much briefer than that of the standard and the primed EAST (1500 ms). However, the interstimulus interval moderates the degree of carryover from one trial to another (see Mierke & Klauer, 2001). Carryover should be higher when there is less time for a passive decay of activation. As the present EAST tried to assess three traits simultaneously on the basis of such a brief interstimulus interval, the three concepts might be simultaneously activated. It is likely that this simultaneous activation leads to interference effects between the concepts, which might result in low reliability due to high measurement errors. Thus, further EAST studies should experimentally manipulate the interstimulus interval in order to find out more about its effects on the EAST scores, the EAST's reliability and validity.

Reversed discrimination tasks
Another difference between the standard and primed EAST versus the present trait EAST concerns the discrimination tasks. In the present EAST, the discrimination tasks of the target and the attributes had to be reversed in order to assess the three traits simultaneously. However, there are some arguments against this inversion suggesting the assessment of only one trait, i.e. *either* shyness *or* anxiousness *or* angriness. First, one could benefit from the EASTs' advantage of assessing concepts separately and independently from their counterpart, because the discrimination tasks of the target and the attribute concepts would not need to be reversed (see above). Therefore, identical with the standard and the primed EAST, the response keys would become associated with the attribute concept (e.g. 'nonshy key' versus 'shy key') instead of the target concept ('others key' versus 'me key'). As a result, the association strength between *me* and the trait (e.g. *shy*) could be assessed separately, regardless of the association strength between *others* and the same trait. In the case of trait EASTs, however, benefiting from the advantage of assessing simple association strengths rules out the EASTs' advantage of assessing multiple traits simultaneously. Second, comparable to Ellwart et al. (manuscript submitted for publication), a single-trait EAST could be preceded by a priming of the interesting trait,

which might increase the EAST's reliability and validity (e.g. for the trait of speech anxiety, the primes might be pictures of a microphone, an auditorium, or a speaker's desk).

Conclusions

What does the present study contribute to the initial question of how personality traits can be best assessed through indirect methods? All in all, the newly developed EAST methodology seemed to provide a promising indirect measurement tool, overcoming important problems of the IATs. However, the present findings clearly dash these hopes. The multiple trait EAST did not meet general standards of psychometric properties that are required in the field of personality assessment. Thus, the present study confirmed De Houwer's (2003a) misgivings that 'EAST scores are not reliable enough to detect interindividual differences' (De Houwer, 2003a, p. 83), which seems to be even more true for trait EASTs.

So far, even indirect measures such as the IATs or IAPs that reach an acceptable internal consistency seem to be inferior to direct measures with regard to parallel test reliability, test–retest reliability, and convergent validity. Their psychometric properties suffer from the relatively high amount of specific method variance (see Schnabel, unpublished dissertation). Nevertheless, indirect measures seem to provide a unique access to the implicit personality self-concept and were shown to add incremental validity to the prediction of behaviour (see e.g. Asendorpf et al., 2002; Dovidio, Kawakami, & Gaertner, 2002; Egloff & Schmukle, 2002; McConnell & Leibold, 2001). Therefore, applying indirect measures appears to be a worthwhile, perhaps even necessary complementary approach to using traditional direct measures. However, as long as indirect measures do not meet the needed psychometric requirements, they are not applicable to the fields of applied psychology. Therefore, further research on indirect measures should concentrate on improving these methodological weaknesses to reach a psychometrically satisfactory level. For this purpose, reliable indirect measures should be developed in order to increase the reliability and validity of the assessment of the implicit personality self-concept. With regard to the EAST, we have to await future studies that attempt to increase the multiple trait EASTs' reliability through enforced semantic processing, as proposed by Egloff and Schmukle, or to develop at least single trait EASTs that are sufficiently reliable by an increased number of trials, a prolonged interstimulus interval, and/or prior priming.

ACKNOWLEDGEMENTS

REFERENCES

Asendorpf, J. B., Banse, R., & Mücke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality and Social Psychology, 83*, 380–393.

Banse, R. (1999). Automatic evaluation of self and significant others: Affective priming in close relationships. *Journal of Social and Personal Relationships, 16*, 803–821.

Banse, R. (2001). Affective priming with liked and disliked persons: Prime visibility determines congruency and incongruency effects. *Cognition and Emotion, 15*, 501–520.

Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift für Experimentelle Psychologie, 48*, 145–160.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer, & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 1, pp. 1–40). Hillsdale, NJ: Erlbaum.

Beringer, J. (1994). *Experimental Run Time System* (Version 3.18). Frankfurt: BeriSoft.

Blair, I., & Banaji, M. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology, 70*, 1142–1163.

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology, 79*, 631–643.

Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology, 81*, 760–773.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Cattell, J. M. (1886). The time it takes to see and name objects. *Mind, 11*, 63–65.

Chaiken, S., & Trope, Y. (1999). *Dual process theories in social psychology*. New York: Guilford.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measure: Consistency, stability, and convergent validity. *Psychological Science, 12*, 163–170.

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology, 37*, 443–451.

De Houwer, J. (2003a). The Extrinsic Affective Simon Task. *Experimental Psychology, 50*, 77–85.

De Houwer, J. (2003b). A structural analysis of indirect measures of attitudes. In J. Musch, & K. C. Klauer (Eds.), *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 219–244). Mahwah, NJ: Erlbaum.

De Houwer, J., & Eelen, P. (1998). An affective variant of the Simon paradigm. *Cognition and Emotion, 12*, 45–61.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82*, 62–68.

Egloff, B., & Schmukle, S. C. (2002). Predictive validity of an Implicit Association Test for assessing anxiety. *Journal of Personality and Social Psychology, 83*, 1441–1455.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology, 71*, 390–405.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69*, 1013–1027.

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54*, 297–327.

Fraisse, P. (1969). Why is naming longer than reading? *Acta Psychologica, 30*, 96–103.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*, 4–27.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellot, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109*, 3–25.

Greenwald, A. G., Draine, S. C., & Abrams, R. L. (1996). Three cognitive markers of unconscious semantic activation. *Science, 273*, 1699–1702.

Greenwald, A. G., & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology, 79*, 1022–1038.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*, 1464–1480.

Greenwald, A. G., & Nosek, B. A. (2001). Health of the Implicit Association Test at age 3. *Zeitschrift für Experimentelle Psychologie, 48*, 85–93.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197–216.

Hogarth, R. M. (2001). *Educating intuition*. Chicago, IL: The University of Chicago Press.

Kawakami, K., & Dovidio, J. F. (2001). The reliability of implicit stereotyping. *Personality and Social Psychology Bulletin, 27*, 212–225.

Laux, L., Glanzmann, P., Schaffner, P., & Spielberger, C. D. (1981). *Das State–Trait-Angstinventar (The State-Trait-Anxiety-Inventory)*. Weinheim: Beltz.

Lieberman, M. D. (2000). Intuition: A social cognitive neuroscience approach. *Psychological Bulletin, 126*, 109–137.

Livingston, R. W., & Brewer, M. B. (2002). What are we really priming? Cue-based versus category-based processing of facial stimuli. *Journal of Personality and Social Psychology, 82*, 5–18.

Lord, C. G., Lepper, M. R., & Mackie, D. (1984). Attitude prototypes as determinants of attitude-behavior consistency. *Journal of Personality and Social Psychology, 46*, 1254–1266.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*, 163–203.

McConnell, A. R., & Leibold, J. M. (2001). Relations among the Implicit Association Test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of Experimental Social Psychology, 37*, 435–442.

Mierke, J., & Klauer, K. C. (2001). Implicit association measurement with the IAT: Evidence for effects of executive control processes. *Zeitschrift für Experimentelle Psychologie, 48*, 107–122.

Mierke, J., & Klauer, K. C. (2003). Method-specific variance in the Implicit Association Test. *Journal of Personality and Social Psychology, 85*, 1180–1192.

Mitchell, J. A., Nosek, B. A., & Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General, 132*, 455–469.

Pratto, F., & John, O. P. (1991). Automatic vigilance. The attention grabbing power of negative social information. *Journal of Personality and Social Psychology, 61*, 380–391.

Rothermund, K., & Wentura, D. (2001). Figure–ground asymmetries in the Implicit Association Test (IAT). *Zeitschrift für Experimentelle Psychologie, 48*, 94–106.

Schwarz, N., & Bohner, G. (2001). The construction of attitudes. In A. Tesser, & N. Schwarz (Eds.), *Blackwell handbook of social psychology: Intraindividual processes* (pp. 436–457). Malden, MA: Blackwell.

Schwenkmezger, P., Hodapp, V., & Spielberger, C. D. (1992). *Das State–Trait-Ärgerausdrucks-Inventar STAXI (The state–trait angriness inventory STAXI)*. Göttingen: Huber.

Spielberger, C. D. (1988). *State-Trait-Anger-Expression-Inventory (STAXI)* (research ed.). Odessa, FL: Psychological Assessment Resources.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *STAI. Manual for the State–Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists.

Strack, F., & Deutsch, R. (in press). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*.

Wentura, D., Rothermund, K., & Bak, P. (2000). Automatic vigilance: The attention-grabbing power of approach- and avoidance-related social information. *Journal of Personality and Social Psychology, 78*, 1024–1037.

Wigboldus, D. (2003). *Single target implicit association test*. Paper presented at the Third Workshop on Implicit Diagnostics, Bonn.

Williams, J. M. G., Mathews, A., & MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin, 120*, 3–24.

Wilson, T., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107*, 101–126.

Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology, 72*, 262–274.