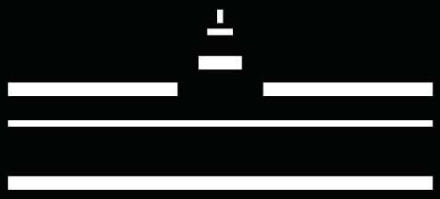
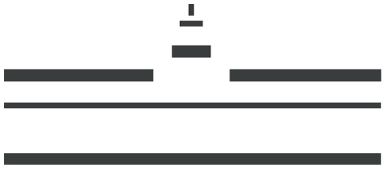


Westfälische
Wilhelms-Universität
Münster

Psychometric Properties of Indirect Assessment Procedures

Reliability and Validity of Implicit Self-Esteem Measures

Sascha Krause



Westfälische Wilhelms-Universität Münster
Fachgebiet: Psychologie

Psychometric Properties of Indirect Assessment Procedures: Reliability and Validity of Implicit Self-Esteem Measures

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Philosophischen Fakultät
der Westfälischen Wilhelms-Universität
zu Münster (Westf.)

vorgelegt von

Sascha Krause
aus Merseburg

im Oktober 2010

Tag der mündlichen Prüfung: 30.11.2010

Dekan: Prof. Dr. Christian Pietsch

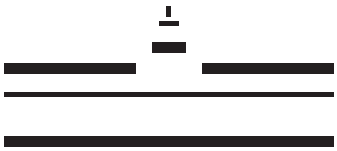
Referent: Prof. Dr. Stefan C. Schmukle

Korreferent: Prof. Dr. Boris Egloff

Sascha Krause

Psychometric Properties of Indirect Assessment Procedures

MV WISSENSCHAFT



Westfälische
Wilhelms-Universität
Münster

Wissenschaftliche Schriften der WWU Münster

Reihe VIII

Band 3

Sascha Krause

Psychometric Properties of Indirect Assessment Procedures

Reliability and Validity of Implicit Self-Esteem Measures

MV WISSENSCHAFT

Wissenschaftliche Schriften der WWU Münster
herausgegeben von der Universitäts- und Landesbibliothek Münster
http://www.ulb.uni-muenster.de

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Dieses Buch steht gleichzeitig in einer elektronischen Version über den Publikations- und Archivierungsserver der WWU Münster zur Verfügung.
http://www.ulb.uni-muenster.de/wissenschaftliche-schriften

Sascha Krause
„Psychometric Properties of Indirect Assessment Procedures. Reliability and Validity of Implicit Self-Esteem Measures“
Wissenschaftliche Schriften der WWU Münster, Reihe VIII, Band 3

Die Reihe „Wissenschaftliche Schriften der WWU Münster“ erscheint im Verlagshaus Monsenstein und Vannerdat OHG Münster
www.mv-wissenschaft.com

ISBN 978-3-8405-0036-7 (Druckausgabe)
URN urn:nbn:de:hbz:6-75429430244 (elektronische Version)

Satz: Sascha Krause
Umschlag: MV-Verlag
Druck und Bindung: MV-Verlag

Danksagung

Diese Dissertation bildet den Abschluss meiner dreijährigen Forschung zur Untersuchung der psychometrischen Qualitäten indirekter Messverfahren. An der Erstellung dieser Arbeit haben einige Personen direkt und indirekt mitgewirkt. Bei ihnen möchte ich mich an dieser Stelle ganz herzlich bedanken!

Mein größter Dank geht an meinen Doktorvater Prof. Dr. Stefan Schmukle. Er hat mich bei der Erstellung dieser Schrift umfassendst unterstützt und fachlich exzellent betreut. Ich möchte mich unter anderem für das sorgfältige Lesen von Artikelentwürfen, die vielen hilfreichen Beratungsrunden, die jederzeit offene Tür und die unmittelbare Beantwortung von E-Mails ganz herzlich bedanken. Es sind aber nicht nur die betreuungsrelevanten Tätigkeiten, die mir in den letzten Jahren sehr geholfen haben. Ich bin dir, Stefan, außerordentlich dankbar, dass du mich gelehrt hast, wie man wissenschaftliche Fragestellungen und Probleme analysiert, wie man komplizierte statistische Auswertungsprogramme in die Knie zwingt, wie man sorgfältig Seminare vorbereitet oder Vorträge zuhörerfreundlich strukturiert. Diese Aufzählung beinhaltet bei weitem nicht alles, was du mich gelehrt hast. Deshalb einfach danke dafür, dass ich Teil deines Arbeitsteams sein darf.

Ganz viel Dank geht an die „Mainzer“ Prof. Dr. Boris Egloff und Prof. Dr. Mitja Back. Ihr habt durch eure fachlich versierte Beratung einen großen Anteil an der Erfüllung meines Traumes, der Erstellung eines wissenschaftlichen Artikels in Erstautorenschaft. Danke für konstruktive Hinweise, Strukturierungs- und Formulierungsvorschläge. Danke aber auch für so manch lustigen Spruch zu Reviewerkommentaren und anderen „Ärgernissen“ des Wissenschaftlerdaseins.

Ganz lieber Dank geht an meinen Bruder Jan, der sich in einer für ihn schwierigen Zeit aufopferungsvoll um die Programmierung der Anagramm-Aufgabe gekümmert hat. Natürlich danke ich auch meiner lieben Frau Antje für Geborgenheit und Zuwendung in den stressigeren Tagen. Danke für deine Geduld und das Verständnis an längeren Arbeitstagen.

Table of Content

Zusammenfassung

Seit einem Vierteljahrhundert wird der Fachöffentlichkeit eine zunehmende Anzahl von verschiedensten indirekten Messverfahren präsentiert. Mit dieser neuen Klasse von Erhebungsinstrumenten sollen mehrere Schwächen des auf Introspektion beruhenden Selbstberichts überwunden werden. Trotz aller Bemühungen leiden indirekte Messverfahren, im Gegensatz zu ihren Gegenspielern den *direkten* Verfahren, im unterschiedlichen Ausmaß unter niedrigen Reliabilitäten. Beispielsweise können aus Gründen der psychometrischen Zuverlässigkeit bisher nur zwei indirekte Verfahren—der Implizite Assoziationstest (*Implicit Association Test*; IAT) und die Namensbuchstaben-Präferenz Aufgabe (*Name-Letter Task*; NLT)— im Bereich der indirekten Selbstwertmessung eingesetzt werden.

In der ersten Studie wird untersucht, inwieweit die interne Konsistenz und die 4-Wochen Stabilität der sechs gegenwärtig populärsten indirekten Maße durch materielle, strukturelle und analytische Innovationen verbessert werden können. Folgende Verfahren, jedes adaptiert für die indirekte Messung von Selbstwert, kommen dabei zum Einsatz: der IAT, eine Kurzversion des IAT (*Brief Implicit Association Test*; BIAT), ein auf Reaktionszeiten beruhendes affektives Priming-Verfahren (*Affective Priming Task*; APT), ein auf Fehlerraten beruhendes affektives Priming-Verfahren (*Response-Window Affective Priming Task*; RW-APT), das Identifizierungs-extrinsische affektive Simon Verfahren (*Identification-Extrinsic Affective Simon Task*; ID-EAST) und die NLT. Im Einklang mit bestehenden Befunden ergeben sich in dieser Studie die höchsten Reliabilitätsschätzungen für die bereits etablierten indirekten Selbstwertmessverfahren, den IAT und die NLT. Es zeigt sich jedoch auch, dass mit den empfohlenen Innovationen zwar nicht die Reliabilitäten der etablierten Verfahren, wohl aber die interne Konsistenz und die zeitliche Stabilität beider Priming-Verfahren und des ID-EAST deutlich verbessert werden können.

In der zweiten Studie wird die prädiktive Validität der besonders vielversprechenden RW-APT untersucht. Zu diesem Zweck wird den Versuchspersonen eine Annagramm-Aufgabe präsentiert, die schwieriger als erwartet ist. Auf Basis des induzierten Misserfolgs bei der Aufgabenbe-

wältigung können drei Selbstwert-Kriterien abgeleitet werden: 1) die Positivitätsverzerrung in den Erwartungen für die Lösungshäufigkeit vor der Annagramm-Augabe, 2) die Positivitätsverzerrung für die selbsteingeschätzte Leistung nach der Annagramm-Aufgabe und 3) die Positivitätsverzerrung in den Erwartungen für die Lösungshäufigkeit einer weiteren und ähnlich schwierigen Annagramm-Aufgabe. Die Ergebnisse der zweiten Studie belegen die prädiktive Validität der Selbstwert RW-APT: Jedes der drei Kriterien wird unabhängig von etablierten direkten Selbstwertmessverfahren vorhergesagt.

Als Fazit dieser Dissertation können zwei Erkenntnisse formuliert werden: Erstens, der Bestand verfügbarer und akzeptabel reliabler impliziter Selbstwertmaße kann um die verbesserten APTs und den ID-EAST erweitert werden. Zweitens, die *Response-Window* (Antwortfenster) Technik hat das Potential, zumindest für affektives Priming, die psychometrischen Eigenschaften indirekter Messverfahren substantiell zu verbessern.

Summary

In the past quarter century, a considerable number of indirect assessment procedures has been introduced. This new class of measurement techniques aims to overcome some of the limitations of introspective self-reports. Nevertheless, in contrast to their counterparts, indirect measures suffer from low reliabilities to different degrees. In self-esteem research, for example, only two indirect measures—the Implicit Association Test (IAT) and the Name-Letter Task (NLT)—have been currently established for reliability reasons.

In the first study, I investigated how recently introduced material, structural, and analytic innovations might increase the internal consistencies and the 4-week stabilities of six currently popular indirect measures: the IAT, the Brief Implicit Association Test (BIAT), the standard Affective Priming Task (APT), the Response-Window Affective Priming Task (RW-APT), the Identification-Extrinsic Affective Simon Task (ID-EAST), and the NLT. As expected, Study 1 revealed that the highest reliabilities were obtained for the established self-esteem IAT and the NLT. However, whereas recently presented innovations did not further improve the reliabilities of these measures, they substantially increased the internal consistencies and temporal stabilities of the standard APT, the RW-APT, and the ID-EAST.

In the second study, I investigated the predictive validity of the particularly promising RW-APT. Hence, an anagram task was presented containing anagrams that were more difficult than expected. Based on this failure experience, three self-esteem criteria were obtained: pre-task expectancy bias, perceived performance bias, and post-task expectancy bias. As expected, the self-esteem RW-APT predicted each of the three criteria over and above established self-report measures.

Overall, the results of this thesis point to two novel findings: First, the improved APTs and ID-EAST may broaden the arsenal of reasonably reliable implicit self-esteem measures. Second, the response-window technique is a valuable tool that has the potential to increase, at least for the APT, the psychometric properties of indirect assessment procedures.

1. General Introduction

To find out what people are really thinking and feeling is an ultimate goal of psychology. Asking them directly via questionnaires, interviews, and so forth, although often done, may not be the best way to capture this information because people may be reluctant to reveal or may try to conceal their true thoughts and feelings. Since the beginning of psychology, researchers have tried to overcome this limitation of direct measures by developing so-called indirect methods that avoid reliance on introspective (i.e., deliberative) self-reports. In the distant past, the most prominent representatives of these assessment methods have been Freud's technique of free associations (Breuer & Freud, 1895), the Rorschach Test (Rorschach, 1921/1942), and the Thematic Apperception Test (TAT; Murray, 1943). Nevertheless, all of these methods have in common that the unobtrusively obtained information is again based on verbal reports. Thus, some of the same limitations of direct measures may occur, that is, people may try to obscure the content of their actual beliefs by reporting something that does not correspond to the true content of their thoughts.[1]

About a quarter century ago, a promising solution for the various weaknesses of verbal reports was developed by Fazio, Sanbonmatsu, Powell, and Kardes (1986). They introduced a priming procedure for the measurement of attitudes that was based on technological advances of the computer era. Different from previous indirect assessment approaches, participants responded by pressing a button on the keyboard of a personal computer instead of answering verbal statements. Specifically, after the presentation of unambiguously *good* and *bad* prime words, participants were asked to indicate as quickly as possible the connotation of unambiguously *positive* and unambiguously *negative* target adjectives. The key feature of this priming task is how the prime facilitates (vs. inhibits) the correct evaluation of the target adjectives. According to the priming paradigm, it is expected that the presentation of congruent prime target pairs

[1] It is not guaranteed that a person who looks at a Rorschach inkblot will actually describe the thoughts that come to mind. It is possible that for several reasons an individual will not report his or her true perceptions. For example, someone might be embarrassed about the content of his/her cognitions. Such situations may result in people hiding their true thoughts and reporting something else.

(i.e., good prime/positive adjective or bad prime/negative adjective) leads to shorter response latencies than the presentation of incongruent prime-target pairs (i.e., good prime/negative adjective or bad prime/positive adjective). Thus, facilitation scores were computed by using reaction times as the main dependent variable.

The main finding of Fazio et al.'s (1986) novel approach was that some sorts of attitudes can be automatically (i.e., spontaneously and inescapably) activated from memory. Consequently, they described the finding as the *automatic attitude activation effect*. Furthermore, because participants were not informed about which attitude was assessed, the priming procedure served as an unobtrusive measure that should therefore eliminate some of the limitations of verbal reports.

Nevertheless, it took an additional 9 years before two papers gave the field of indirect measurement techniques its initial attraction: First, Fazio, Jackson, Dunton, and Williams (1995) introduced a variant of the priming task, the affective priming task (APT), which predicted race-related judgments and behaviors among white students. That is, specifically for this group, presenting photos of black people as primes facilitated, on average, the identification of negative adjectives, whereas priming with white photos elicited a facilitation effect for the identification of positive adjectives. Individual differences in the automatic attitude activation effect of negativity were related to behavior toward a black experimenter. That is, those white participants with more negative attitudes toward blacks behaved in a less friendly and less interested manner than those white students with more positive attitudes toward blacks. Another important finding of this study was that the self-report measure did not capture this prejudiced behavior (Study 1 of Fazio et al., 1995). Thus, the pattern of results suggested that indirect measures such as the affective priming task might be valuable complements to questionnaires for the assessment of racial attitudes because some people might be (a) unaware of their true sentiments toward minorities or (b) unwilling to report their negativity toward this group.

Second, in the same year, another influential contribution to the field of indirect measurement was published by Greenwald and Banaji (1995). They reviewed research on different domains of social psychol-

ogy, namely attitudes, prejudice, and self-esteem, and came to the conclusion that social behavior often operates in an unconscious manner. To explain the findings of their review, Greenwald and Banaji (1995) adapted ideas from cognitive psychology and coined the term *implicit social cognition*. The key feature of its definition is that "traces of past experience affect some performance, even though the influential earlier experience is not remembered in the usual sense—that is, it is unavailable to self-report or introspection" (Greenwald & Banaji, 1995, pp. 4-5). Their review ended with the proposal to develop indirect measures that would allow for the investigation of individual differences in implicit social cognition as a routine. That is, these new assessment procedures should meet the requirements of reliability and validity in the assessment of implicit (i.e., unconscious) aspects of behavior control.

Since 1995, various indirect assessment procedures have been introduced. Before shedding light on the main focus of my thesis—the psychometric properties of indirect measures (see Chapter 2 and Chapter 3)—in this chapter, I will address some important questions regarding their theoretical considerations. First, I will focus on an existing debate about the actual use of two important terms that was raised by both papers: Whereas Fazio et al. (1986, 1995) used the term *automatic* to describe the mode of attitude expression in the affective priming task, Greenwald and Banaji (1995) used the term *implicit* to describe the mode of operation of participants' attitudes (self-esteem and stereotypes) under certain circumstances. Second, I will illustrate the theoretical background of the indirect measurement domain. Despite much research activity, the application of indirect measures has been surprisingly atheoretical for an extended period of time. Third, I will try to justify the necessity of using indirect measures as complements to direct measures. Fourth and last, I will give a short overview about the use of indirect measures in psychological research.

1.1 The Automatic Attitude Activation Effect versus Implicit Social Cognition

The difference between the terms "automatic" and "implicit" might often be hard to understand for readers unfamiliar with the relevant papers. Both terms have their roots in two intellectual traditions from cognitive psychology: Whereas Fazio et al. (1986, 1995) saw their results as being in line with a distinction between automatic and controlled processes offered by cognitive psychologists (e.g., Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977), Greenwald and Banaji (1995) linked their findings to research on implicit memory, which differentiates between implicit (unconscious) and explicit (conscious) processes (Graf & Schacter, 1985; for a critical review, see Schacter, 1987).

For the sake of clarity, Fazio et al.'s (1986) "automatic attitude activation effect" means that attitudes can be activated spontaneously and without any conscious effort. According to this, in their seminal paper, the term *unconscious* was not used. The key feature of the automatic attitude activation effect was named inescapability. This means that "upon presentation of an attitude object, an individual's attitude would be activated despite the lack of any reflection whatsoever on his or her part" (Fazio et al., 1986, p. 229). Controlled processing, by contrast, was defined as demanding attention that requires the active concentration of the individual. Insofar, this process is seen as reflective and active in nature because the individual needs to retrieve a previously stored evaluation of the attitude object or may actively construct such an attitude on the spot (Fazio et al., 1986).

Nevertheless, Greenwald and Banaji's (1995) influential review led to a shift in the use of the automatic-controlled distinction to the implicit-explicit dichotomy—interpreted as synonyms for the terms "unconscious" and "conscious." Based on findings of implicit memory, they defined implicit attitudes as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects" (Greenwald & Banaji, 1995, p. 8). The key feature of this definition is that implicitness is equated with

unconsciousness because people are not able to recall the true trace that led to the attitude expression in a certain situation.

Note that the definition of implicit attitudes by Greenwald and Banaji (1995) is not equal to definitions of implicit memory. These memory studies showed that prior exposure to stimuli (e.g., words in a list) facilitated the completion of word fragments or word stems. This effect occurred despite the fact that participants were not able to recall or recognize words from the earlier list (for a review, see Schacter, 1987). That is, implicit memory studies provided objective evidence that prior exposure to stimuli had a positive effect on later performance although participants did not remember the presentation of the stimuli. However, implicit social cognition researchers were not able to provide clear evidence of unmemorable or unconscious past experiences that gave rise to attitudes, stereotypes, or self-esteem. By contrast, the available evidence clearly speaks against this assumption of unconscious attitudes (Gawronski, Hofmann, & Wilbur, 2006).

Nevertheless, claims that indirect measures assess unconscious aspects of attitudes or traits remain common. In order to come to an agreement on the debate about the difference between the terms automatic and implicit, I propose using the terms interchangeably. This is line with De Houwer (2006; De Houwer & Moors, 2010; De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009) who suggested that researchers "replace the concept of implicit with the concept of automatic" (Moors, Spruyt, & De Houwer, 2010, p. 19) for theoretical and practical reasons.

Since the introduction of the term "implicit social cognition," another discussion has been raised about the proper use of the term implicit regarding the characterization of the effect of an assessment procedure or the assessment procedure itself. Specifically, some researchers have applied the term implicit to describe the fact that participants were not asked to verbally report their attitude (e.g., Fazio & Olson, 2003). In this view, an implicit measure is a procedure that does not require an introspective self-report. The main advantage of this measure is that it is able to circumvent response set problems associated with questionnaire measures (see section 1.3). Other researchers, however, have used the term implicit to describe the construct itself, which is assessed by a reaction-time-based

measure. In this view, implicit measures do not require conscious introspection, and therefore might reflect psychological structures that are introspectively inaccessible (e.g., Banaji, 2001).

With regard to a common application of the terms, the most comprehensive conceptual analysis has been provided by De Houwer (2006). He suggested that the terms "implicit" and "explicit" should describe features of the psychological attributes that are assessed by different measurement procedures. To describe features of the measurement procedure, the terms "direct" and "indirect" should be used (e.g., De Houwer et al., 2009). As an illustration of the different applications of the four terms, consider self-esteem: A direct measure (e.g., the Rosenberg Self-Esteem Scale; RSES; Rosenberg, 1965) requires deliberative responses to self-esteem-relevant questionnaire items. Because participants are aware of the measurement purpose and can control their responses, this self-esteem scale would capture explicit self-esteem. An indirect measure of self-esteem (e.g., the self-esteem APT) requires fast, and thus spontaneous, reactions to target adjectives after the presentation of self-related stimuli (e.g., a person's own name or own picture). Because participants may not be aware of the measurement purpose and may not control their responses, a self-esteem APT would capture implicit self-esteem.

1.2 Theoretical approaches

In recent years, several "two-process" theories of information processing have been offered. They follow the assumption that individuals process information about themselves and their environment in not only an explicit (i.e., controlled) but also an implicit (i.e., automatic) mode (Epstein, 1994; Fazio, 1990; Greenwald & Banaji, 1995; Greenwald et al., 2002; Strack & Deutsch, 2004; Wilson, Lindsey, & Schooler, 2000). Note that the explicit-implicit distinction does not rest on conscious versus unconscious process, nor does it rest on the postulation of structurally distinct brain mechanisms. That distinction, instead, rests most firmly on findings that have provided evidence for two classes of assessment methods that have different empirical correlates (Banse & Greenwald, 2007). Two of the most influential dual process models that explain which under-

lying processes explain which behavior are the Reflective-Impulsive Model (RIM; Strack & Deutsch, 2004) and the MODE-model (Fazio, 1990).

1.2.1 The Reflective-Impulsive Model (RIM)

The RIM explains how social behavior can be predicted by underlying associative and/or propositional processes. Both kinds of processes are seen as distinct but also interdependent. Associative processes are related to associations in memory (which are stored in a so-called "associative system") that were activated automatically in response to a relevant attitude object. Thus, this kind of process requires little cognitive capacity to evaluate a specific stimulus. Propositional processes, by contrast, are responsible for validating the automatically activated evaluation of the associative system. Accordingly, evaluations that stem from such deliberative processes are assumed to be stored in a so-called "reflective system." Consequently, the main hypothesis of the RIM is that the reflective system is subordinate to the associative store because it transforms inputs from the associative system into a propositional format (e.g., a negative affective reaction toward X is translated into the proposition "I dislike X"). The central difference between the two kinds of processes is their dependency on truth values and accuracy, respectively. That is, whereas automatically activated associations need to be accurate because they occur independently of whether a person considers these contents to be true or false, deliberative evaluations resulting from propositional processes of the reflective system are not bounded by this degree of accuracy because people have the opportunity to validate the automatically activated evaluation as true or false.

The distinction between different processes of an associative system and a reflective system have allowed for the theoretical interpretation of outcomes resulting from direct and indirect assessment procedures (e.g., Gawronski & Bodenhausen, 2006; Strack & Deutsch, 2004). Whereas indirect attitude measures are assumed to tap evaluations that have their roots in associative processes, explicit attitude measures are assumed to tap evaluations that have their roots in propositional processes. In accor-

dance with this hypothesis, previous research has provided support for these assumptions by showing that (a) indirect measures have been able to predict automatic (e.g., spontaneous) behavior that was not predicted by corresponding self-report measures and (b) direct measures have been able to predict controlled behavior that was not predicted by corresponding indirect measures (double dissociation; Asendorpf, Banse, & Mücke, 2002; Egloff & Schmukle, 2002).

1.2.2 The MODE Model

In contrast to the dual representation assumption implied by the RIM model, a second class of model approach is characterized by a single representation assumption. Based on the automatic attitude activation effect, Fazio (1990) introduced the MODE model: MODE is an acronym for *Motivation* and *Opportunity* as *DEterminants* of how the attitude-to-behavior process is primarily formed. The central assumption of the MODE model is that indirect and direct measures capture behavioral effects of the same underlying mental representation. Whereas indirect measures capture automatically activated attitudes, direct measures capture the same attitude, but responses are more controlled by the participant. Thus, the main difference between indirect and direct measures is the degree of control participants have over their responses. Whereas indirect measures constrain participants' opportunities to control their responses, direct measures are contaminated by influences of motivated processes because responses on self-report measures can be easily altered. In an exaggerated view from the MODE-model, responses to direct measures may be an exercise in self-presentation (Olson, Fazio, & Hermann, 2007). Thus, they may not offer an accurate portrait of automatically activated self-evaluations.

According to theoretical assumptions of the MODE model, automatically activated attitudes will guide judgments and/or behavior if either the motivation or the opportunity to engage in deliberations is low. If, however, both the motivation and the opportunity for deliberations are high, the influence of automatically activated evaluations on self-report measures may be attenuated. In line with the assumptions of the MODE

model, the predictive validity of a given measure depends on the processing conditions that existed during the complementation of a measure and the circumstances that existed during the execution of certain behavior. Hence, as specified for the RIM-model, whereas indirect measures should be better in predicting automatic (e.g., spontaneous) behavior, direct measures should be better predictors of controlled (e.g., deliberative) behavior. Previous research has provided evidence for these assumptions (e.g., Asendorpf et al., 2002; Egloff & Schmukle, 2002; Fazio et al., 1995; Greenwald & Farnham, 2000).

Note that the MODE-model differs from the RIM in one important way: Whereas the RIM postulates two distinct albeit related processes for the explanation of social behavior, the MODE model explicitly proposes the possibility of "mixed" processes that are neither purely automatic nor purely controlled. Such mixed processes are essential elements of social reality because individual behavior is frequently not completely automatic or completely controlled. Previous studies that have been able to show that direct and indirect measures independently predicted the same behavior-relevant criterion have provided empirical support for this assumption (e.g., Back, Schmukle, & Egloff, 2009; Back, Krause, et al., 2009).

1.3 On the Necessity of Indirect Assessment Procedures for Psychological Research

Traditionally, self-report measures have been the method of choice for obtaining information from individuals. These often-used methods have a number of advantages, including high economy and high reliability. Nevertheless, they often suffer from limitations that can be roughly divided into two categories (see e.g., Greenwald et al., 2002). First, direct measures rely on the participant's willingness to report private knowledge. Mainly in socially sensitive situations (e.g., the measurement of prejudice), factors such as social desirability (Crowne & Marlowe, 1960), various response sets (Cronbach, 1946, 1950), self-presentation styles (Paulhus, 1984), demand characteristics (Orne, 1962), or intentional faking (Cronbach, 1990) may cause distorted self-disclosure.

Indirect measures, however, seem to be resistant—or at least less susceptible—to such motivational distortions (e.g., Fazio et al., 1995; Greenwald, Nosek, & Banaji, 2003; Hermans, Vansteenwegen, Crombez, Baeyens, & Eelen, 2002; Wiers, Van Woerden, Smulders, & De Jong, 2002). This interpretation is based on findings that participants have greater difficulty in intentionally faking reactions on these new indirect measures than they have for standard self-report measures (e.g., Banse, Seise, & Zerbes, 2001; Egloff & Schmukle, 2002; Kim, 2003; Steffens, 2004; but also see Fiedler & Bluemke, 2005). Furthermore, because participants are presumably unaware of the relationship between these measures and their attitudes, indirect measures also minimize the incentives and opportunities for strategic responding. Thus, the development of indirect measures seems to be a promising way to obtain undisguised information about participants. Accordingly, 15 years ago, Fazio et al. (1995) spoke euphorically about a bona fide pipeline "to get inside the head of the participant" (p. 1014) without asking them to consider their attitudes.

Second, direct measures depend on a subject's ability to report such knowledge accurately. However, prior research has suggested that the self-view can be introspectively limited, that is, people may have no conscious access to processes that guide their evaluations and/or behavior (Greenwald & Banaji, 1995; Nisbett & Wilson, 1977; Wilson et al., 2000). Thus, when participants are asked to respond to self-report measures, they might rely on naive theories, which may or may not be correct (Gawronski, LeBel, & Peters, 2007). Therefore, some scientists have proposed using the new indirect assessment technique for measuring implicit (i.e., unconscious) representations that are not available for self-report or introspection (e.g., Asendorpf et al., 2002; Baccus, Baldwin, & Packer, 2004; Dijksterhuis, 2004; Spalding & Hardin, 1999).

However, note that the assumption that indirect measures can provide information about people in cases in which they are unaware of their true attitude is controversially debated. To date, there is no empirical evidence that indirect measures assess unconscious representations. By contrast, indirect measurement techniques tend to show quite substantial correlations with questionnaire measures when methodological factors (e.g., conceptual correspondence, measurement error) are taken into account

(Cunningham, Preacher, & Banaji, 2001; Gawronski et al., 2007). Furthermore, correlations between direct and indirect measures have been diminished for just those dispositions for which self-presentation concerns exists. In accordance with this assumption, Olson et al. (2007) were recently able to show that discrepancies between direct and indirect measurements of self-esteem were the result of reporting tendencies.

Another weakness of direct measures consists of the lack of predictive validity for social behavior. This might be a consequence of both limitations that were mentioned above (i.e., response sets and/or introspective limited access). This phenomenon is particularly apparent in socially sensitive domains such as prejudice or self-esteem. Here, correlations between questionnaire measures and behavior-relevant criteria are low or zero (e.g., Bosson, Swann, & Pennebaker, 2000; Fazio et al., 1995). However, research with the new class of assessment procedures has shown that indirect measures are able to predict behavior independently from or beyond direct measures (e.g., Back et al., 2009; Back, Krause, et al., 2009; Fazio et al., 1995; Greenwald & Farnham, 2000). Such findings are the most convincing arguments for the necessity of the use of indirect measures in addition to direct ones.

1.4 On the Use of Indirect Assessment Procedures in Psychological Research

At the end of their seminal paper, Greenwald and Banaji (1995) predicted "a new industry of research on implicit cognitive aspects of personality and social behavior" (p. 20) when indirect measures will be available to allow for the investigation of individual differences in implicit social cognition as a routine practice. In the first part of their prediction, they were totally right: Various indirect assessment procedures—most of them relying on response latencies—have been introduced in the last 15 years (for an overview, see Fazio & Olson, 2003). These include the APT, the Name-Letter Task (NLT; Kitayama & Karasawa, 1997), semantic priming (Wittenbrink, Judd, & Park, 1997), the Implicit Association Test (IAT; Greenwald, McGhee, & Schwarz, 1998), the Response-Window Affective Priming Task (RW-APT; Draine & Greenwald, 1998;

see also Wentura, Kulfanek, & Greve, 2005), the Stroop Color-Naming Task (SCNT; Bosson et al., 2000); the Go/No-Go Association Task (GNAT; Nosek & Banaji, 2001), the Extrinsic Affective Simon Task (EAST; De Houwer, 2003a), stereotypic explanatory bias (Sekaquaptewa, Espinoza, Thompson, Vargas, & von Hippel, 2003), the Stimulus Response Compatibility Task (SRCT; Mogg, Bradley, Field, & De Houwer, 2003), the Affect Misattribution Procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005), the Single-Category IAT (SC-IAT; Karpinski & Steinman, 2006), the Implicit Association Procedure (IAP; Schnabel, Banse, & Asendorpf, 2006), the Single Association Test (SAT; Blanton, Jaccard, Gonzales, & Christie, 2006), the Approach-Avoid Task (AAT; Rinck & Becker, 2007), the Identification-Extrinsic Affective Simon Task (ID-EAST; De Houwer & De Bruycker, 2007a), the Word Association Test (WAT; Stacy, Ames, & Grenard, 2007), the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes, Hayden, Barnes-Holmes, & Stewart, 2008), the Single-Block IAT (SB-IAT; Teige-Mocigemba, Klauer, & Rothermund, 2008), the Single-Target IAT (ST-IAT; Bluemke & Friese, 2008), the Brief IAT (BIAT; Sriram & Greenwald, 2009), the IAT-recoding free (IAT-RF; Rothermund, Teige-Mocigemba, Gast, & Wentura, 2009), and the Sorting Paired Features Task (SPF; Bar-Anan, Nosek, & Vianello, 2009).

At present, so many indirect measures are available that it has become difficult to differentiate one measure from another (for a structural analysis of prominent indirect measures, see De Houwer, 2003b). Moreover, the scope of their application is not limited to issues of social psychology such as attitudes and stereotypes (e.g., Fazio et al., 1986, 1995; Greenwald et al., 1998). To date, indirect assessment methods have been applied in various domains of psychological research, for example, clinical psychology (e.g., Teachman, Gregg, & Woody, 2001), self-concept of personality (e.g., Asendorpf et al., 2002; Back et al., 2009), marketing (e.g., Brunel, Tietje, & Greenwald, 2004), and health psychology (e.g., Wiers et al., 2002).

However, despite the various use of this new class of assessment methods, there is still one unresolved problem: Indirect measures suffer in different degrees from low reliabilities that prohibit using this kind of

measurement for the investigation of individual differences (e.g., Bosson et al., 2000; Rudolph, Schröder-Abé, Schütz, Gregg, & Sedikides, 2008). This core message stemmed from one of the most important papers in the field that was published by Bosson et al. (2000). They investigated the internal consistencies and the temporal stabilities of different indirect assessment procedures that were adopted for the measurement of self-esteem. In sum, Bosson et al. (2000) reported disappointingly low reliability estimates for most of the indirect assessment procedures. In particular, their test-retest correlations did not reach the level of standard questionnaire measures (e.g., the RSES). Nevertheless, two indirect measures of self-esteem, namely the self-esteem IAT (Greenwald & Farnham, 2000) and the NLT, demonstrated good internal consistency[2] as well as a respectable level of temporal stability.

At the end of their paper, Bosson et al. (2000) stated that researchers should further explore ways to improve the indirect assessment of self-esteem. This 10-year-old requirement constitutes the main motivation of this thesis. Thus, in the first study, I have addressed this point by investigating the internal consistencies and test-retest reliabilities of different indirect assessment procedures. According to Bosson et al. (2000), each of them was adopted for the measurement of implicit self-esteem.

1.5 Aims of the First Study

In the first Study, I shed light on the reliability of different indirect measures of self-esteem. Self-esteem is one of the most important and most studied constructs in psychological research. Moreover, past research has established the existence of two facets of self-esteem, namely explicit self-esteem and implicit self-esteem (Bosson et al., 2000; Greenwald & Banaji, 1995; Spalding & Hardin, 1999). Whereas explicit self-esteem mirrors the controlled evaluation of a person's own self-worth, implicit self-esteem reflects the automatic self-evaluation.

[2] Note that for the NLT, the reported internal consistency of .57 was a correlation based on only two items. Thus, this is a reasonably high reliability estimate compared to the inter-item correlations of questionnaires that are normally found.

For the assessment of explicit self-esteem, various direct measures are available (for a review, see Demo, 1985). They all tend to reach satisfactory reliability estimates, that is, Cronbach's alphas and test-retest correlations are typically in the range of .80 to .90 (Bosson et al., 2000; Brown & Zeigler-Hill, 2004). For the assessment of implicit self-esteem, likewise, various indirect measures have been adopted. However, reliability estimates, in particular for temporal stability, have been substantially lower compared to explicit self-esteem measures (Bosson et al., 2000; Rudolph et al., 2008).[3]

Altogether, to establish indirect measurement methods for the assessment of interindividual differences, it is absolutely necessary to increase their reliability. With regard to this problem, several material, structural, and analytic innovations for different indirect assessment procedures have been found in the literature in the last decade (see e.g., Baccus et al., 2004; Borkenau & Mauer, 2007; De Houwer & De Bruycker, 2007a; Greenwald et al., 2003; Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007; Krause, Schmukle, Back, & Egloff, 2008; Sriram & Greenwald, 2009; Voss, Leonhart, & Stahl, 2007). Accordingly, the primary aim of the first study was to investigate the impact of these innovations on the internal consistencies and the test-retest correlations of the currently most popular indirect assessment procedures. For this purpose, participants responded twice—with a time lag of 4 weeks—to six different implicit self-esteem measures: the IAT, the BIAT, the standard APT, the RW-APT, the ID-EAST, and the NLT.

[3] Please note that the internal consistencies of different variants of the self-esteem IAT show values in the neighborhood of those typically reported for explicit self-esteem measures (Bosson et al., 2000; Rudolph et al., 2008).

2 Reliability of Implicit Self-Esteem Measures Revisited[4]

2.1 Introduction

Self-esteem is one of the most extensively investigated constructs in personality and social psychology. According to dual process models of personality and information processing (e.g., Back et al., 2009; Epstein, 1994; Greenwald et al., 2002; Strack & Deutsch, 2004; Wilson et al., 2000), one can distinguish between explicit self-esteem (i.e., deliberate evaluations of the self as assessed by direct self-esteem measures) and implicit self-esteem (i.e., automatic self-evaluations as assessed by indirect self-esteem measures; e.g., Back, Krause, et al., 2009; Greenwald & Farnham, 2000; Spalding & Hardin, 1999).

From a psychometric perspective, there is still one unresolved problem: Direct self-esteem measures regularly show high reliability coefficients. By contrast, reliability estimates for many indirect measures do not reach the magnitude of those reported for direct measures (Fazio & Olson, 2003). In a seminal paper, Bosson et al. (2000) analyzed the reliabilities of different implicit self-esteem measures. Whereas the self-esteem IAT and the NLT showed good internal consistencies as well as at least respectable levels of temporal stability, other indirect measures, namely the SCNT and a subliminal and a supraliminal APT (Hetts, Sakuma, & Pelham, 1999; Spalding & Hardin, 1999), showed rather poor estimates of reliability. Recently, Rudolph et al. (2008) reported results for different implicit self-esteem measures with regard to their internal consistency and test-retest reliability. Here, also, the standard self-esteem IAT and the NLT outperformed the other measures, whereas, in particular, the more recently introduced EAST substantially lagged behind.[5]

---

[4] Chapter 2 (without section 2.7) is reprinted from *European Journal of Personality* (in press), Krause et al., Copyright © 2010, with permission from John Wiley & Sons, Ltd.

[5] Although the SC-IAT slightly outperformed the standard self-esteem IAT with regard to internal consistency, its 6-month temporal stability was substantially lower. In the case of the NLT, the best reliability coefficients were estimated for a modification of the standard procedure in which all letters were presented twice (the so-called *D-IPT*). Rudolph et al. (2008) also reported a satisfying internal consistency for a recently introduced variant of

This is a suboptimal situation: Reliability is a prerequisite for the application of any kind of personality measure (Perugini & Banse, 2007). As is the case for explicit self-esteem measures, a variety of different reliable implicit self-esteem measures should be available to fully capture all aspects of the construct. Thus, as stated by Bosson et al. (2000), "if researchers […] persist in their efforts to perfect the measurement of implicit self-esteem, they may ultimately find ways to improve implicit measures to the point of greater utility" (p. 642). Our study aimed to show whether or not this step has been taken in the last 10 years.

2.1.1 Candidate Measures

In the last decade, several indirect measures have been developed, and improvements in procedures and scoring algorithms of established methods have been suggested. In the next section we give a short overview of analytic innovations of the most prominent implicit self-esteem measures included in our study, namely, the IAT and the NLT. We describe how the recently introduced BIAT and the ID-EAST can be adopted to measure implicit self-esteem. Furthermore, we illustrate how material, procedural, and analytic innovations of a very prominent indirect measure, the APT allow for a more reliable assessment of implicit self-esteem.[6]

---

the EAST, the ID-EAST. Unfortunately, contrary to the other measures, no stability coefficients were published.

[6] Because it was impossible to include all available indirect measures in our study, we selected the most prominent ones. For the self-esteem IAT, we included only the standard variant, which showed higher stability coefficients than the SC-IAT and the GNAT (see Rudolph et al., 2008). Furthermore, other recently developed variants of the IAT, which aimed to eliminate the two-block structure (the SB-IAT, the IAT-RF) were not selected because their estimates of internal consistency were lower than those reported for the standard IAT. For the affective priming task, another prominent variant, the AMP was recently introduced. However, because it showed low internal consistency coefficients in measuring implicit self-esteem (Schmukle, Hirschmüller, Back, & Egloff, 2007), the AMP was not included in our study.

2.1.1.1 Implicit Association Test (IAT)

The self-esteem IAT measures the relative strength of pairs of associations by comparing response times on two combined discrimination tasks. The underlying assumption of the IAT is that if two concepts are highly associated (e.g., self-concept and positivity in the case of a person with high self-esteem), the sorting task will be easier (i.e., faster) when the two associated concepts share the same response key (i.e., *self-positive* vs. *other-negative*) than when they share different response keys (i.e., *self-negative* vs. *other-positive*). At present, the self-esteem IAT is the most prominent implicit self-esteem measure, and it is, with respect to reliability, the benchmark against which all other implicit self-esteem measures are compared.

In contrast to Bosson et al. (2000), we used two recently introduced innovations of the standard scoring procedure. First, self-esteem IAT effects were computed by using an improved scoring algorithm, taking the individual standard deviation of response times into account (Greenwald et al., 2003). Second, based on findings that IAT effects are contaminated by several confounding factors (e.g., Back, Schmukle, & Egloff, 2005; Fiedler, Messner, & Bluemke, 2006), we additionally performed a diffusion-model analysis for two-choice response time tasks (see, e.g., Ratcliff, 1978; Wagenmakers, van der Maas, & Grasman, 2007). As described by Klauer et al. (2007), the diffusion model disentangles construct-related and construct-unrelated variance in the IAT effect. As a result, this scoring procedure leads to an IAT score that is thought to more purely reflect the construct-related aspects of IAT outcomes. The question of the reliability of this score has, however, not been addressed yet. Thus, we investigate whether this analytic strategy leads to an improvement in internal consistency and temporal stability of the IAT effect.

2.1.1.2 Brief Implicit Association Test (BIAT)

There are some potential advantages of this shortened IAT variant: First, the BIAT reduces the length of an IAT measure to one third of the number of required trials. Second, the BIAT is designed with the intention of reducing spontaneous variation in subject strategy by forcing respon-

dents to focus on just two of the four IATs' category-response mappings in each combined task (Sriram & Greenwald, 2009; for a critical view, see Rothermund & Wentura, 2010). Despite these modifications, the BIAT seems to retain the favorable psychometric properties of the original IAT, as indicated by a high internal consistency and moderate short-term stability (Sriram & Greenwald, 2009). In our study, we adopted the BIAT to measure implicit self-esteem. Additionally, we addressed the question of the long-term (4-week) test-retest reliability of the BIAT.

2.1.1.3 Affective Priming Tasks (APT)

Self-esteem APTs measure the extent to which exposure to self-related versus other-related prime stimuli facilitates evaluations of unambiguously valenced target words into a positive and negative category, respectively. Due to the experience of positive affect following self-related stimuli, people with high self-esteem are thought to be faster (or show more correct reactions) when a positive stimulus is preceded by a self-related prime and slower (or show more false reactions) when a negative stimulus is preceded by a self-related prime. Unfortunately, standard self-esteem APTs have not shown the expected positivity bias and have demonstrated disappointingly low internal consistency and temporal stability coefficients (see Bosson et al., 2000).

Five material, structural, and analytic innovations can be applied for improving the reliability of the self-esteem APT: (a) increasing the number of trials,[7] (b) using pictures of the participant instead of words as self-related primes (Krause et al., 2008), (c) individually trimming of response latencies (Borkenau & Mauer, 2007; Wilcox, 1998), (d) including error trials in the analysis by replacing error latencies with values that functioned as error penalties (Greenwald et al., 2003), (e) using an adaptive response-window procedure to circumvent effects of individually different speed-accuracy tradeoffs (Draine & Greenwald, 1998; Wentura et al., 2005).

[7] In contrast to Bosson et al.'s (2000) supraliminal APT in which only 4 trials were involved for the calculation of the implicit self-esteem score, we used 120 trials.

2.1.1.4 Identification – Extrinsic Affective Simon Task (ID-EAST)

The EAST was created to overcome some of the limitations of the IAT (e.g., block structure); however, the EAST has not demonstrated satisfactory reliability (De Houwer & De Bruycker, 2007b; Rudolph et al., 2008; Schmukle & Egloff, 2006; Teige, Schnabel, Banse, & Asendorpf, 2004). De Houwer & De Bruycker (2007a) recently introduced the ID-EAST—a procedural modification of the standard EAST—which considerably enhances the internal consistency of this type of measure (e.g., Richetin, Perugini, Adjali, & Hurling, 2007). As in the standard EAST, participants are asked to decide as quickly as possible whether an unambiguously positive versus negative attribute stimulus has a positive or negative valence. However, in contrast to the standard EAST where the target stimuli (green vs. blue words) do not need to be processed immediately, the ID-EAST requires identification of the target feature before participants can select the correct response. For implicit self-esteem, target stimuli are self-relevant versus non-self-relevant words, which are either presented in lowercase or uppercase letters. The first evidence for a more reliable assessment of implicit self-esteem with the ID-EAST as compared to the EAST was provided by Rudolph et al. (2008). To date, however, the question of the test-retest reliability of the ID-EAST has not been addressed.

2.1.1.5 Name-Letter Task (NLT)

The NLT is based on the well-documented finding that people prefer the letters of their own names, especially their initials, over the remaining letters of the alphabet (Nuttin, 1985, 1987). The degree of this preference has been successfully employed as an indirect measure of self-esteem (e.g., Jones, Pelham, & Mirenberg, 2002; Koole, Dijksterhuis, & van Knippenberg, 2001; Zeigler-Hill, 2006). Since its introduction 20 years ago, different computational procedures have been developed to estimate implicit self-esteem scores. Recently, LeBel and Gawronski (2009) recommended using the *ipsatized double-correction algorithm* (Baccus et al., 2004), which controls for both differences in the likeability of the dif-

ferent letters and individual differences in likeability ratings of letters in general. By contrast, the most widely used *baseline-corrected algorithm* (Kitayama & Karasawa, 1997) controls for only the former compound. Whereas internal consistencies of both scoring algorithms were investigated by LeBel and Gawronski (2009), the question of the temporal stability of the I-algorithm has yet not been addressed.

2.1.2 Goals of This Research

Ten years after Bosson et al.'s (2000) report of disappointingly low reliability coefficients for most implicit self-esteem measures, the present research was designed to reinvestigate the reliabilities of different indirect measurement procedures. In particular, we aimed to show how recently introduced material, procedural, or analytic innovations have the potential to increase the internal consistencies and temporal stabilities of most prominent indirect measures. We therefore conducted a study in which we presented each implicit self-esteem measure at two measurement occasions with a time lag of 4 weeks. Additionally, we examined intercorrelations between self-esteem measures.

2.2 Method

2.2.1 Participants

One hundred one students (78 women, 23 men) participated in the study in exchange for research participation credit or monetary compensation (25 €). Their average age was 23.5 years (SD = 3.5).

2.2.2 Procedure

The procedure consisted of two sessions with a time lag of 4 weeks (M = 28.31 days, SD = 1.26 days). To optimize the reliability of individual differences, participants completed the self-esteem measures in a fixed order: RSES, BIAT, APT, response-window APT, NLT, IAT, and ID-EAST. Stimuli for the different implicit self-esteem measures were German words (see Appendix A for the complete list of stimuli). Attribute- or target-adjectives were selected according to their valences from German

norm tables (Hager & Hasselhorn, 1994). All implicit self-esteem measures were administered on personal computers using Inquisit software (2006). There were two filler tasks: one after the APT and another after the NLT. In these filler tasks, participants were asked to write down their thoughts about a picture from the TAT. At the beginning of the first session, the face of each participant was photographed. At the end of the second session, participants were debriefed carefully about the purpose of the study.

Instead of a standard computer keyboard, participants used external response pads to complete the implicit self-esteem measures. This technical detail should reduce error variance in the recorded latencies because key presses on a standard computer keyboard are buffered in the keyboard hardware before they are signaled to the computer (Voss et al., 2007).

2.2.3 Self-Esteem Measures

2.2.3.1 Rosenberg Self-Esteem Scale (RSES)

The RSES (German version by von Collani & Herzberg, 2003) was employed as a direct measure of people's conscious feelings of global self-worth. Scores were obtained from 10 items, each measured on a 4-point Likert scale (1 = *strongly disagree* to 4 = *strongly agree*). Additionally, observer ratings of self-esteem were obtained by using a peer report version of the RSES. Each participant was asked to nominate two friends or family members to complete the RSES as a peer report. To ensure that the well-known others made anonymous self-esteem evaluations of the participant, peer reports were returned via mail in pre-paid envelopes.[8]

2.2.3.2 Implicit Association Test (IAT)

The self-esteem IAT is a computerized categorization task that measures automatic associations of self-relevant and non-self-relevant

[8] For peer-reported self-esteem, we included only participants who were evaluated by both other people ($N = 89$). Three evaluations were excluded from analysis because one of the peers stated knowing the participant only partially.

words with pleasant and unpleasant words.[9] Five self-related stimuli (e.g., *me*) and other-related stimuli (e.g., *your*) were presented, as well as five stimuli representing the positive (e.g., *cheerful*) and negative (e.g., *vain*) categories. The IAT procedure followed the five-block structure described by Nosek, Greenwald, and Banaji (2005). In the fourth block, participants had to perform twice as many single-categorization trials (40) as in the first two practice blocks to reduce undesirable order effects of both combined judgment blocks (Nosek et al., 2005). The critical blocks 3 and 5 consisted of 80 trials each. Stimulus presentation alternated between the target and the attribute dimension. We kept critical block order constant to avoid confounding order with individual differences in ISE in completing the self-esteem IAT. Therefore, each participant in the present study completed the Self + Pleasant block followed by the Self + Unpleasant block.

Participants were asked to categorize, as quickly and accurately as possible, stimulus words that appeared in the center of their computer screens. When an incorrect categorization was chosen, a red "X" appeared on the screen. Participants had to press the correct button to continue to the next judgment. The computer recorded elapsed time between the start of each stimulus presentation and the correct response. This created a built-in error penalty, which is a property of standard IAT measures. Intertrial intervals were set to 150 ms after correct categorization.

[9] There are currently a few different self-esteem IAT variants used in self-esteem research: *self/other idiographic* (Greenwald & Farnham, 2000, Study 1), *self/other nomothetic* (Greenwald & Farnham, 2000, additional test-retest reliability study), *self/not-self* (Jordan, Spencer, Zanna, Hoshino-Browne, & Correll, 2003), and *self/object* (Jordan, Whitfield, & Zeigler-Hill, 2007). These variants have evolved because of the discovery that the particular contrast category chosen in the SE-IAT can have a large impact on the resulting SE-IAT scores (Karpinski, 2004). Nevertheless, the focus of our study was to select a version of the self-esteem IAT that (a) is established in implicit self-esteem research and (b) shows satisfying reliabilities according to the focus of our study. Unfortunately, whereas Jordan et al. (2003) did not report effect sizes nor internal consistencies of their *self/not-self* IAT version, the relatively low internal consistencies of their *self/object* IAT version (Spearman-Brown coefficient = .62 in Study 1, and Spearman-Brown coefficient = .73 in Study 3) did not convince us to use it for our study. Furthermore, we think that a *self/other* IAT has been established in implicit self-esteem research as a measure with satisfying reliabilities and, more important, predictive validities (see Back, Krause, et al., 2009; Bosson et al., 2000; Greenwald & Farnham, 2000; Grumm, Erbe, von Collani, & Nestler, 2008). Last but not least, because Greenwald and Farnham (2000) reported a test-retest reliability only for the *self/other nomothetic* IAT, we selected this variant for our study.

Self-esteem IAT scores were computed following an improved scoring algorithm (D_1 measure) described by Greenwald et al. (2003): (a) trials with latencies greater than 10,000 ms were eliminated; (b) error trials were included in the analysis by using the latency until the correct response was given (built-in error penalty); (c) the mean latency for the critical trials of Block 3 (self + pleasant) was subtracted from the mean latency for critical trials of Block 5 (self + unpleasant); (d) the resulting difference was divided by the individual-respondent reaction-time standard deviation of Blocks 3 and 5. Scores were calculated such that the higher the IAT effect, the higher the implicit self-esteem of a given participant.

In line with Klauer et al. (2007), we performed a diffusion-model analysis of the self-esteem IAT. Because participants did not produce high enough error rates, which are needed to fit the standard Ratcliff diffusion model, we applied a simplified version of the diffusion model, the EZ-diffusion model (EZ-DM; Wagenmakers et al., 2007), which does not seek to address RT-distribution of the error trials. The EZ-DM determines only the most relevant parameters of the Ratcliff diffusion model, namely boundary separation (a), non-decision time (t_0), and drift rate (v). Whereas parameter a mirrors participants' response conservativeness—that is, high values of a indicate slow response times, but also highly accurate performance—in the decision process, parameter t_0 captures the nondecision component of reaction times—for example, the process of stimulus encoding or preparation of motor responses. Parameter v reflects a participant's performance in the decision process itself, that is, for people with high implicit self-esteem, high values of v are expected in the Self + Pleasant block and low values of v are expected in the Self + Unpleasant block. Thus, to compute the self-esteem IAT-effect based on the EZ-DM, the drift rate of the Self + Unpleasant block was subtracted from the drift rate of the Self + Pleasant block. Because v determines the construct-related component of the IAT-effect, we report results for only this main component of the EZ-DM.

2.2.3.3 Brief Implicit Association Test (BIAT)

The instructions and procedure of the self-esteem BIAT followed the approach described in Experiment 1 by Sriram and Greenwald (2009). Thus, the self-esteem BIAT consisted of only 32 critical trials per block. Each IAT stimulus appeared twice in each critical block (one stimulus per category was deleted: *own, their, active, vain*). A 32-trial practice block using items of non-social categories preceded the self-esteem BIAT.

We presented three critical blocks for the assessment of self-esteem: first, a Self + Positive block; second, an Other + Positive block; and third, a Self + Negative block. Because individual differences were the focus of the present research, all participants completed the BIAT blocks in the same order as described above. This procedural approach allowed for the comparison of two different self-esteem BIAT scores dependent on which categories were focal in both critical blocks. According to the naming convention of Sriram and Greenwald (2009), the first self-esteem BIAT score was named *self-other/positive-(negative)*: Here, the mean latency of the Self + Positive block was subtracted from the mean latency of the Other + Positive block. The second self-esteem BIAT score was named *positive-negative/self-(other)* because the mean latency of the Self + Positive block was subtracted from the mean latency of the Self + Negative block. According to the order of the categories listed in the names, higher scores indicated greater strength of association between the categories listed first and third. Thus, the higher the BIAT effects, the higher the implicit self-esteem of a given participant. Both self-esteem BIAT scores were computed using the D_1 measure (Greenwald et al., 2003). Additionally, as we did for the IAT, we report reliabilities of both self-esteem BIAT scores based on the EZ-DM.

2.2.3.4 Affective Priming Task (APT) – Standard Procedure

The self-esteem APT is a computerized categorization task that quantifies the degree to which exposure to self-related versus other-related prime stimuli facilitates judgments of affective target words. Pictures of participants' faces were presented as self-related primes. Pictures of the faces of two different unknown persons of the same sex who did not par-

ticipate in the experiment served as other-related primes. During the task, participants were asked to identify unambiguously valenced adjectives as pleasant (e.g., *honest*) or unpleasant (e.g., *mean*) as quickly and accurately as possible. The time course of the APT was as follows: Each judgment trial started with the presentation of a prime, which remained on the screen for 67 ms and was immediately replaced by a blank screen for 33 ms. Then the target word was presented with a constant prime-target SOA of 100 ms and remained on the screen until a response key had been pressed. When an incorrect response was chosen, a red X appeared on the screen for 300 ms. The intertrial interval was 1000 ms following the response. An initial practice block was followed by six critical blocks of 20 trials each (each target word was paired once with the self-related and once with an other-related prime). Critical trials were presented in a fixed order to optimize the reliability of individual differences (Banse, 2001).

Self-esteem priming scores were calculated by excluding trials in which an incorrect response was given (5.30%). In line with data treatment in previous research (e.g., Spruyt, Hermans, De Houwer, Vandekerckhove, & Eelen, 2007), for each of the four crucial prime-target conditions, response latencies that deviated by more than 2.5 standard deviations from a participant's mean latency were discarded (2.90%). Further analyses were conducted on log transformed data to reduce the skewness associated with response latencies. For each participant, an index of the relative positivity was calculated for each prime by subtracting the average response latency for pleasant targets from the average response latency for unpleasant targets that followed the same prime. Implicit self-esteem priming effects were computed by subtracting the positivity index for other-related primes from the positivity index for the self-related prime. Higher composite scores thus indicated higher implicit self-esteem.

Typically, self-esteem priming measures have demonstrated low reliabilities (see Bosson et al., 2000). In the past, two promising innovations of alternative scoring algorithms for response time tasks were suggested: First, Borkenau and Mauer (2007) have recently shown that the use of individually trimmed means (Wilcox, 1998) enhances the internal consistency of latency measures. Accordingly, we checked how reliabilities of the self-esteem APT might be improved by deleting, separately for each

prime-target condition, each participant's 10% slowest and 10% fastest responses. The deletion rate was exactly the same as the one used by Borkenau and Mauer (2007).

Second, for the IAT, Greenwald et al. (2003) provided compelling evidence that including error trials in analyses has some advantages compared to deleting these trials. To explore how including error latencies might improve the reliabilities of the self-esteem APT in addition to the individually trimmed response latencies, the following procedure—as suggested by Greenwald et al. (2003)—was used: Error trial latencies were replaced by the mean of the correct responses in the prime-target condition in which the error had occurred plus a penalty of twice the standard deviation of correct responses in the respective prime-target condition.

Third, as we did for the IAT, we report effect sizes and reliabilities of the self-esteem APT based on the EZ-DM. To compute self-esteem APT effects, the drift rate difference for other-related primes ($v_{other\text{-}pleasant} - v_{other\text{-}unpleasant}$) was subtracted from the drift rate difference for self-related primes ($v_{self\text{-}pleasant} - v_{self\text{-}unpleasant}$).

2.2.3.5 Affective Priming Task – Response-Window Procedure (RW-APT)

After a filler task that followed the standard APT, participants completed the self-esteem RW-APT. In the RW-APT, the same prime and target stimuli as in the standard priming procedure were used. The time course of the stimuli sequence was also identical, but 300 ms after the target stimulus had appeared, a response window was opened for 150 ms. Participants were instructed to respond to the target stimulus within this time limit. The appearance of a white exclamation point defined the beginning of the response window. If the participant's reaction fell within the window, the white exclamation point turned green. The exclamation point did not change color and disappeared when the participant reacted too slowly. If the participant reacted before the start of the window, the exclamation point never appeared. No feedback was given for incorrect responses. The intertrial interval was set to 1,000 ms.

At the beginning of the RW-APT, participants completed three practice blocks in order to become familiarized with the task. After the practice blocks, six critical blocks of 20 trials each followed. After each block, a feedback summary was given, indicating the percentage of responses within the window, the percentage of correct responses, and the mean reaction time. Additionally, participants were told that they should maximize the rate of responses within the window. Simultaneously, to avoid that participants respond by chance, they were asked to keep the percentage of correct responses in a range of about 70-80%. Depending on the participant's average accuracy and speed in the preceding critical block, the response window was adjusted by default (see Draine & Greenwald, 1998). The onset of the response window varied between 167-400 ms across participants and blocks.

Self-esteem RW-APT scores were calculated analogously to the standard reaction-time-based self-esteem priming, but instead of response latencies, error rates served as dependent variables. Implicit self-esteem priming effects were computed by subtracting the mean error rate difference (unpleasant minus pleasant targets) for the other-related primes from the mean error rate difference (unpleasant minus pleasant targets) for the self-related primes. Higher composite scores indicated higher implicit self-esteem. Analogous to the IAT and the APT, we report effect sizes and reliabilities of the self-esteem RW-APT effects based on the EZ-DM. The scoring procedure was the same as the one used for the diffusion model analysis of the self-esteem APT.

2.2.3.6. Identification Extrinsic Affective Simon Task (ID-EAST)

The ID-EAST was adopted to measure implicit self-esteem by presenting target words related to self (e.g., *my*) and other (e.g., *you*) intermixed with unambiguously positive (e.g., *nice*) and negative (e.g., *cruel*) attribute words. All words were presented equally often in uppercase and lowercase letters to ensure that participants would identify the contents of the stimuli. If attribute stimuli appeared on the screen, participants were instructed to evaluate all adjectives according to their valence by pressing the left key for positive words and the right key for negative words. The

purpose of these trials was to assure that the response keys became extrinsically associated with a positive versus a negative valence. If target stimuli appeared on the screen, participants were asked to select the response on the basis of the letter case in which the pronoun was presented. If a self-related or other-related word appeared in uppercase letters, participants had to press the left key, the "positive key." By contrast, if they saw a self-related or other-related word in lowercase letters, participants had to press the right key, the "negative key."

Following De Houwer and De Bruycker (2007a), the ID-EAST started with two blocks in which the attribute and target discrimination tasks were practiced separately. Next, participants completed four combined critical blocks of 50 trials involving two practice attribute trials. In the remaining 48 critical trials of each test block, each of the target words was presented twice in lower case, and twice in upper case, whereas each attribute word appeared once in each letter case. In each trial, a black fixation cross appeared for 400 ms before the stimulus word was displayed. If a participant's selection was incorrect, a red *X* appeared on the screen for 400 ms, and the next trial started. An intertrial interval of 500 ms was used.

Self-esteem ID-EAST scores reflect the ease with which participants press the positive versus negative key when a self-related versus other-related pronoun appeared on the screen. Four implicit self-esteem scores were obtained: First, based on correct response latencies, the standard score was computed. Second, based on error frequencies (9.52%), the error score was computed (see De Houwer & De Bruycker, 2007a, for both scoring algorithms). Additionally, we computed two self-esteem ID-EAST scores by using the alternative scoring algorithms of the standard APT. Thus, implicit self-esteem scores based on individually trimmed response latencies as well as error penalties were computed. Furthermore, we report effect sizes and reliabilities of the self-esteem ID-EAST based on the EZ-DM.

2.2.3.7 Name-Letter Task (NLT)

In the NLT, participants evaluated how much they liked each letter of the alphabet on response scales ranging from 1 (*I dislike this letter very much*) to 7 (*I like this letter very much*). We computed two different scoring algorithms to calculate preference scores for the participants' initials: first, the most widely used *baseline-corrected algorithm* (B-algorithm; Kitayama & Karasawa, 1997); and second, the *ipsatized double-correction algorithm* (I-algorithm; Baccus et al., 2004), which is recommended for both theoretical and empirical reasons (LeBel & Gawronski, 2009). For computing the B-algorithm, normative letter baselines were calculated by averaging the letter ratings for individuals whose first and last initials did not include the letter. Next, the respective letter baselines were subtracted from each participant's ratings of his or her own initials. The I-algorithm was computed by additionally considering the mean liking score that participants gave to the remaining letters of the alphabet. That is, the average of the participant's rating of all non-initials letters was subtracted from the participant's initials ratings. Next, similar to the B-algorithm, ipsatized letter baselines were subtracted from this difference.

2.2.4 Treatment of Outliers

Data from four participants were excluded for the following measures: The IAT data of one participant was deleted because his mean response latency on critical trials at session 1 was more than 5.5 *SD* above the mean. For the APT, two participants with extreme error values were excluded: Participant 1 made more than 25% errors on the standard APT at session 2. Participant 2 made more than 50% errors (i.e., she reacted by pure chance) on the RW-APT at session 2. For the NLT, the data for one participant were deleted because his name letter effect was more than 3 *SD* below the mean name letter effect.

2.3 Results

2.3.1 Descriptive Statistics

For both measurement occasions, descriptive statistics, internal consistencies, and test-retest correlations of the self-esteem measures can be found in Table 1. All explicit and implicit self-esteem measures exhibited a statistically significant positivity bias, which was moderate to large in magnitude (all *d*s > 0.40). People's general tendency to have positive implicit and explicit self-evaluations (Greenwald & Banaji, 1995; Yamaguchi et al., 2007) was thus reflected in each self-esteem measure.

2.3.2 Internal Consistencies

We estimated split-half reliabilities of the IAT, BIAT, APT, RW-APT, ID-EAST by applying the respective scoring algorithm separately to two mutually exclusive subsets of the critical trials. As recommended by Schmukle and Egloff (2006), critical trials were divided by using an odd-even method in which the trials of the two test halves were equally distributed over the task. The Spearman-Brown corrected correlation between these two part-measures across participants provided a measure of internal consistency. For the NLT, the Spearman-Brown corrected correlation between participants' preferences for their first and last initials served as a measure of internal consistency. Cronbach's α for the RSES was based on items.

As expected, the internal consistency of the RSES was high for both the self-report and the peer report. With regard to the standard scoring algorithms of the implicit self-esteem measures, the highest internal consistencies were obtained for both IAT variants. However, the self- esteem BIAT required only about one third the number of critical trials to reach the same level of internal consistency.

Table 1
Descriptive Statistics, Positivity Biases, Internal Consistencies, and Test-Retest Correlations of the Self-Esteem Measures

	First occasion of measurement				Second occasion of measurement				
	M	*SD*	*d*	*Rel*	*M*	*SD*	*d*	*Rel*	r_{tt}
Direct Measures									
Rosenberg Self-Esteem Scale - self-report	3.24	0.52	1.41**	.88	3.28	0.58	1.36**	.92	.88**
Rosenberg Self-Esteem Scale - peer-report	3.21	0.48	1.49**	.89	-	-	-	-	-
Indirect Measures									
Implicit Association Test									
D_1-Measure	0.58	0.29	1.99**	.75	0.53	0.30	1.77**	.83	.54**
EZ-DM	0.10	0.08	1.22**	.37	0.10	0.09	1.11**	.52	.38**
Brief Implicit Association Test									
Self-Other/Positive-(Negative)									
D_1-Measure	0.27	0.41	0.66**	.75	0.59	0.36	1.64**	.67	.13
EZ-DM	0.07	0.15	0.49**	.33	0.14	0.14	1.00**	.25	.11
Positive-Negative/Self-(Other)									
D_1-Measure	0.33	0.43	0.77**	.74	0.53	0.37	1.43**	.67	.22*
EZ-DM	0.08	0.15	0.57**	.55	0.13	0.16	0.83**	.40	.21*
Affective Priming Task									
Standard Procedure									
RT score (log)	0.11	0.10	1.09**	.55	0.08	0.10	0.81**	.54	.29**
Individually trimmed	0.09	0.09	1.03**	.67	0.08	0.09	0.83**	.66	.30**
Individually trimmed with error-penalty	0.10	0.10	1.03**	.72	0.08	0.10	0.83**	.69	.32**
EZ-DM	0.09	0.17	0.52**	.03	0.08	0.22	0.37**	.26	-.03

(continued)

Table 1 (continued)

Response-Window Procedure									
Error score	0.08	0.21	0.38**	.67	0.12	0.28	0.42**	.72	.43**
EZ-DM	0.08	0.23	0.35**	.58	0.11	0.28	0.38**	.68	.37**
Identification-Extrinsic Affective Simon Task									
RT score (log)	0.10	0.15	0.67**	.51	0.07	0.13	0.55**	.34	.13
Individually trimmed	0.12	0.15	0.77**	.64	0.07	0.13	0.50**	.63	.20*
Individually trimmed with error-penalty	0.16	0.19	0.87**	.75	0.12	0.17	0.70**	.77	.29**
Error score	0.12	0.16	0.72**	.63	0.12	0.15	0.79**	.65	.44**
EZ-DM	0.11	0.16	0.72**	.28	0.13	0.17	0.69**	.05	.18†
Name-Letter Task									
Baseline-corrected algorithm	0.70	1.20	0.58**	.51	0.80	1.24	0.64**	.53	.38**
Ipsatized double-correction algorithm	0.72	1.20	0.60**	.51	0.81	1.20	0.68**	.51	.50**

Note. $N = 101$; M = mean; d = Cohen's effect size; *Rel* = Spearman-Brown corrected split-half reliability; r_{tt} = 4-week test-retest reliability. EZ-DM = EZ diffusion-model analysis (drift rate).
† $p < .10$. * $p < .05$. ** $p < .01$.

The diffusion-model analysis of the self-esteem IAT showed that the internal consistency of the IAT effect based on drift rates was substantially lower than those found for the established D_1 measure. A similar decrease in internal consistency was obtained also for all other two-choice response-time tasks, with the exception of the RW-APT, for which scores based on the diffusion model showed internal consistencies similar to those found for error scores.

Split-half reliabilities for the standard scoring algorithms of the APT, NLT, and ID-EAST were moderate und slightly lower compared to the IAT variants. However, the use of individually trimmed response latencies provided a substantial increase in the internal consistencies of the APT and ID-EAST. By further adding error penalties, split-half reliabilities were reached, which were on the same level as reported for the IAT variants (mean internal consistency coefficients were > .70).

An important increase in the internal consistencies of the APT and ID-EAST occurred when error rates were considered instead of reaction times. For the APT, the use of the adaptive response-window procedure led to a substantial increase in split-half reliability coefficients, which showed values comparable to those reported for both IAT variants. Most notably, no critical trials, as for the improved APT, had to be discarded. A similar result was observed for the ID-EAST: The internal consistency of the error-based score was clearly superior to the internal consistency of the standard reaction-time-based score.

For the NLT, the use of the I-algorithm did not substantially improve the internal consistency compared to the B-algorithm. This result resembled those reported in LeBel and Gawronski (2009).

2.3.3 Test-Retest Stabilities

As often reported in the literature, the RSES possessed a high stability coefficient. The 4-week stabilities of the implicit self-esteem measures were lower than that of the explicit self-esteem measure and lower than the corresponding split-half reliabilities. Consistent with previous findings, the highest stability coefficient was found for the IAT (e.g., Egloff, Schwerdtfeger, & Schmukle, 2005). The size of the test-retest correlation

resembled previous results for standard self-esteem IATs (see Greenwald & Farnham, 2000; Rudolph et al., 2008; but see also Bosson et al., 2000 for an unusually high temporal stability of the self-esteem IAT).

Furthermore, the temporal stability of the IAT effect based on drift rates was substantially lower than that of the D_1 measure. A similar result was obtained for the diffusion-model-based test-retest correlation of the self-esteem APT. For all other implicit self-esteem measures, namely the BIAT, the RW-APT, and the ID-EAST, temporal stability coefficients for the diffusion model analysis showed values in the neighborhood of the standard scoring algorithms.

Surprisingly, in contrast to the IAT, the stability coefficients of the BIATs were disappointingly low. For the *self-other/positive-(negative)* BIAT score, the effects of both sessions did not correlate significantly. With regard to the *positive-negative/self-(other)* BIAT score, the stability coefficient, although significant, was relatively low.

The test-retest reliability of the reaction-time-based APT was notably higher than that reported for a supraliminal APT by Bosson et al. (2000). Improved scoring (trimming plus error penalty), although enhancing the internal consistency, had only a slight effect on temporal stability. The use of the adaptive response-window procedure, however, yielded a test-retest correlation that was substantially higher than that of the reaction-time-based APT.

For the ID-EAST, the test-retest correlation of the reaction-time-based score did not differ significantly from zero. However, the improved scoring algorithm had a substantial impact on the temporal stability: With the use of individually trimmed response latencies and by adding error penalties, we found a correlation between both measurement occasions that was comparable to the stability of the RW-APT.

The 4-week stability of the NLT was different for both scoring algorithms. Here, the test-retest reliability of the I-algorithm was substantially better compared to the B-algorithm. Nevertheless, the general stability level was lower compared to the one reported by Bosson et al. (2000).

2.3.4 Supplementary Analyses

Table 2 presents intercorrelations among the explicit and implicit self-esteem measures. We report results for the averaged effects across both sessions because the interrelations between the trait-relevant aspects of the implicit and explicit self-esteem measures were of primary interest. In order to give a comprehensive but concise overview, we selected only the scores for which we found the highest reliability coefficients.

Whereas self-reported and peer-reported self-esteem were strongly correlated with one another, explicit and implicit self-esteem were mostly uncorrelated (average r = .07). Only the self-esteem IAT and the NLT correlated significantly with the RSES in the expected direction. For the peer report, little evidence of convergent validity was shown by a marginally positive correlation with the error score of the self-esteem ID- EAST. Note that the correlation between NLT and peer-reported RSES is reduced to r = .13, p = .24, after controlling for self-reported RSES. The significantly negative correlation between the *positive-negative/self-(other)* BIAT score and the observer ratings was clearly unexpected.

Whereas self-reported and peer-reported self-esteem were strongly correlated with one another, explicit and implicit self-esteem were mostly uncorrelated (average r = .07). Only the self-esteem IAT and the NLT correlated significantly with the RSES in the expected direction. For the peer report, little evidence of convergent validity was shown by a marginally positive correlation with the error score of the self-esteem ID-EAST. Note that the correlation between NLT and peer-reported RSES is reduced to r = .13, p = .24, after controlling for self-reported RSES. The significantly negative correlation between the *positive-negative/self-(other)* BIAT score and the observer ratings was clearly unexpected.

Intercorrelations among implicit self-esteem measures were generally low (average r = .10). Contrary to previous implicit self-esteem research, however, we found at least some significant correlations between *different* tasks. In particular, the RW-APT was significantly correlated with the ID-EAST and the NLT. Additionally, there were marginally significant relations of the ID-EAST with the IAT and the reaction-time-

Table 2

Correlations Among the Explicit and Implicit Self-Esteem Measures

	2	3	4	5	6	7	8	9	10
Direct measures									
1. RSES – self-report	.58**	.21*	.02	.00	-.03	.03	.03	.06	.21*
2. RSES – peer-report		.13	-.03	-.23*	.08	.11	.13	.21†	.24*
Indirect measures									
3. IAT (D_1-Measure)			.27**	.30**	-.19†	-.08	.20†	.08	.06
4. BIAT - Self-Other/Positive				.66**	.07	.11	-.04	.09	.15
5. BIAT - Positive-Negative/Self					-.11	-.10	-.09	-.04	.05
6. APT – RT-score (improved scoring algorithm)						.57**	.14	.19†	.08
7. APT – Response-Window Procedure (error-score)							.28**	.25*	.24*
8. ID-EAST – RT-score (improved scoring algorithm)								.63**	-.01
9. ID-EAST – error-score									.03
10. NLT – Ipsatized double-correction algorithm									-

Note. RSES = Rosenberg Self-Esteem Scale; IAT = Implicit Association Test; BIAT = Brief Implicit Association Test; APT = Affective Priming Tasks; ID-EAST = Identification Extrinsic Affective Simon Task; NLT = Name-Letter Task.

† $p < .10$. * $p < .05$. ** $p < .01$.

based APT. The marginally significant negative correlation between the IAT and the reaction-time-based APT was clearly unexpected.

Note that the high intercorrelations between the different scores of the BIAT and of the ID-EAST were a logical consequence of the respective calculation procedure, because the computation was, at least partly, based on the same trials. Contrary to this, the high intercorrelation of the APT standard procedure and the RW-APT ($r = .57$, $p < .01$) provided evidence of convergent validity because these two tasks were presented separately. Similar results were expected for the interrelations between the self-esteem IAT and both self-esteem BIATs, but intercorrelations were only moderate. The poor temporal stabilities of the self-esteem BIATs seem to be an obvious explanation.

2.4 Discussion

In this study, we aimed to increase the reliability of different implicit self-esteem measures. We started by nominating candidate measures for the reliable indirect measurement of self-esteem. Indeed, we found that the internal consistencies as well as the temporal stabilities of these measures were improved by changes in material, procedural, and analytic strategies. We will now evaluate each implicit self-esteem measure in more detail.

2.4.1 Implicit Association Test

The standard self-esteem IAT displayed the best internal consistency as well as the best temporal stability. However, in line with previous reports, whereas the split-half reliability of the IAT reached a level comparable to that of traditional questionnaire measures, its test-retest reliability did not reach the magnitude of those reported for direct measures (Egloff & Schmukle, 2002; Egloff et al., 2005; Nosek, Greenwald, & Banaji, 2007). Although this pattern of results suggests that the IAT captures state-specific variance in addition to trait-specific variance, this assumption could not be confirmed, at least for the IAT-anxiety (Schmukle & Egloff, 2004). Thus, other occasion-specific effects seem to contribute to the moderate stability of IAT effects (Schmukle & Egloff, 2005). For ex-

ample, thinking about the purpose of the self-esteem IAT as well as changes in test-taking strategy from session one to session two might decrease its test-retest reliability.

With regard to scoring procedures, the D_1 measure (Greenwald et al., 2003) clearly showed better psychometric properties than an IAT score based on a diffusion-model analysis. This leads us to ask: What factors might have caused the low reliability of the latter scoring procedure, which should actually be superior for theoretical reasons (Klauer et al., 2007)? First, although we already used the EZ-DM, which is particularly suited for individual subject analyses with only a moderate amount of data (Wagenmakers et al., 2007), more critical trials might nevertheless be necessary for a reliable measurement of IAT scores based on drift rates. Monte Carlo simulations by Wagenmakers et al. (2007) showed that drift rate estimates based on 50 (in contrast to 250) trials tended to be rather imprecise even when the EZ-DM was used (in our study, drift-rate estimates on the IAT were based on 80 trials). Second, although the EZ-DM does not explicitly address the RT-distribution of error responses, the usual error rate on the IAT might still be too low even for the EZ-DM. Wagenmakers et al. (2007) recommended that error rates be in the range of at least 5-10%, whereas in our study, 66% of participants had error rates less than 5%.[10] Thus, future studies using the diffusion model for analyzing the IAT should aim to increase both the number of critical trials and the error rate. The first condition could easily be fulfilled by lengthening the IAT; however, to increase the error rate, it would be necessary to design a more difficult IAT procedure (e.g., by using an adaptive response window procedure that forces decisions to be made under time pressure).

2.4.2 Brief Implicit Association Test

Whereas the internal consistencies of both implicit self-esteem BIAT scores were acceptable, their test-retest correlations were surprisingly low. We assume that the reduction of practice trials, especially the

[10] This recommendation is corroborated by an analysis of the 34 participants who made at least 5% errors on the self-esteem IAT. Mean internal consistency and test-retest reliability of the diffusion-model-based IAT score increased to moderate levels of .65 and .44, respectively.

absence of self-esteem-relevant practice trials, might have contributed to the instability of the self-esteem BIATs. This conclusion is supported by the observation of considerable differences in effect sizes for both sessions. For both self-esteem BIATs, the effect sizes from the first session were significantly lower than the effect sizes from the second session (for the *self-other/positive-[negative]* BIAT: $t(100) = -6.35$, $p < .001$; for the *positive-negative/self-[other]* BIAT: $t(100) = -4.03$, $p < .001$). Note that the effect sizes from only the second session were in the range reported for the standard self-esteem IAT. Thus, this pattern of results suggests that participants need either more practice trials or the full experience of the attitude-relevant BIAT procedure. Because the latter is less applicable, future research should explore whether including additional self-esteem-relevant practice trials improves the long-term stability of the self-esteem BIAT.

2.4.3 Affective Priming Tasks

In contrast to Bosson et al. (2000), who reported disappointing results for self-esteem APTs, we revealed two essential new findings. First, both APTs demonstrated the expected positivity bias, which is a characteristic feature of self-esteem measures. Second, reliability coefficients were substantially higher than those that have been reported in the present implicit self-esteem literature. Notably, compared to the established IAT, the quantitative difference in internal consistency is nearly eliminated, and for temporal stability, is substantially reduced.

The use of different material, procedural, and analytical innovations improved the reliabilities of both self-esteem APTs. Specifically, methodological improvements that contributed to the enhanced reliabilities consisted of enhancing the number of critical trials, presenting pictures of participants' faces as prime stimuli, presenting trials in a fixed order, and using supraliminal priming. For the reaction-time-based standard APT, the use of the combination of individually trimmed response latencies and the inclusion of error penalties in the analysis provided the best reliability coefficients. In the future, we suggest considering this new analytical ap-

proach for the computation of reaction-time-based affective priming effects.

The adaptive response-window priming technique for measuring implicit self-esteem yielded an even more promising result. For this method, we found an adequate internal consistency as well as a temporal stability coefficient with a magnitude in the neighborhood of the IAT. Please note that in contrast to the reaction-time-based APT, no further critical trials had to be eliminated to reach these levels of reliabilities.[11] Furthermore, the RW-APT was the only implicit self-esteem measure that showed convergent validity with structurally different indirect measures, namely the ID-EAST and the NLT. Thus, in the future, we suggest considering the adaptive response-window procedure of the APT as a valuable tool for assessing implicit self-esteem.

2.4.4 Identification – Extrinsic Affective Simon Task

The reaction-time-based ID-EAST scores did not show acceptable reliability coefficients. Although improved scoring led to better internal consistencies and temporal stabilities, the latter remained unsatisfactory (below .30). By contrast, the error score measure showed temporal stabilities as high as the RW-APT. The self-esteem ID-EAST—in particular the error score—is a promising implicit self-esteem measure, which should be further analyzed in future studies.

2.4.5 Name-Letter Task

The findings for the NLT resembled those of LeBel and Gawronski (2009). For both scoring algorithms, the internal consistency coefficients were comparably low. However, by considering that the computation of the NLT's split-half reliabilities is based on only two items, the magnitude of the observed measurement error is comparable to that reported for tra-

[11] As one reviewer pointed out, we do not know whether and how the standard priming procedure that preceded the RW-APT had an effect on the reliability of the second measure. Nevertheless, research using the response-window procedure in other areas of psychological research showed internal consistency coefficients similar to those found for our self-esteem RW-APT (see e.g., Cunningham et al., 2001; Steffens, Kirschbaum, & Glados, 2008).

ditional self-esteem scales by taking only two items into account (see LeBel & Gawronski, 2009, for further argumentation). With regard to temporal stability, the I-algorithm demonstrated its superiority over the B-algorithm. Thus, in line with LeBel and Gawronski (2009), we too recommend using the I-algorithm for the computation of self-esteem NLT scores.

2.4.6 Diffusion Model Analysis

Computing implicit self-esteem scores using a diffusion model analysis improved neither internal consistencies nor test-retest reliabilities for any of our two-choice response time tasks. In particular, reliabilities of scores based on a diffusion model analysis were even substantially lower than those found when using conventional scoring for the IAT, the BIAT, the standard APT, and the ID-EAST. Only for the RW-APT did a diffusion model analysis lead to rather similar reliabilities. As discussed for the IAT, we assumed that both the number of critical trials and the error rates might have been too low for a reliable diffusion model analysis. The comparatively better results for the RW-APT indicate that increasing the error rate of a task by using a response-window procedure might indeed lead to higher reliability estimates of diffusion-model-based scores. Future research should explore whether enhancing the number of critical trials might additionally improve reliability scores based on drift rates.

2.4.7 Intercorrelations of Self-Esteem Measures

Significant intercorrelations between self-esteem measures were observable only for some pairs of measures. There was a very sizeable relation between self-report and peer report measures of self-esteem. Nevertheless, the measures seem to assess different facets of explicit self-esteem because the magnitude of their intercorrelation was substantially lower than the stability coefficient of the RSES. Similar to previous research (e.g., Bosson et al., 2000; Greenwald & Farnham, 2000; Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005), explicit and implicit self-esteem measures were barely correlated.

In line with previous findings, we found mostly low or nonexistent intercorrelations between structurally different implicit self-esteem measures (see Bosson et al., 2000; Rudolph et al., 2008). Nevertheless, some intercorrelations among the more stable implicit self-esteem measures, in particular for the RW-APT, were noteworthy. However, future research should replicate these results before giving too much weight to them. In sum, all intercorrelations were considerably smaller in magnitude than those typically reported for explicit self-esteem measures. In light of the improved internal consistencies and temporal stabilities, the often discussed lack of reliability of indirect measures does not explain our pattern of results.

We rather assume that different implicit self-esteem measures "are likely to assess a broader range of disparate processes" (Gawronski & Bodenhausen, 2007, p. 273). Recently, Back, Krause, et al. (2009) provided empirical support for this hypothesis by showing that three prominent self-esteem measures (namely the RSES, the IAT, and the APT) independently predicted the same self-esteem-relevant criterion (metaperceptions of liking). Thus, Back, Krause, et al. (2009) concluded that structurally different self-esteem measures assess separate valid aspects of self-esteem. We suggest that future research shed more light on the differential predictive validities of different indirect measurement methods.

2.5 Limitations

Some limitations of the present study should be noted: First, the important question of whether the suggested material, procedural, and analytic improvements of the specific implicit self-esteem measures translate into increased validity cannot be answered with our data. Future research should shed more light on the predictive validity of the different implicit self-esteem measures presented in this study (for an initial study in this direction, see Back, Krause, et al., 2009).

Second, due to the fixed order of the tasks, we cannot rule out that effects and/or reliabilities of the implicit self-esteem measures were influenced by this specific order. We did not counterbalance the sequence of the tasks because our aim was to optimize the reliability of individual dif-

ferences (for a detailed argument, see Banse, 2001). At least our results are not suggestive of substantial order effects since effect sizes and reliabilities of the last three implicit self-esteem measures that were presented (NLT, IAT, ID-EAST) showed values in a range typically reported for these measures (Bosson et al., 2000; Rudolph et al., 2008). Moreover, work in other domains has not shown any relevant effect of task order on the reliabilities of personality IATs (Schmukle, Back, & Egloff, 2008).

2.6 Conclusions

Ten years after Bosson et al.'s (2000) article reporting weak psychometric properties of most indirect measures, we showed that the use of material, procedural, and analytic innovations substantially improved the reliabilities of different implicit self-esteem measures. In particular, the error scores of the APT and the ID-EAST as well as a double-corrected scoring algorithm of the NLT seem to be—complementary to the established IAT—promising variants for reliably capturing different facets of implicit self-esteem. In sum, our findings contribute to enhancing the development of more reliable implicit self-esteem measures, a prerequisite for analyzing the effects of this important personality dimension.

2.7 Aims of the Second Study

The results of Study 1 suggest that recently introduced material, structural, and analytic innovations can considerably enhance the reliability of different indirect assessment procedures. In particular, the ATP and the ID-EAST may be, at least for reliability reasons, valuable alternatives to the established self-esteem IAT and the NLT. Nevertheless, despite these promising findings, Study 1 did not answer the important question of whether the suggested improvements also translate into increased validity for self-esteem-relevant outcomes.

Study 2 marks the first step in this direction by investigating the predictive validity of the RW-APT, which was the most reliable alternative implicit self-esteem measure in the previous study. To explore its validity, a failure feedback task was presented by giving participants anagrams that were more difficult than expected. There were two reasons that

justified this approach: First, past research has shown that individual differences in self-esteem are more pronounced after experiencing failure (Brown & Dutton, 1995; for a review, see Baumeister, Campbell, Krueger, & Vohs, 2003). Second, the use of a negative feedback condition is in accordance with a similar approach that was applied by Greenwald & Farnham (2000) to determine the criterion validity of their—meanwhile established—self-esteem IAT.

During the anagram task, participants were asked to respond to three questions that served as self-esteem criteria: (a) before starting the task, participants were asked to indicate how many of the 20 anagrams they expected to solve (pre-task expectancy); (b) after completing the task, participants were asked to estimate their rate of correct responses (perceived performance); and (c) finally, participants were asked to indicate how many of 20 anagrams with similar difficulty on an upcoming test block they would expect to solve (post-task expectancy).

I expected that each of the criteria above would be predicted by the self-esteem RW-APT. Additionally, because participants might have the opportunity to control the three estimates of how many anagrams they would expect to solve (or did solve) correctly, direct measures of self-esteem should also predict each of the three criteria. Furthermore, it was assumed that explicit and implicit self-evaluations would be uncorrelated because direct and indirect assessment procedures measure distinct representations of self-esteem (i.e., controlled vs. automatic self-evaluations). Consequently, each kind of self-esteem measure should independently explain criterion variance. Thus, Study 2 aimed to provide evidence for (a) the predictive validity of the self-esteem RW-APT and, more importantly, (b) the necessity of using indirect measures as complements to direct measures.

3 Predictive Validity of the Response-Window Affective Priming Task for Assessing Self-Esteem

3.1 Introduction

Over the past 15 years, the use of a new class of indirect measurement techniques has become increasingly important in many fields of psychological research (for a review, see Fazio & Olson, 2003). These so-called *implicit measures* might be valuable complements to self-report questionnaires because they (a) do not rely on the respondent's willingness to report self-relevant knowledge and (b) assess evaluations that are activated automatically (i.e., without conscious effort) in response to an attitude object.

The measurement of self-esteem is one of the most important application areas for these new assessment methods. According to theoretical considerations of distinct implicit and explicit self-esteem constructs, implicit self-esteem (ISE) measures assess how people automatically evaluate their own self-worth, whereas explicit self-esteem (ESE) measures assess how people deliberatively evaluate the self (Greenwald & Banaji, 1995). Studies have provided empirical support for these assumptions by showing that ISE measures have predicted relevant outcomes above and beyond ESE measures (e.g., Back, Krause, et al., 2009; Greenwald & Farnham, 2000).

Whereas a variety of questionnaires can be used for the assessment of ESE, for ISE so far, only two measures—the self-esteem IAT and the NLT—meet the requirements of reliability and validity (Bosson et al., 2000). An interesting alternative class of indirect measurement is the APT (for an overview, see Fazio, 2001) because—in contrast to many other indirect measures—a well-established theoretical understanding of how affective priming effects arise is available (Fazio & Olson, 2003). However, whereas APTs are one of the most prominent implicit attitude measures (De Houwer et al., 2009; Fazio & Olson, 2003), they are seen as less applicable for the assessment of ISE due to low reliabilities (Bosson et al., 2000). Therefore, we think it is worthwhile to design an adaption of the self-esteem APT that—compared to previous versions—shows both a bet-

ter reliability and, consequently, a better predictive validity for self-esteem-relevant outcomes.

3.1.1 The Self-Esteem Affective Priming Task

In a typical self-esteem APT, participants categorize unambiguously valenced target stimuli into a positive or negative category. The key feature of the APT is that each target is preceded by a self-related or other-related prime. Self-esteem APTs quantify the degree to which the presentation of a prime facilitates the categorization task. The basic assumption is that—due to the experience of positive affect following self-related stimuli—people with high self-esteem are thought to be faster when a positive stimulus is preceded by a self-related prime and slower when a negative stimulus is preceded by the same prime. As the first evidence of construct validity, the outcome of a self-esteem APT should exhibit a positivity bias because most people keep in mind a positive explicit and implicit self-evaluation (Bosson et al., 2000).

We used four material and procedural innovations for improving the usually disappointing reliability of the self-esteem APT: (a) increasing the number of critical trials; (b) using a picture of the participant as a self-related prime instead of words; (c) presenting critical trials in a fixed order (Banse, 2001); and (d) using an adaptive response-window technique (Draine & Greenwald, 1998) to circumvent effects of individually different speed-accuracy-tradeoffs. In the previous study, I was able to show that these innovations substantially improved the internal consistency and the temporal stability of the RW-APT. As a next step, the goal of the present study was to analyze the predictive validity of this improved self-esteem RW-APT.

3.1.2 Self-Esteem in the Face of Failure

Past research has shown that there are particularly pronounced differences between individuals with high or low self-esteem in dealing with the experience of failure (e.g., Brown & Dutton, 1995). That is, whereas all persons, independent of their self-esteem, tend to accept success, those with high self-esteem are more apt to reject the negative implications of

failure than those with low self-esteem. Thus, high self-esteem can function as a buffer against negative feedback (Brown, in press; Brown & Dutton, 1995; Greenwald & Farnham, 2000).

Previous studies that relied on consequences of self-reported self-esteem before and after negative feedback in achievement tasks have concluded that ESE (a) is positively related to performance expectancies (Dutton & Brown, 1997; McFarlin & Blascovich, 1981; Shrauger, 1975), (b) predicts perceived performance after failure (Brown & Dutton, 1995; Greenwald & Farnham, 2000), and (c) predicts the confidence of performing well on a subsequent task after failure (Baumeister & Tice, 1985; Greenwald & Farnham, 2000; McFarlin & Blascovich, 1981). According to these findings, giving participants failure feedback seems to be well suited for the investigation of individual differences in ESE.

There is now growing evidence that behavior-related responses are determined not only by deliberative but also by automatic processes. In their seminal paper, Greenwald and Farnham (2000) found evidence for the predictive validity of the—meanwhile established—implicit self-esteem IAT for responses to a failure feedback (for more research on predicting relevant outcomes by ISE, see Back, Krause, et al., 2009; McGregor & Jordan, 2007; for research on predicting behavior by explicit and implicit self-concepts of personality, see Back et al., 2009). Therefore, in accordance with these results, we expected that each of the three criteria above would be predicted by both ESE *and* ISE.[12]

To explore the predictive validity of our adopted self-esteem RW-APT, we used an anagram task, which is a popular and easy way to induce a failure experience by giving participants anagrams that are more difficult than expected (e.g., Baumeister & Tice, 1985; Egloff & Krohne, 1996). As we assume that reactions to failure are affected by both ESE and ISE, both types of self-esteem should demonstrate predictive validity for responses before and after solving the anagram task. Moreover, be-

[12] There has also been research that has examined the predictive validity of the interaction of ESE and ISE for relevant outcomes: For example, Jordan et al. (2003) revealed that individuals with defensive high self-esteem (i.e., the combination of high ESE and low ISE) showed the highest levels of narcissism and behaved more defensively than individuals with secure high self-esteem (i.e., high ESE and high ISE).

cause ISE and ESE measures capture different aspects of self-esteem (Back, Krause, et al., 2009; Bosson et al., 2000), it is plausible that the use of both types of measures should result in a better prediction of reactions to failure. Thus, we expected that the two types of self-esteem measures would show incremental validity to each other. The focus of the current study was therefore whether the self-esteem RW-APT would predict relevant behavior over established ESE measures.

3.2 Method

3.2.1 Participants

Students from different fields of study participated in exchange for research participation credit or monetary compensation. The data of two subjects were discarded for having error rates on the RW-APT in excess of 45%. One participant was dropped for having 0% hits within the response window. The data of another subject were eliminated for reporting an exaggerated performance expectation (number of correct answers = 100%) for the test block of the anagram task. The average age of the remaining 97 participants (74 women) was 23.6 years (*SD* = 3.6).

3.2.2 Procedure

Upon arrival at the laboratory, participants were placed in separate cubicles. Up to six participants were tested at the same time. First, participants completed self-esteem measures, starting with ESE tasks. Subsequently, participants were asked to complete the anagram task. Participants' changes in affective state were obtained by a mood adjective scale. At the end of the session, participants were debriefed carefully about the purpose of this study and the objective difficulty of the anagram task.

3.2.3 Self-Esteem Measures

3.2.3.1 Explicit self-esteem

For obtaining a comprehensive measure of ESE, participants completed two different questionnaires: The Self-Attributes Questionnaire

(SAQ; Pelham & Swann, 1989) and the Multidimensional Self-Esteem Scale (MSES; Schütz & Sellin, 2006). The SAQ measures respondents' confidence about their standing relative to same-sex and same-age students. Participants estimated their standing on different domains relative to their peers using 10-point scales ranging from A (*bottom 5%*) to J (*top 5%*). The MSES—a German adaption of Fleming and Courtney's (1984) Self-Rating Scale—was applied as a second direct measure of people's conscious feelings of self-esteem. The total score was obtained from 32 items, each measured on a 7-point Likert scale ranging from 1 (*never*) to 7 (*always*).

3.2.3.2 Response-Window Affective Priming Task (RW-APT)

Pictures of participants' faces were presented as self-related primes. Pictures of faces of two different unknown persons of the same sex who did not participate in the experiment served as other-related primes. During the task, participants were asked to identify unambiguously valenced adjectives as pleasant (*fair, honest, healthy, beautiful, sincere*) or unpleasant (*mean, envious, cowardly, vicious, ruthless*) as accurately as possible. The time course of the RW-APT was as follows: Each judgment trial started with the presentation of a prime, which remained on the screen for 67 ms and was immediately replaced by a blank screen for 33 ms. Then, the target word was presented with a constant prime-target SOA of 100 ms, and 300 ms after the target stimulus had appeared, a response window was opened for 150 ms. Participants were instructed to respond to the target stimulus within this response window using external response pads. No feedback was given for incorrect responses. The intertrial interval was set to 1000 ms. Three practice blocks were followed by six critical blocks of 20 trials each. Depending on the participant's average accuracy and speed in the preceding critical block, the response-window was adjusted by default (Draine & Greenwald, 1998).

Because participants were required to respond within a 150-ms response window, responses that occurred 50 ms before the start of the response window as well as responses that occurred 250 ms after the end of the response window were deleted (5.14%). For each participant, a posi-

tivity index for the self-related prime was calculated by subtracting the average error rate for pleasant targets from the average error rate for unpleasant targets that followed the same prime. Likewise, a positivity index for other-related primes was computed to control for the preference of positive versus negative target adjectives in the analysis. To yield a single self-esteem priming effect, the positivity index for the self-related prime was regressed on the positivity index for other-related primes, and the residual was used as the ISE score.

3.2.4 Criterion Measure

3.2.4.1 Anagram task

After completing the self-esteem measures, participants had to solve an anagram task that was presented as a test of verbal ability (Egloff & Krohne, 1996). Before the start of the task, participants filled out a mood scale that consisted of nine adjectives (*dissatisfied, displeased, happy, enraged, dejected, tense, excited, humiliated, balanced*) as a baseline measure of their emotional state ($\alpha = .83$). To create the experience of failure in the anagram task, the difficulty of the items was manipulated: Whereas three fourths of the anagrams were difficult to solve, only a quarter of the anagrams was easy to solve (see Appendix B for the complete list of the German anagrams). A pilot study confirmed the actual success rate of 5 out of 20 anagrams. Each anagram was presented on the screen for 7 s. When the 7 s had elapsed, participants were asked to give their solution within a time span of 5 s. Then the correct solution was presented for 3 s. After a break of 1 s, the next anagram followed. Participants first completed four training anagrams to become familiar with the task. Two very easy and two very difficult exemplars were presented to strengthen the plausibility of the instructions (for exact wording, see Egloff & Krohne, 1996).

Before starting the test, *pre-task expectancy* was obtained by asking participants how many of the 20 anagrams they expected to solve. Next, 5 easy and 15 difficult anagrams, each consisting of five letters, were presented in a fixed random order. After completing the test block, partici-

pants were asked to estimate their rate of correct responses (*perceived performance*). Then feedback about their actual performance was given. Next, participants were asked to indicate how many of 20 anagrams with similar difficulty on an upcoming test block they expected to solve (*post-task expectancy*). After answering this question, participants were given the adjective mood scale to measure their affective state following the anagram task ($\alpha = .91$).

3.2.4.2 Dependent Measures

Based on our theoretical considerations and previous findings, we created three self-esteem criteria: (a) *Pre-task expectancy bias* – To derive a criterion to measure a self-serving bias with regard to performance expectancy, test performance was subtracted from the pre-task expectancy. Based on our failure feedback condition, we hypothesized that participants with high self-esteem would show a higher difference between their pre-task expectancy and their actual performance than participants low in self-esteem. (b) *Perceived performance bias* – To derive a measure of self-serving bias with regard to past test performance, we subtracted the participants' number of correct responses from their estimated number of correctly solved anagrams. Whereas positive scores on this index indicate high self-esteem, negative scores indicate low self-esteem. (c) *Post-task expectancy bias* – As a further measure of a self-serving bias regarding future performance expectancy after failure experience, we subtracted the participants' number of correct responses from the number of anagrams they expected to solve in a second test block. Again, positive scores indicate high self-esteem, and negative scores are seen as an indicator of low self-esteem. According to findings of weak or nonexistent correlations between explicit and implicit self-esteem (e.g., Back, Krause, et al., 2009; Bosson et al., 2000), and because behavior-related responses are determined by both deliberative and automatic processes, we expected that ESE and ISE measures would independently predict each of the three criteria above.

3.3 Results

3.3.1 Self-Esteem Measures

To yield a single index of ESE, scores on the SAQ ($M = 6.03$, $SD = 1.06$) and the MSES ($M = 4.64$, $SD = 0.95$) were first standardized and then averaged together ($\alpha = .76$). For the self-esteem RWP, the positivity index for the self-related prime—that is, the difference between the mean error rate for unpleasant targets ($M = 0.32$, $SD = 0.17$) and the mean error rate for pleasant targets ($M = 0.19$, $SD = 0.10$)—differed significantly from zero ($M = 0.13$, $SD = 0.20$), $t(96) = 6.69$, $p < .001$, $d = 0.96$.[13] Thus, the expected positivity bias was obtained, indicating that people's general tendency to have a positive self-view was thus reflected in this measure.

The internal consistency of the self-esteem RWP was estimated by separately applying the scoring algorithm to the six critical blocks. Note that the positivity index for other-related primes was regressed on the positivity index for the self-related prime, and the residual served as a single index of ISE in each block. In line with previous findings (see Chapter 2 of this thesis), Cronbach's alpha of the newly introduced self-esteem RW-APT was satisfactory ($\alpha = .75$) and showed a value typically reported for ESE measures. As expected, ISE and ESE measures were not significantly correlated, and could thus be considered independent indicators of self-esteem ($r = -.01$).

3.3.2 Anagram Task

Two results suggest that the manipulation of item difficulty successfully induced an experience of failure: First, on average participants solved 5 out of 20 anagrams ($M = 5.04$, $SD = 3.19$). Second, compared to pre-task mood ($M = 2.39$, $SD = 0.77$), participants reported a significantly more negative mood after completing the anagram task ($M = 2.84$, $SD = 1.03$), $t(96) = -5.16$, $p < .001$, $d = 0.50$. According to our anagram instruc-

[13] The positivity index for other-related primes—mean error rate for unpleasant targets ($M = 0.25$, $SD = 0.14$) minus the mean error rate for pleasant targets ($M = 0.23$, $SD = 0.13$)—did not differ significantly from zero ($M = 0.02$, $SD = 0.18$), $t(96) = 1.15$, $p = .25$, $d = 0.16$.

tions, participants' pre-task expectations of their performance (M = 9.43, SD = 3.24) were only slightly lower than the announced success rate of 10 anagrams, $t(96) = -1.72$, $p = .09$, $d = 0.17$. Due to our failure feedback, a highly significant difference between pre-task expectations and actual performance was obtained, $t(96) = 11.29$, $p < .001$, $d = 1.36$. Participants' perceived mean level of performance rated after the anagram task (M = 4.84, SD = 2.91) did not differ significantly from their mean level of actual performance, $t(96) = 1.53$, $p = .13$, $d = 0.07$. As expected, participants' post-task expectations of success in a future anagram task (M = 5.23, SD = 2.98) was adjusted and did not differ significantly from their actual performance in the test block, $t(96) = -1.38$, $p = .17$, $d = 0.06$.

3.3.3 Predictive Validities

The first four columns of Table 3 show descriptive statistics of our self-esteem criteria derived from the anagram task. To examine the predictive validity of ESE and self-esteem RW-APT scores, we first computed correlations between the self-esteem criteria and ESE and self-esteem RW-APT. As can be seen in Column 5 of Table 1, the direct measure of self-esteem was significantly correlated with all criterion variables (for perceived performance, the correlation fell short of the conventional level of significance with $p = .07$). Moreover, the same pattern of results was obtained for the self-esteem RW-APT (Column 6 of Table 1). Thus, ESE and ISE were both significant predictors of reactions before and after the anagram task.[14]

According to our incremental validation strategy, we examined the amount of variance in dependent variables actually accounted for by ISE in addition to ESE. Therefore, we performed hierarchical regressions with pre-task expectancy bias, perceived performance bias, and post-task expectancy bias as separate criteria. In each of the three regressions, the ESE score was entered in Step 1, and the self-esteem RW-APT score was entered in Step 2. As can be seen in the eighth Column of Table 1, the self-esteem RW-APT significantly predicted each validity criterion above and

[14] The interaction of ESE and ISE did not significantly predict any of the three criteria: for pre-task expectancy bias, $\beta = .04$, $t(93) = 0.40$, $p = .69$; for perceived performance bias, $\beta = .08$, $t(93) = 0.82$, $p = .41$; for post-task expectancy bias, $\beta = .00$, $t(93) = -0.03$, $p = .97$.

Table 3

Predictive Validities of Explicit and Implicit Self-Esteem Measures

	Descriptives				Correlations		Hierarchical regression		
							Step 1 ESE	Step 2 RW-APT	ESE+RW-APT
Self-esteem criteria	Min	Max	*M*	*SD*	ESE	RW-APT	R^2	ΔR^2	R^2
Pre-task expectancy bias	-4.00	13.00	4.39	3.83	.35**	.27**	.12**	.07*	.19**
Perceived performance bias	-5.00	3.00	-0.21	1.33	.19†	.27**	.03†	.07*	.11**
Post-task expectancy bias	-5.00	3.00	0.19	1.33	.30**	.34**	.09**	.12**	.21**

Note. ESE = Explicit self-esteem measures; RW-APT = Response-Window Affective Priming Task.
$^{\dagger}p < .10$. *$p < .05$. **$p < .01$.

beyond ESE. The last column of Table 1 shows for each validity criterion the amount of variance explained by both types of self-esteem.[15]

3.4 Discussion

The present study confirmed the usefulness of a recently introduced adaption of the self-esteem APT, the adaptive response-window technique, for the assessment of ISE. In line with previous findings (see study 1), the self-esteem RW-APT demonstrated both the expected positivity bias and, more importantly, a satisfactory internal consistency. Moreover, this study provides compelling evidence for the predictive validity of the self-esteem RW-APT. According to our incremental validation strategy, the self-esteem RW-APT predicted several self-esteem criteria independently of ESE measures.

These findings are in line with multiple process accounts of self-esteem (e.g., Back, Krause, et al., 2009; Greenwald & Banaji, 1995). Moreover, they extend previous findings on the intrapersonal consequences of self-esteem in a performance context. Both ESE and ISE independently predicted the evaluation of a person's own (low) performance as well as expectations of success prior to and after the experience of failure:

(a) Both explicit and implicit self-esteem were positively related to the expectation of success. Whereas people with high self-esteem generally expect to be successful, those with low self-esteem are more modest to prevent negative or disappointing outcomes (Dutton & Brown, 1997; Shrauger, 1975).

(b) Compared to their actual performance, people with high explicit and implicit self-esteem perceived their performance more favorably than those with low self-esteem. Thus, people

[15] As an alternative method of analysis, we residualized each criterion for the participant's actual test performance. Subsequently, ESE and ISE were correlated with each residualized score. Results were highly similar for all self-esteem criteria, $r = .25$, $p < .05$ (implicit) and $r = .32$, $p < .01$ (explicit) for pre-task expectancy, $r = .25$, $p < .05$ (implicit) and $r = .14$, $p = .18$ (explicit) for perceived performance, and $r = .33$, $p < .01$ (implicit) and $r = .26$, $p < .05$ (explicit) for pre-task expectancy.

with high self-esteem are more generous in their performance appraisals than people with low self-esteem (Brown & Dutton, 1995).

(c) After the experience of failure, people with high explicit and implicit self-esteem expected to be more successful on a subsequent task than people with low self-esteem. Thus, people with high self-esteem compensate for negative feedback better by responding with a more optimistic outlook in a subsequent task (McFarlin & Blascovich, 1981). Low self-esteem people, however, accept negative feedback more because such feedback is consistent with their chronic self-evaluation (Shrauger, 1975).

Taken together, our results indicate that high ISE as well as high ESE leads to a more optimistic outlook on future performances and buffers reactions to negative feedback. Insofar, high ISE seems also to be an important protective factor in the face of failure.

3.5 Conclusion

Since the influential study of Bosson et al. (2000), the APT was mostly considered to be inappropriate for ISE research. In this paper, we demonstrated that the adaptive response-window technique is a valuable tool for assessing ISE with the APT. Our self-esteem RW-APT was satisfactorily reliable and, more importantly, showed predictive validity for self-esteem-relevant outcomes before and after experiencing failure. Moreover, the self-esteem RW-APT incrementally predicted each criterion over established ESE measures. In summary, we suggest that this indirect assessment method has the potential to complement the arsenal of implicit self-esteem measures: The RW-APT might be an important addition to self-reports when it comes to predicting self-esteem-relevant outcomes.

4 Global Discussion

Since the beginning of my research activity in 2006, the use of indirect measures for assessing self-esteem has become even more popular. A search in "Web of Science" with the keyword "implicit self-esteem" yielded an increase of almost 100% of related papers from 2006 (11) to 2007 (21). After a decrease in 2008 (13), a new peak of published studies was reached in 2009 (25). Thus, there exists a growing interest in assessing how people automatically (i.e., without deliberative effort) evaluate their self-worth.

However, to date, most of these studies have used only two indirect assessment procedures to measure people's automatic self-evaluations: the self-esteem IAT and the NLT. The limited application of only two measures from the variety of indirect assessment procedures is at least partially due to the results of a very influential study by Bosson et al. (2000). In their study, the self-esteem IAT and the NLT showed the highest internal consistencies and test-retest reliabilities. Consequently, Bosson et al.'s (2000) findings have resulted in the widespread opinion in the research community that other indirect measurement procedures are not able to reliably assess implicit self-esteem (and other implicit attitudes; see e.g., Asendorpf et al., 2002; Fazio & Olson, 2003; Nosek et al., 2007; Payne & Gawronski, 2010). This often-cited belief constitutes the main motivation for my thesis. My primary aim was to investigate whether and how internal consistencies and test-retest reliabilities of different indirect measures may be enhanced by using the latest knowledge of the past decade. To enable optimal comparability to the study of Bosson et al. (2000), the same time lag of 4 weeks between measurement occasions was used.

In the first study, I investigated how recently introduced material, structural, and analytic innovations influence the reliability estimates for six of the currently most popular indirect measurement techniques. Beside the IAT and NLT, a shortened variant of the IAT, the BIAT, two forms of the APT—namely the reaction-time-based standard APT and the error-based RW-APT—and, finally, the ID-EAST were adopted for the indirect assessment of self-esteem.

From a psychometric perspective, the results of the first study revealed that, in the last decade, no really important news was delivered for the established self-esteem IAT and the NLT: The internal consistencies and test-retest reliabilities reached expected and known values. Moreover, both kinds of reliability estimates were slightly lower as reported by Bosson et al. (2000). For the newly introduced BIAT, internal consistencies but not temporal stabilities were on the same level as they were for the standard IAT. In summary, the first study provided further evidence that the standard self-esteem IAT and the NLT may be reliable complements to direct measures of self-esteem. However, the use of the two assessment procedures for psychological research questions is still limited because their 4-week stabilities, although moderate, did not reach the level that is required for individual diagnostic assessment. Nevertheless, the reliabilities were good enough to use these measures for correlational studies.

Whereas the results for the established indirect measures were not surprising, the first study provided promising reliability estimates for less-established implicit self-esteem measures, namely the ID-EAST and two variants of the APT (i.e., the standard APT and the RW-APT). For each of the three measures, the internal consistencies and temporal stabilities were substantially improved by using material, structural, and analytic innovations that had been introduced in the last decade.

For further research with the self-esteem ID-EAST, I recommend analyzing errors instead of response latencies because the former scoring procedure was substantially more stable. However, if someone wants to compute effects and reliabilities by using reaction times, it would be more appropriate to take individually trimmed response latencies and error penalties for error responses into account. Taken together, effect sizes and reliability estimates of the self-esteem ID-EAST—computed by two alternative scoring procedures—were comparable to the values of the established implicit self-esteem measures. For further use in the self-esteem domain, it is necessary to investigate the ID-East's predictive validity for self-relevant outcomes. If this evidence can be provided, the ID-EAST has the potential to become an alternative tool for the indirect assessment of self-esteem.

Even more promising results were obtained for the APT. Both the internal consistencies and the test-retest correlations showed values equivalent to those of the established implicit self-esteem measures. For further application of the self-esteem APT, I recommend using (a) pictures of the participants to prime the automatic self-related evaluation, (b) an adequate number of critical trials, which should be presented in a fixed randomized order, (c) the response-window technique (i.e., the RW-APT) for reliability reasons. If someone decides to apply the standard reaction-time-based APT, effects and reliabilities should be computed by using both individually trimmed response latencies and error penalties for error trials.

Regarding the scoring algorithm for the APT, there was a change in the computation of self-esteem priming effects from Study 1 to Study 2. That is, instead of calculating two difference scores—one for the positivity index for self-related primes and one for the positivity index for other-related primes—that were subtracted from each other in Study 1, the positivity index for the self-related prime was regressed on the positivity index for the other-related primes. For Study 2, this new method of computing priming effects resulted in higher reliability estimates and, more importantly, higher validity coefficients in comparison to the standard scoring procedure used in Study 1.

A subsequent analysis of the impact of this new scoring algorithm on the reliability estimates of Study 1 yielded increases for both variants of the APT. That is, for the standard APT, the average internal consistency of both sessions increased from .71 to .75.[16] For the RW-APT, the mean internal consistency increased from .70 to .71. More importantly, the new scoring algorithm had a considerable impact on the test-retest reliabilities, in particular for the response-window priming: Whereas the 4-week stability of the standard APT increased from .32 to .37, the test-retest correlation of the RW-APT increased from .43 to .58. These results show that using this new scoring algorithm leads to reliability coefficients of the self-esteem RW-APT that are equal to the values of the self-esteem

[16] As reference values for the standard reaction-time-based ATP, I report the results of the improved scoring algorithm with individually trimmed response latencies and error penalties for error trials.

IAT (its mean internal consistency was .79, its 4-week stability was .54). Thus, especially for reliability reasons, response-window priming seems to be an actual alternative to the currently most popular implicit self-esteem measures.

To establish the RW-APT for the self-esteem domain, it was necessary to validate this measure on relevant outcomes.[17] Accordingly, I employed an anagram task in which participants received anagrams that were more difficult than expected. This approach was chosen because previous research has shown that self-esteem differences are more pronounced after experiencing failure (for a review, see Baumeister et al., 2003). During the task, three self-esteem criteria were obtained by asking participants how many anagrams they (a) expected to solve before the start of the task (pre-task expectancy), (b) thought they had solved after the task (perceived performance), and (c) expected to solve in an upcoming test block with similar difficulty (post-task expectancy).

As hypothesized, the results of Study 2 showed that the new self-esteem RW-APT demonstrated predictive validity for each of the three self-esteem-relevant outcomes. Moreover, the findings of this study confirmed the positive results of the previous reliability study because the internal consistency of the self-esteem RW-APT was satisfying and (nearly) identical to that of the direct measures of self-esteem. Taken together, these results suggest that the response-window technique is an important innovation, at least for the APT, for the reliable and valid assessment of implicit self-esteem.

Additionally, the pattern of results of Study 2 provided further evidence for the application of direct *and* indirect measures in self-esteem research: First, as expected, explicit self-esteem also predicted each of the three relevant outcomes. Second, explicit self-esteem was uncorrelated with implicit self-esteem. Thus, direct and indirect measures of self-esteem independently predicted each of the three criteria. In other words, both types of self-esteem demonstrated incremental validity suggesting that the combination of direct and indirect measures may increase the probability of explaining the highest amount of variance in self-esteem

[17] For the validation of the standard reaction-time-based APT, see Back, Krause, et al., 2009.

relevant outcomes. In summary, the findings of Study 2 determined the usefulness of employing indirect measures as complements to direct measures.

In order to summarize the results of this thesis, two essential new findings were derived regarding the measurement of implicit self-esteem: First, whereas recently introduced innovations did not improve the reliability estimates for established and often-used indirect measures of self-esteem, namely the IAT and the NLT, the improvements were substantial for less-established implicit self-esteem measures, namely the ID-EAST and the APT. Consequently, the results of the first study suggest that the reliability estimates of all four assessment methods justify their use in future research settings.

Second, the response-window procedure is a promising assessment technique, at least for affective priming, that should be more frequently employed in the future. The internal consistencies and test-retest reliabilities of the self-esteem RW-APT were equal to the values of the self-esteem IAT. Moreover, the RW-APT demonstrated predictive validity for relevant outcomes in an experimental setting that was almost identical to that used to determine the validity of the IAT. In the future, it would be very fruitful to investigate how the response-window technique may improve the psychometric properties of other reaction-time-based measures of self-esteem.

5 General Conclusions

Fifteen years ago, Greenwald and Banaji (1995) concluded in their seminal paper that the development of indirect measures could be the hallmark of a new industry of research as long as these measures meet the usual psychometric standards for individual difference measures. In the implicit self-esteem domain, currently there have been no indirect measures that have fulfilled all criteria, whereby most of them have suffered from low test-retest correlations. Nevertheless, with the introduction of the self-esteem IAT in 2000 by Greenwald and Farnham, this research exploded because the IAT nearly met the required criteria. In my thesis, I was able to show that the material, structural, and analytical innovations of the last 10 years have substantially improved the reliabilities of alternative indirect measures of self-esteem. Moreover, the results of Studies 1 and 2 suggest that the self-esteem RW-APT is just as reliable and valid as the popular self-esteem IAT. I hope that future research will frequently employ this measure and provide more evidence for its reliability and validity.

6 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.

Baccus, J. R., Baldwin, M. W., & Packer, D. J. (2004). Increasing implicit self-esteem through classical conditioning. *Psychological Science, 15*, 498-502.

Back, M. D., Krause, S., Hirschmüller, S., Stopfer, J., Egloff, B., & Schmukle, S. C. (2009). Unraveling the three faces of self-esteem: A new information-processing sociometer perspective. *Journal of Research in Personality, 43*, 933-937.

Back, M. D., Schmukle, S. C., & Egloff, B. (2005). Measuring task-switching ability in the Implicit Association Test. *Experimental Psychology, 52*, 167-179.

Back, M. D., Schmukle, S. C., & Egloff, B. (2009). Predicting actual behavior from the explicit and implicit self-concept of personality. *Journal of Personality and Social Psychology, 97*, 533-548.

Banaji, M. R. (2001). Implicit attitudes can be measured. In H. L. Roediger & J. S. Nairne (Eds.), *The nature of remembering: Essays in remembering Robert G. Crowder* (pp. 117-150). Washington, DC: American Psychological Association.

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., & Greenwald, A. G. (2007). Personality and implicit social cognition research: Past, present and future. *European Journal of Personality, 21*, 371-382.

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.

Bar-Anan, Y., Nosek, B. A., & Vianello, M. (2009). The sorting paired features task: A measure of association strengths. *Experimental Psychology, 56*, 329-343.

Barnes-Holmes, D., Hayden, E., Barnes-Holmes, Y., & Stewart, I. (2008). The Implicit Relational Assessment Procedure (IRAP) as a response-time and event-related-potentials methodology for testing natural verbal relations: A preliminary study. *Psychological Record, 58*, 497-515.

Baumeister, R. F., Campbell, J. D., Krueger, J. I., & Vohs, K. D. (2003). Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? *Psychological Science in the Public Interest, 4*, 1-44.

Baumeister, R. F., & Tice, D. M. (1985). Self-esteem and responses to success and failure: Subsequent performance and intrinsic motivation. *Journal of Personality, 53*, 450-467.

Blanton, H., Jaccard, J., Gonzales, P. M., & Christie, C. (2006). Decoding the Implicit Association Test: Implications for criterion prediction. *Journal of Experimental Social Psychology, 42*, 192-212.

Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single-Target IAT (ST-IAT): Assessing automatic affect towards multiple attitude objects. *European Journal of Social Psychology, 38*, 977-997.

Borkenau, P., & Mauer, N. (2007). Well-being and the accessibility of pleasant and unpleasant concepts. *European Journal of Personality, 21*, 169-189.

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.

Breuer, J., & Freud, S. (1895). *Studien ueber Hysterie*. Leipzig, Germany: Deuticke.

Brown, J. D. (in press). High self-esteem buffers negative feedback: Once more with feeling. *Cognition and Emotion*.

Brown, J. D., & Dutton, K. A. (1995). The trill of victory, the complexity of defeat: Self-esteem and people's emotional reactions to success and failure. *Journal of Personality and Social Psychology, 68*, 712-722.

Brown, R. P., & Zeigler-Hill, V. (2004). Narcissism and the non-equivalence of self-esteem measures: A matter of dominance? *Journal of Research in Personality, 38*, 585-592.

Brunel, F. F., Tietje, B. C., & Greenwald, A. G. (2004). Is the Implicit Association Test a valid and valuable measure of implicit consumer social cognition? *Journal of Consumer Psychology, 14*, 385-404.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6*, 475-494.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*, 3-31.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York; NY: Harper Row.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349-354.

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

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; NY: Erlbaum.

De Houwer, J. (2006). What are implicit measures and why are we using them. In R. W. Wiers & A. W. Stacy (Eds.), *The handbook of implicit cognition and addiction* (pp. 11-28). Thousand Oaks, CA: Sage Publishers.

De Houwer, J., & De Bruycker, E. (2007a). The identification-EAST as a valid measure of implicit attitudes toward alcohol-related stimuli. *Journal of Behavior Therapy and Experimental Psychiatry, 38*, 133-143.

De Houwer, J., & De Bruycker, E. (2007b). The Implicit Association Test outperforms the extrinsic affective Simon task as an implicit meas-

ure of inter-individual differences in attitudes. *British Journal of Social Psychology, 46*, 401-421.

De Houwer, & Moors, A. (2010). Implicit measures: Similarities and differences. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 176-193). New York, NY: Guilford Press.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin, 135*, 347-368.

Demo, D. H. (1985). The measurement of self-esteem: Refining our methods. *Journal of Personality and Social Psychology, 48*, 1490-1502.

Dijksterhuis, A. (2004). I like myself but I don't know why: Enhancing implicit self-esteem by subliminal evaluative conditioning. *Journal of Personality and Social Psychology, 86*, 345-355.

Draine, S. C., & Greenwald, A. G. (1998). Replicable unconscious semantic priming. *Journal of Experimental Psychology: General, 127*, 286-303.

Dutton, K. A., & Brown, J. D. (1997). Global self-esteem and specific self-views as determinants of people's reactions to success and failure. *Journal of Personality and Social Psychology, 73*, 139-148.

Egloff, B., & Krohne, H. W. (1996). Repressive emotional discreteness after failure. *Journal of Personality and Social Psychology, 70*, 1318-1326.

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.

Egloff, B., Schwerdtfeger, A., & Schmukle, S. C. (2005). Temporal stability of the Implicit Association Test-Anxiety. *Journal of Personality Assessment, 84*, 82-88.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist, 49*, 709-724.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna

(Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75-109). San Diego, CA: Academic Press.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion, 15*, 115-141.

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.

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitude. *Journal of Personality and Social Psychology, 50*, 229-238.

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the Implicit Association Tests. *Basic and Applied Social Psychology, 27*, 307-316.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I", the "A", and the "T": A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology, 17*, 74-147.

Fleming, J. S., & Courtney, B. E. (1984). The dimensionality of self-esteem II: Hierarchical facet model for revised measurement scales. *Journal of Personality and Social Psychology, 46*, 404-421.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*, 692-731.

Gawronski, B., & Bodenhausen, G. V. (2007). What do we know about implicit attitude measures and what do we have to learn? In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 265-286). New York: Guilford.

Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are „implicit“ attitudes unconscious? *Consciousness and Cognition, 15*, 485-499.

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us? Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science, 2*, 181-193.

Graf, P., & Schacter, D. (1985). Implicit and explicit memory for new associations in normal and amnesic subjects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 501-518.

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., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109*, 3-25.

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., & Banaji, M. R., (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*, 197-216.

Grumm, M., Erbe, K., von Collani, G., & Nestler, S. (2008). Automatic processing of pain: The change of implicit pain associations after psychotherapy. *Behaviour Research and Therapy, 46*, 701-714.

Hager, W., & Hasselhorn, M. (1994). *Handbuch deutschsprachiger Wortnormen* [Handbook of German word norms]. Göttingen, Germany: Hofgrefe.

Hermans, D., Vansteenwegen, D., Crombez, G., Baeyens, F., & Eelen, P. (2002). Expectancy-learning and evaluative learning in human classical conditioning: Affective priming as an indirect and unobtrusive measure of conditioned stimulus valence. *Behaviour Research and Therapy, 40*, 217-234.

Hetts, J. J., Sakuma, M., & Pelham, B. W. (1999). Two roads to positive regard: Implicit and explicit self-evaluation and culture. *Journal of Experimental Social Psychology, 35*, 512-559.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the Implicit Association Test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31*, 1369-1385.

Inquisit 2.0.60616 [Computer Software]. (2006). Seattle, WA: Millisecond Software LLC.

Jones, J. T., Pelham, B. W., & Mirenberg, M. C. (2002). Name letter preferences are not merely mere exposure: Implicit egotism as self-regulation. *Journal of Experimental Social Psychology, 38*, 170-177.

Jordan, C. H., Spencer, S. J., Zanna, M. P., Hoshino-Browne, E., & Correll, J. (2003). Secure and defensive high self-esteem. *Journal of Personality and Social Psychology, 85*, 969-978.

Jordan, C. H., Whitfield, M., & Zeigler-Hill, V. (2007). Intuition and the correspondence between implicit and explicit self-esteem. *Journal of Personality and Social Psychology, 93*, 1067-1079.

Karpinski, A. (2004). Measuring self-esteem using the Implicit Association Test: The role of the other. *Personality and Social Psychology Bulletin, 30*, 22-34.

Karpinski, A., & Steinman, R. B. (2006). The Single Category Implicit Association Test as a measure of implicit social cognition. *Journal of Personality and Social Psychology, 91*, 16-32.

Kim, D. Y. (2003). Voluntary controllability of the Implicit Association Test (IAT). *Social Psychology Quarterly, 66*, 83-96.

Kitayama, S., & Karasawa, M. (1997). Implicit self-esteem in Japan: Name letters and birthday numbers. *Personality and Social Psychology Bulletin, 23*, 736-742.

Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the Implicit Association Test: A diffusion-model analysis. *Journal of Personality and Social Psychology, 93*, 353-368.

Koole, S. L., Dijksterhuis, A., & van Knippenberg, A. (2001). What's in a name: Implicit self-esteem and the automatic self. *Journal of Personality and Social Psychology, 80*, 669-685.

Krause, S., Schmukle, S. C., Back, M. D., & Egloff, B. (2008, July). *Assessing implicit self-esteem with affective priming procedures: Improving psychometric properties by using pictures of participants as primes.* Paper presented at the XXIX International Congress of Psychology, Berlin.

LeBel, E. P., & Gawronski, B. (2009). How to find what's in a name: Scrutinizing the optimality of five scoring algorithms for the name-letter task. *European Journal of Personality, 23*, 85-106.

McFarlin, D. B., & Blascovich, J. (1981). Effects of self-esteem and performance feedback on future affective preferences and cognitive expectations. *Journal of Personality and Social Psychology, 40*, 521-531.

McGregor, I., & Jordan, C. H. (2007). The mask of zeal: Low implicit self-esteem, threat, and defensive extremism. *Self and Identity, 6*, 223-237.

Mogg, K., Bradley, B. P., Field, M., & De Houwer, J. (2003). Eye movements to smoking-related pictures in smokers: Relationship between attentional biases and implicit and explicit measures of stimulus valence. *Addiction, 98*, 825-836.

Moors, A, Spruyt, A., & De Houwer (2010). In search of a measure that qualifies as implicit: Recommendations based on a decompositional view of automaticity. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 19-37). New York, NY: Guilford Press.

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

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

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition, 19*, 625-664.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin, 31*, 166-180.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior* (pp. 265-292). New York, NY: Psychology Press.

Nuttin, J. M., Jr. (1985). Narcissism beyond Gestalt and awareness: The name letter effect. *European Journal of Social Psychology, 15*, 353-361.

Nuttin, J. M., Jr. (1987). Affective consequences of mere ownership: The name letter effect in twelve European languages. *European Journal of Social Psychology, 17*, 381-402.

Olson, M. A., Fazio, R. H., & Hermann, A. D. (2007). Reporting tendencies underlie discrepancies between implicit and explicit measures of self-esteem. *Psychological Science, 18*, 287-291.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*, 776-783.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*, 598-609.

Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, D. S. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*, 277-293.

Payne, B. K., & Gawronski, B. (2010). A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going? In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 1-15). New York; NY: Guilford Press.

Pelham, B. W., & Swann, W. B., Jr. (1989). From self-conceptions to self-worth: On the sources and structure of global self-esteem. *Journal of Personality and Social Psychology, 57*, 672-680.

Perugini, M., & Banse, R. (2007). Personality, implicit self-concept and automaticity. *European Journal of Personality, 21*, 257-261.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59-108.

Richetin, J., Perugini, M., Adjali, I., & Hurling, R. (2007). The moderator role of intuitive versus deliberative decision making for the predictive validity of implicit and explicit measures. *European Journal of Personality, 21*, 529-546.

Rinck, M., & Becker, E. S. (2007). Approach and avoidance in fear of spiders. *Journal of Behavior Therapy and Experimental Psychiatry, 38*, 105-120.

Rorschach, H. (1942). *Psychodiagnostik*. Bern, Switzerland: Bircher. (Original work published 1921)

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton: University Press.

Rothermund, K., Teige-Mocigemba, S., Gast, A., & Wentura, D. (2009). Minimizing the influence of recoding in the IAT: The IAT-RF. *Quarterly Journal of Experimental Psychology, 62*, 84-98.

Rothermund, K., & Wentura, D. (2010). It's brief but is it better? An evaluation of the Brief Implicit Association Test (BIAT). *Experimental Psychology, 57*, 233-237.

Rudolph, A., Schröder-Abé, M., Schütz, A., Gregg, A. P., & Sedikides, C. (2008). Through a glass, less darkly? Reassessing convergent and discriminant validity in measures of implicit self-esteem. *European Journal of Psychological Assessment, 24*, 273-281.

Schacter, D. (1985). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 501-518.

Schmukle, S. C., Back, M. D. & Egloff, B. (2008). Validity of the five-factor model for the implicit self-concept of personality. *European Journal of Psychological Assessment, 24*, 263-272.

Schmukle, S. C., & Egloff, B. (2004). Does the Implicit Association Test for assessing anxiety measure trait and state variance? *European Journal of Personality, 18*, 483-494.

Schmukle, S. C., & Egloff, B. (2005). A latent state-trait analysis of implicit and explicit personality measures. *European Journal of Psychological Assessment, 21*, 100-107.

Schmukle, S. C., & Egloff, B. (2006). Assessing anxiety with extrinsic Simon tasks. *Experimental Psychology, 53*, 149-160.

Schmukle, S. C., Hirschmüller, S., Back, M. D., & Egloff, B. (2007). Evaluation of an affect misattribution procedure for assessing self-esteem. In J. Henning, M. Reuter, C. Montag & P. Netter (Eds.), *IS-SID 07 Gießen* (pp. 147-148). Göttingen: Hofgrefe.

Schnabel, K., Banse, R., & Asendorpf, J. (2006). Employing automatic approach and avoidance tendencies for the assessment of implicit personality self-concept: The Implicit Association Procedure (IAP). *Experimental Psychology, 53*, 69-76.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*, 1-66.

Schütz, A., & Sellin, I. (2006). *Die Multidimensionale Selbstwertskala (MSWS)* [The multidimensional self-esteem scale]. Göttingen, Germany: Hofgrefe.

Sekaquaptewa, D., Espinoza, P., Thompson, M., Vargas, P., & von Hippel, W. (2003). Stereotypic explanatory bias: Implicit stereotyping as a predictor of discrimination. *Journal of Experimental Social Psychology, 39*, 75-82.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127-190.

Shrauger, J. S. (1975). Responses to evaluation as a function of initial self-perception. *Psychological Bulletin, 82*, 581-596.

Spalding, L. R., & Hardin, C. D. (1999). Unconscious unease and self-handicapping: Behavioral consequences of individual differences in implicit and explicit self-esteem. *Psychological Science, 10*, 535-539.

Spruyt, A., Hermans, D., De Houwer, J., Vandekerckhove, J., & Eelen, P. (2007). On the predictive validity of implicit attitude measures: Prediction of consumer choice behavior on the basis of affective prim-

ing in the picture-picture naming task. *Journal of Experimental Social Psychology, 43*, 599-610.

Sriram, N., & Greenwald, A. G. (2009). The Brief Implicit Association Test. *Experimental Psychology, 56*, 283-294.

Stacy, A. W., Ames, S. L., & Grenard, J. L. (2007). Word associations tests of associative memory and implicit processes: Theoretical and assessment issues. In R. W. Wiers & A. W. Stacy (Eds.), *The handbook of implicit cognition and addiction* (pp. 75-90). Thousand Oaks, CA: Sage Publishers.

Steffens, M. C. (2004). Is the Implicit Association Test immune to faking? *Experimental Psychology, 51*, 165-179.

Steffens, M. C., Kirschbaum, M., & Glados, P. (2008). Avoiding stimulus confounds in Implicit Association Tests by using the concepts as stimuli. *British Journal of Social Psychology, 47*, 217-243.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*, 220-247.

Teachman, B. A., Gregg, A. P., & Woody, S. R. (2001). Implicit associations for fear-relevant stimuli among individuals with snake and spider fears. *Journal of Abnormal Psychology, 110*, 226-235.

Teige-Mocigemba, S., Klauer, K. C., & Rothermund, K. (2008). Minimizing method-specific variance in the IAT: A single block IAT. *European Journal of Psychological Assessment, 24*, 237-245.

Teige, S., Schnabel, K, Banse, R, & Asendorpf, J. B. (2004). Assessment of multiple implicit self-concept dimensions using the extrinsic affective Simon task (EAST). *European Journal of Personality, 18*, 495-520.

von Collani, G., & Herzberg, P. Y. (2003). Eine revidierte Fassung der deutschsprachigen Skala zum Selbstwertgefühl von Rosenberg. *Zeitschrift für Differentielle und Diagnostische Psychologie, 24*, 3-7.

Voss, A., Leonhart, R., & Stahl, C. (2007). How to make your own response boxes: A step-by-step guide for the construction of reliable and inexpensive parallel-port response pads from computer mice. *Behavior Research Methods, 39*, 797-801.

Wagenmakers, E.-J., van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review, 14*, 3-22.

Wentura, D., Kulfanek, M., & Greve, W. (2005). Masked affective priming by name letters: Evidence for a correspondence of explicit and implicit self-esteem. *Journal of Experimental Social Psychology, 41*, 654-663.

Wiers, R. W., Van Woerden, N., Smulders, F. T. Y., & De Jong, P. J. (2002). Implicit and explicit alcohol-related cognitions in heavy and light drinkers. *Journal of Abnormal Psychology, 111*, 648-658.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53*, 300-314.

Wilson, T. D., 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.

Yamaguchi, S., Greenwald, A. G., Banaji, M. R., Murakami, F. Chen, D., Shiomura, K., Kobayashi, C., Cai, H., & Krendl, A. (2007). Apparent universality of positive implicit self-esteem. *Psychological Science, 18*, 498-500.

Zeigler-Hill, V. (2006). Discrepancies between implicit and explicit self-esteem: Implications for narcissism and self-esteem instability. *Journal of Personality, 74*, 119-143.

Appendix A – List of Word Stimuli Used in Study 1

Target- and attribute-stimuli for the IAT (BIAT), APT (RW-APT), and ID-EAST (German equivalents are in parenthesis)

IAT		APT		ID-EAST	
Target-stimuli		**Prime-stimuli**		**Target-stimuli**	
Me	**Other**	**Me**	**Other**	**Me**	**Other**
I (ich)	other (andere)			I (ich)	other (andere)
me (mir)	their (ihr)	picture of	pictures of	me (mir)	those (jene)
my (meine)	those (jene)	participant	two unknown	my (meine)	you (euch)
own (eigene)	you (euch)		persons	self (selbst)	your (eure)
self (selbst)	your (eure)				
Attribute-stimuli		**Target-adjectives**		**Attribute-stimuli**	
Positive	**Negative**	**Pleasant**	**Unpleasant**	**Positive**	**Negative**
active (aktiv)	brutal (brutal)	beautiful (schön)	cowardly (feige)	caring (sozial)	arrogant (arrogant)
cheerful (fröhlich)	dishonest (verlogen)	fair (fair)	envious (neidisch)	competent (fähig)	cruel (grausam)
evenhanded (gerecht)	inertial (träge)	healthy (gesund)	mean (gemein)	gentle (zärtlich)	labile (labil)
gifted (begabt)	vain (eitel)	honest (ehrlich)	vicious (bösartig)	nice (lieb)	nasty (fies)
human (human)	weary (lustlos)	sincere (herzlich)	ruthless (rabiat)		

Appendix B – List of German Anagrams Used in Study 2

German anagrams, their solutions, and item difficulty

Anagram	Solution	Item difficulty
EAFLP	Apfel	high
TWEES	Weste	medium
PAILR	April	high
OLVEG	Vogel	medium
KUREG	Gurke	high
URZEK	Kreuz	high
LMDEU	Mulde	high
AABUN	Anbau	high
THAPU	Haupt	medium
DRNEU	Runde	high
REEFN	Ferne	high
THIEZ	Hitze	medium
NIPZR	Prinz	high
BFREA	Farbe	high
FKATR	Kraft	high
UARFN	Anruf	high
EGBTO	Gebot	medium
SPIET	Piste	high
ALUEN	Laune	high
ESURT	Streu	high

Psychometric Properties of Indirect Assessment Procedures

Sascha Krause

Reaktionszeitbasierte Verfahren zur indirekten Erfassung des Selbstwerts erfüllen häufig nicht die psychometrischen Anforderungen, die an psychologische Tests gestellt werden. Insbesondere die Kennwerte der Reliabilität sind niedriger im Vergleich zu etablierten Selbstwertfragebogen. In der ersten Studie wird gezeigt, wie die interne Konsistenz und die Stabilität verschiedener indirekter Verfahren durch Optimierungen des Stimulusmaterials, der Darbietungsmethode und der statistischen Auswertung erhöht werden können. In der zweiten Studie wird gezeigt, dass eine entsprechend optimierte Version des affektiven Primings — das sogenannte Response-Window Priming — verschiedene selbstwertrelevante Kriterien unabhängig vom Selbstbericht vorhersagen kann. Die Ergebnisse beider Studien belegen die Nützlichkeit des Response-Window Primings für die zukünftige Selbstwertforschung.

ISBN 978-3-8405-0036-7 EUR 10,50
01050
9 783840 500367