INTRODUCTION

The concept of implicit self-esteem (ISE) was originally introduced by Greenwald and Banaji (1995) as an “introspectively unidentified (or inaccurately identified) effect of the self-attitude on the evaluation of self-associated and self-dissociated objects” (p. 11). The authors described a series of effects that indicate an automatic, unconscious positive evaluation of self-associated objects (objects, persons, and symbols). In doing so, they initially remained agnostic about the underlying processes and structures of the effects. Later on, ISE was described in terms of semantic network models as the totality of the direct and mediated associations between the concept of “self” with a valence node (Greenwald et al., 2002). For most individuals this association is positive. Regardless of the choice
of implicit self-positivity (Bosson et al., 2000; Yamaguchi et al., 2007). This finding soon inspired theorizing on the determinants and effects of an individual’s ISE and research attempting to validate the theoretical claims about ISE. The hope was to identify cognitive and affective processes determining self-confident behavior and cognition as well as psychological and physical well-being beyond what self-esteem measured in a self-report could explain. An early investigation into the measurement and correlates of ISE, however, revealed crucial problems in this line of research (Bosson et al., 2000). First, most supposed ISE measures used at this point did not display levels of reliability necessary to study individual differences in ISE. Second, correlations between supposed ISE measures were negligible, casting doubt on the assumption that they actually measure the same construct. Third, supposed measures of ISE showed small or even null correlations with directly measured (i.e., explicit) global self-esteem (ESE) and plausible correlates of self-esteem like interpretations of ambiguous statements, seeking of positive feedback, number of doctor visits, or academic ability.

These early findings caused researchers to focus on the two measures that stood out in terms of reliability: the Self-Esteem Implicit Association Test (SE-IAT; Greenwald & Farnham, 2000) and the Name Letter Task (NLT; Kitayama & Sarasawa, 1997; Koole et al., 2001). While other measures like the response window affective priming procedure were also developed, improved, and validated (Krause et al., 2011, 2012, 2016), most literature on ISE conventionally uses the NLT and the SE-IAT. In the SE-IAT, self-esteem is estimated based on performance differences between compatible and incompatible conditions in a double categorization task. In the NLT, individuals rate the attractiveness of letters and an ISE estimator is calculated as the difference in ratings between the initials of the subject and non-initial letters (LeBel & Gawronski, 2009). Despite the concentration on these two comparatively reliable measures and weeding out problematic measures, the validity of both the idea and the measures of ISE is as open to question today as they were 20 years ago (Buhrmester et al., 2011; Schimmack, 2019).

Granted, research on covariates of ISE has produced some findings which hint towards the validity of the conventional measures of ISE. These include, for example, the positive correlation of ISE-ESE discrepancies with anger suppression, physical and mental health, attributional style, and self-enhancement tendencies (Bosson et al., 2003; Schröder-Abé et al., 2007; Zeigler-Hill, 2011); the connection of ISE and self-confidence in social situations (Rudolph et al., 2010; Spalding & Hardin, 1999); the negative relation of depressive symptoms and ISE (e.g., Franck, de Raedt, & de Houwer, 2007; Franck, de Raedt, Dereu, et al., 2007; Risch et al., 2010; van Tuijl et al., 2020, for a meta-analysis, see Phillips et al., 2010), the prediction of reactions to negative events (Dijksterhuis, 2004; Haeffel et al., 2007), and the moderating role of ISE in the relation of positive and negative events and the drinking behavior in college students (DeHart et al., 2009). On the other hand, more sobering or counter-intuitive results have also been reported frequently (e.g., Bos et al., 2010; De Jong, 2002; Mota et al., 2020; Raedt et al., 2006; Schimmack & Diener, 2003; Wegener et al., 2015). In their comprehensive review, Buhrmester et al. (2011) point to the lack of robust correlations of supposed ISE measures with psychological well-being, depression, transient affect, physical health, emotional instability, or positive interpretation of ambiguous statements indicating a lack of criterion validity of these measures. They argue that measures of a concept entailing the term “self-esteem” should at least have some relation to ESE and its covariates. Correlations with ESE as well as self-esteem assessed by others, however, are usually small to negligible (Krzan & Suls, 2008; Pietschnig et al., 2018). Also, the SE-IAT and the NLT were shown to correlate only at .08 on average, with correlations being small and sometimes even negative (Bosson et al., 2000; Krause et al., 2011; Rudolph et al., 2008; for a review, see Buhrmester et al., 2011). What is more, in a meta-analysis Greenwald et al. (2009) demonstrated that self-esteem is the only domain in which the IAT cannot significantly predict behavior. Research on covariates of supposed ISE measures thus yielded evidence for the validity of these measures that is mixed at best and questionable at worst (Falk & Heine, 2015; Falk et al., 2015). So much so, that the review by Buhrmester et al. (2011) even called for an abandonment of the NLT and the SE-IAT altogether.

In our view, it is still too early to draw final conclusions from the existing findings, since the options of optimizing research on the validity of these ISE measures are not exhausted yet. In the following, we will point to two crucial obstacles in this line of research that are underappreciated in the current literature: the situational variability of ISE measures and construct irrelevant variance in ISE measures. Addressing these obstacles might be fruitful in validating supposed ISE measures and the concept of ISE itself. The present study will attempt to do so.

### 1.1 Situational variability in ISE measures

Authors investigating correlates of ISE often more or less explicitly regard ISE as a stable personality trait. Meteorologically speaking, ISE is sometimes considered to be the “self-evaluative climate”, whereas ESE is
Much like other implicit associations, it is viewed as an overlearned self-attitude, the development of which is attributed to the quality of early interpersonal interactions and which is automatically activated when encountering self-associated stimuli (e.g., DeHart et al., 2006; Koole et al., 2001). There is, in fact, some evidence for this view. As DeHart et al. (2006) demonstrated, the parenting behavior of parents proved to be predictive for NLT scores measured in teenagers. Teenagers of caring and not overprotecting parents had higher NLT scores than their peers. Also, twin studies demonstrated a heritability of 22% and 35% of SE-IAT and NLT scores respectively (Cai & Luo, 2017; Stieger et al., 2017).

Despite these findings attesting to the existence of a trait component in ISE measures, their situational malleability has also been demonstrated repeatedly: SE-IAT and NLT scores have been shown to be malleable by relatively simple short-term interventions. These interventions include classical conditioning (Baccus et al., 2004; Espinosa et al., 2018), written emotional disclosure (O’Connor et al., 2011), subliminal evaluative conditioning (Dijksterhuis, 2004; Grumm et al., 2009; Jraidi & Frasson, 2010), social exclusion (Stewart et al., 2017), and unilateral hand contractions (Quirin et al., 2018). Importantly, most of these interventions are single session manipulations that cannot be expected to permanently change a stable personality trait.

Studies explicitly modeling stable trait and occasion-specific state components of repeated ISE measurements were recently conducted. They found that about half of the reliable variance in SE-IAT scores can be attributed to a stable trait and half can be attributed to situational factors (Dentale et al., 2019). In two studies on the NLT, this ratio was shown to lie between 70/30 and 80/20 (Perinelli et al., 2018). The SE-IAT thereby showed higher dependency on situational factors than measures of ESE, while the NLT was close to ESE in this regard (Wang et al., 2018), for which the estimated proportion of reliable variance attributable to a stable trait was found to be .77 and .82 respectively.

In sum, the evidence for occasion-specific variance in supposed ISE measures reveals the first obstacle for correlational research on ISE: The stable trait variance aimed at by most research is only part of the variance captured by supposed ISE measures. Another part of the variance can be attributed to measurement error and situational factors. Accordingly, just as predicting an individual’s degree of neuroticism with a single measurement of their mood cannot yield a satisfying estimate of the relation, predicting neuroticism with a single measurement of ISE cannot be expected to produce robust results. Predicting a stable trait with an occasion-specific state simply limits the size of correlations that can be expected. Interestingly, when modeling the latent trait components of ISE and ESE, Dentale et al. (2019) found a moderate and significant correlation between ISE (as measured with the SE-IAT) and ESE. This finding deviates from what is usually found in single session measurements thereby providing some promise about the advantages of adopting multiple measurement designs and modeling of latent variables.

1.2 Construct irrelevant variance

A second problem in the investigation of ISE is that implicit measures are not “process-pure”. That is, they capture not only construct relevant variance but also variance produced by influences other than the to-be-measured construct. Recent research has demonstrated this to be true for the IAT. The NLT, however, due to its format gives fewer opportunities to investigate construct irrelevant sources of variance aside from response biases that can be accounted for by using an appropriate algorithm (LeBel & Gawronski, 2009). For the IAT, non-evaluative as well as evaluative confounding processes have been identified (Teige-Mocigemba et al., 2010). These processes include recoding of the dual categorization task to a simple categorization task in the compatible block (De Houwer et al., 2005; Rothermund et al., 2005, 2009), task switching between task and attribute categorization (Ito et al., 2015; Klauer & Mierke, 2005; Klauer et al., 2010; Mierke & Klauer, 2001, 2003), inhibition of response biases (Conrey et al., 2005), general processing speed (Blanton et al., 2006; McFarland & Crouch, 2002), and differences regarding speed-accuracy tradeoffs in the compatible and incompatible blocks (Klauer et al., 2007). All of these influences may impact the outcome of an IAT to varying degrees depending on the test-taking strategy and the cognitive capabilities of the individual. Besides these non-evaluative processes, the automatic evaluative associations regarding the category “other” that is used as a contrast category in the SE-IAT may also influence the outcome, although these associations are unrelated to the evaluation of the self. While there has been some debate about the right choice of contrast category in the SE-IAT (Karpinski, 2004; Pinter & Greenwald, 2005), there can be no doubt that there is variability in the way individuals evaluate the contrast category. Since the D2 score of the IAT, which is the default effect measure, only reflects differences in attitudes towards the two categories (e.g., “me” and “other”, Blanton et al., 2006), the other-association might mask a link between the criterion and the self-evaluation. The well-known cognitive triad of depression (Beck & Rush, 1979; Beckham et al., 1986), for example, entails automated negative beliefs about the self as well as about others (“the world”). Depressed individuals might, therefore, have the same relative preference of self
over others compared to healthy individuals, but would differ from them with respect to the overall negativity/positivity of evaluative associations towards both target concepts. In sum, the SE-IAT effect must be interpreted as an amalgam of different sources and processes only one of which is the implicit self-evaluation.

1.3 | Identifying construct-relevant sources of trait implicit self-esteem

Combining the two perspectives, a twofold problem has to be solved with regard to the variance in supposed ISE measures: (a) Trans-situationally stable variance has to be distinguished from state-dependent sources of variance, and within each of these components, (b) construct-relevant variance has to be separated from construct-irrelevant sources of variance. Given the complexity of this task, it is apparent that a stable implicit self-evaluation might be the proverbial needle in the haystack of the SE-IAT, with only stable and construct-relevant variance being a candidate for predicting relevant trait-like outcomes.

The present study aims to explicitly address this problem to investigate the incremental validity of the stable trait component of supposed ISE measures in predicting plausible criteria over and above trait ESE. This is done by modeling the respective latent trait components of multiple measurements of ISE using latent state-trait theory (Steyer & Schmitt, 1990), thereby addressing the comparatively low reliability and the occasion-specificity of ISE measurements. What is more, construct-relevant and irrelevant processes in the SE-IAT are modeled using the multinomial ReAL model (Meissner & Rothermund, 2013) thereby extracting a meso-positive association parameter that is not confounded with recoding and the evaluation of the “other” category. The correlation of this parameter with ESE and the NLT score, as well as its capability to predict life satisfaction, neuroticism, interpretation of ambiguous scenarios, feedback-seeking behavior, remembered parenting behavior, and depressive symptoms over and above ESE are investigated.

In the following, we give a brief overview of the two approaches that we used to separate construct-relevant and construct-irrelevant influences in the SE-IAT (the ReAL model), and to distinguish stable from situation-specific variance (latent trait-state theory).

1.4 | The ReAL model

The ReAL model (Meissner & Rothermund, 2013) is a multinomial model that maps multiple processes in the IAT onto distinct model parameters: The Re parameter is an indicator for the role of recoding in the IAT (i.e., redefining and subsuming the target and attribute categories under a common description in the compatible block; Gast & Rothermund, 2010; Rothermund et al., 2009), whereas two separate A parameters reflect evaluative associations towards the two target concepts, thus representing the strength and direction of attitudinal preferences. Finally, the parameter L (i.e., the label-based identification of the response) reflects the probability of responding that is based on the instructed rules of the task. Former studies gathered ample evidence for the internal and external validity of these model parameters (Meissner & Rothermund, 2013). Furthermore, it revealed that the A parameters provide more valid conclusions about evaluative associations than global IAT scores (Jin, 2016; Koranyi & Meissner, 2015; Meissner & Rothermund, 2013, 2015a, 2015b) and that they can outperform the global IAT score in terms of predictive validity (Meissner & Rothermund, 2013). Moreover, since the ReAL model provides separate A parameters for each of the two target concepts, this analysis strategy goes beyond the relative nature of the IAT score: correlates can be determined separately for evaluative associations of both target concepts. Other models were proposed to separate construct-irrelevant from construct-relevant processes in the IAT. These include the quad model (Conrey et al., 2005) and the diffusion model (Klauser et al., 2007), which will also be applied to the IAT data we collected. However, since the ReAL model is the only model specific to the IAT, which captures the most probable contaminants in an SE-IAT (recoding and the other-association), results concerning the ReAL parameter estimates will be prioritized and findings for the other models will be added later on.

1.5 | Latent state-trait theory

Latent state-trait theory (Steyer & Schmitt, 1990) conceptualizes multiple manifest measurements of a random variable \( Y_{it} \) as a sum of a latent true score \( \tau_{it} \) and measurement error \( e_{it} \). Indices \( i \) and \( t \) denote the ID of the measurement (i.e., the item number or the test-half) and the occasion respectively. \( \tau_{it} \) can, in turn, be partitioned into a latent trait, that is stable across all measurements (\( \theta \)), occasion-specific latent state residuals (\( \zeta_{it} \)), and a stable method factor (\( M \)), that is specific to the measurement \( i \). The sum of \( \theta \) and \( \zeta_{it} \) yields the latent state \( \eta_{it} \). Using structural equation modeling, LST theory can be applied to multiple parallel measurements of the same construct. While the details, of course, depend on the exact specification of the model, the general aims of LST modeling are (a) to investigate the proportions of variance in \( Y_{it} \) that are attributable to \( e_{it}, \theta, \zeta_{it} \) and \( M \) respectively and (b) to investigate the relations (i.e., correlations, regressive dependencies) of the latent variables with other latent and manifest variables. In
the present study, we will use LST models to investigate the proportion of variance of ISE estimates that can be attributed to a stable trait (θ) and the relation of this latent trait with criterion measures. The exact specifications of the applied models are detailed below.

1.6 Outline of the study

In four equally spaced measurement occasions across six weeks, we administered the SE-IAT and the NLT and ESE measures. In addition, we measured relevant personality and mental health outcomes. The measures were administered via the internet to capture measurements in a real-life environment which is less prone to artificially increase stability estimates due to the high similarity between measurement situations (e.g., same building, same laboratory, and same interactions with the experimenters). The selection of criterion measures was informed by the criteria discussed by Buhrmester et al. (2011) and was intended to cover a wide range of self-esteem correlates. We measured life satisfaction, neuroticism, interpretation bias, feedback-seeking behavior, and depressive symptoms as well as experienced parenting behavior. Using LST and process modeling we extracted a latent trait component for all constructs. We expected to find a positive correlation between the different supposed ISE measures on the latent trait level. What is more, we expected a positive incremental validity of a latent trait ISE in predicting life satisfaction, positive interpretation bias, and caring parenting behavior over and above ESE and a negative incremental validity in predicting neuroticism, depressive symptoms, and overprotective parenting behavior.

2 METHODS

2.1 Participants and procedure

The procedure of this study was preregistered (see https://osf.io/etv3g) and approved by the Ethical Commission of the Faculty of Social and Behavioral Sciences of the University of Jena (reference number FSV 19/068). 360 participants were recruited via Prolific (www.prolific.co) to take part in four measurement sessions spaced out across six weeks with an average interval of 14. The initial sample size of 360 was decided on to surpass a sample size of 250 after dropout and data exclusion, which Schönbrodt and Perugini (2013) found is a critical sample size for the stabilization of correlations. Participants were informed that they were taking part in a study on the development of emotions across time. Only participants with more than ten previous submissions on the platform and a prolific score of 95 or higher were admitted to this study. Participants were financially compensated for their effort. 302 participants (185 female, 115 male, 2 other) finished all four sessions and made up the final sample. The mean age of participants was 33.04 years, SD = 12.21. In the four sessions supposed measures of ISE, measures of ESE, and criterion measures were administered according to the plan shown in Table 1. Measures of ISE were always completed first to prevent order effects and contaminations of ISE measures with ESE (Bosson et al., 2000). Criterion measures were mostly personality measures that are assumed to be rather stable and therefore only administered on one or two measurement occasions. All measures were administered by virtue of a jsPsych (Leeuw, 2015) routine hosted on a JATOS server (Lange et al., 2015). Participants were fully debriefed at the end of the last session.

We ran a simulation study based on the results of the most complex model including the least reliable measures to determine the lower bound of the size of path coefficients (incremental validity) which grant sufficient power to be detected with this sample size. We found that with N = 302 a power larger than .80 was given for standardized path coefficients that are larger than .15 when testing with one-tailed tests. The higher the reliability of the involved measures the lower this critical value.

2.2 Measures

2.2.1 Implicit self-esteem

**Self-Esteem Implicit Association Test**

We adopted an SE-IAT procedure (Greenwald & Farnham, 2000) that was modified to provide the error
frequencies and the number of trials necessary for the reliable estimation of the model parameters and diffusion models (see Meissner & Rothermund, 2013). Participants were asked to categorize the following words into the target categories (“Me” and “Other”) or the attribute categories (“Positive” and “Negative”) respectively. The following stimuli were used as exemplars for the four categories: Me (I, self, me, my, mine); other (other, they, those, their, others); positive (honest, competent, strong, clever, beautiful); negative (dishonest, incompetent, weak, stupid, ugly). The stimuli were presented in a quasi-randomized order preventing contingencies between trial types and ensuring equal frequencies for task-switch (attribute-target, target-attribute sequences) and task-repetition (attribute-attribute, target-target sequences) trials as well as trials of the four categories. The procedure consisted of five practice blocks (16 trials attribute practice, 16 trials target practice, 20 trials combined practice compatible, 16 trials of target practice reversed, 20 trials of combined practice incompatible) followed by six pairs of compatible and incompatible test blocks of 33 trials each (totaling $12 \times 33 = 396$ trials).

An adaptive response deadline was established in the combined practice to guarantee a sufficient number of erroneous responses (Meissner & Rothermund, 2013). A red frame appeared around the stimulus after a deadline. Participants were instructed to respond before the frame appeared and to accept potential errors in order to respond within the deadline. The deadline and instructions were adaptive to the performance of the participants, aiming to yield approximately 25% errors for each participant (for details, see Meissner & Rothermund, 2013). Using the error data in this task, ReAL and quad model parameters were estimated. A diffusion model was estimated using reaction time and error data. As a reference parameter, an IAT effect was computed using the $D_2$ algorithm (Greenwald et al., 2003). For each session, ISE parameters were estimated for odd and even test block pairs separately to produce two estimators per session.

**Name Letter Task**
The NLT (Koole et al., 2001) was used as a complementary measure of ISE. Participants were asked to rate the twenty-six upper-case letters of the English alphabet on a 9-point scale from “not at all beautiful” to “extremely beautiful”. They were instructed to respond without overthinking their rating but listening to their gut feeling. The ratings were scored using the I-algorithm (LeBel & Gawronski, 2009), which was found to have optimal psychometric properties. Both initials were scored independently per session. Participants were asked to report their initials at the end of the first session and again after being debriefed about the calculation of NLT scores at the end of the last session. For us to be able to exclude participants who intentionally did not report their true initials, the participants were at the same moment explicitly asked if they reported their true initials.

### 2.2.2 Explicit self-esteem

**State Self-Esteem Scale**
The State Self-Esteem Scale (SSES; Heatherton & Polivy, 1991) was administered in each session to capture ESE at the same level of temporal resolution as ISE. This scale was proposed to measure fluctuations in ESE that other measures of global ESE are insensitive to. Participants rated the accuracy of 20 items about their confidence in their performance, their appearance, and their social esteem in the present moment on a 5-point Likert scale. Test-halves were produced by splitting odd- and even-numbered items. Mean ratings for both test-halves were used as manifest indicators of ESE.

**Rosenberg Self-Esteem Scale**
The Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965) was used in the first and the last session. This measure is most often used as a measure of global and stable self-evaluations. It consists of 10 items about the general attitude individuals have towards themselves, which are rated on a 5-point Likert scale from "strongly disagree" to "strongly agree". Test-halves were again produced by splitting odd- and even-numbered items. Mean ratings for both test-halves were used as manifest indicators of ESE.

### 2.2.3 Criterion measures

**Satisfaction with Life Scale**
The Satisfaction with Life Scale (SWLS; Diener et al., 1985) was issued in the first and last session to measure overall life satisfaction. Participants rated five items on a 7-point scale from “strongly disagree” to “strongly agree”. Test-halves were produced by splitting odd- and even-numbered items. Mean ratings for both test-halves were used as manifest indicators of global life satisfaction. The unequal distribution of items into the test-halves was taken into account in the modeling.

**Neuroticism**
The Emotional Stability scale of the Big Five Factor Markers (Goldberg, 1992) was taken from the IPIP (Goldberg et al., 2006) and inversely scored as an indicator of neuroticism (NEU). Participants rated ten items about their emotional stability on a 5-point scale from “very
inaccurate” to “very accurate”. Test-halves were produced by splitting odd- and even-numbered items. Mean ratings for both test-halves were used as manifest indicators of neuroticism.

**Ambiguous Scenario Test**
Two parallel versions of an Ambiguous Scenario Test (AST; Rohrbacher & Reinecke, 2014) were administered in the first and last session respectively to assess positive interpretation bias. Participants were presented with 15 ambiguous descriptions of situations from the domains “self”, “experience” and “past” and were asked to imagine these scenarios as vividly as possible. They were then instructed to rate how the imagined scenario felt on an 11-point scale ranging from “extremely unpleasant” to “extremely pleasant”. Test-halves were produced by splitting odd and even-numbered items. Mean ratings for both test-halves were used as manifest indicators of interpretation bias. The unequal distribution of items into the test-halves was taken into account in the modeling.

**Feedback Seeking Questionnaire**
The Feedback Seeking Questionnaire (FSQ) was adopted from Swann et al. (1992) and completed in the first and last session to assess preferences for positive or negative feedback. Participants were instructed to imagine that the person who knows them best is being interviewed about them. They were presented with five lists of six questions each pertaining to a certain area (Social, Intellectual, Artistic, Physical Appearance, and Sport). They were instructed to choose two questions from each area they would most like the person who knows them best to answer about themselves. One half of each set of questions was formulated to warrant negative feedback (e.g., “What are some signs you have seen that he/she is below average in overall intellectual ability?”) and the other half was formulated to warrant positive feedback (e.g., “In terms of social competence, what is his/her best asset?”). Selections of positive feedback were coded as 1, selections of negative feedback were coded as −1, resulting in a mean score representing a preference for positive or negative feedback indicated by its sign. Test-halves were produced by splitting first and second choices per area. Mean ratings for both test-halves were used as manifest indicators of feedback-seeking behavior.

**Becks Depression Inventory II**
The second version of Becks Depression Inventory (BDI; Beck et al., 1996) was issued in the first and last session. Participants were asked to rate the severity of 21 depressive symptoms on a 4-point scale that was individually labeled for each symptom. Test-halves were produced by splitting odd and even-numbered items. Mean ratings for both test-halves were used as manifest indicators of the severity of depressive symptoms. The unequal distribution of items into the test-halves was taken into account in the modeling.

**Parental Bonding Instrument**
The Parental Bonding Instrument (PBI; Parker et al., 1979) was completed in the third session only. Since it is a retrospective measure of parenting behavior multiple measurement was dismissed as redundant in this investigation. Additionally, sessions one and four were supposed to be as parallel as possible, so the measurement was restricted to the third session. The instrument consists of 20 items belonging to the scales “care” and “overprotection”. Participants rated these items in regard to how like or unlike the described behaviors are to the behavior of their mother and father separately. Ratings were given on a 5-point scale from “very unlike” to “very like”. Ratings for the scales were averaged across mother and father resulting in a “care” (PBI-C) score and an “overprotection” (PBI-O) score per participant.

## 3 RESULTS

The data supporting the following results as well as the syntax for all analyses can be accessed at https://osf.io/ahve (Jusepeitis & Rothermund, 2022a, 2022b).

### 3.1 Data preparation

NLT data for participants who did not report their initials (n = 17) were excluded from the analyses. Furthermore, NLT data of units (a participant at a certain session) who evaluated all letters equally was excluded. This resulted in sample sizes of n₁ = 280, n₂ = 277, n₃ = 265 and n₄ = 268 for NLT scores. IAT data were excluded for 20 units which showed response latencies below 200ms in more than 20% of trials. The exclusion of these data points did not change the outcomes of the presented analyses. Missing data were dealt with by full information maximum likelihood (FIML) estimation since Little’s missing completely at random test indicated that data were missing completely at random for most SE-IAT attitude parameters, all χ²(20) ≤ 15.62, all p ≥ .74, as well as for NLT scores, χ²(19) = 11.55, p = .90. Only D₂-scores were not missing at random, χ²(20) ≤ 41.02, all p < .01. Hence, we compared the FIML estimation for D₂-scores with an estimation based on only the complete data but found no difference and only the FIML estimation will be reported.
3.2 | Multinomial process tree modeling

First, we attempted to estimate ReAL and Quad model parameters by employing the hierarchical Bayesian approach proposed by Klauer (2010) with the TreeBUGS R package (Heck et al., 2018) with session and test-half introduced as within-subject factors. However, the resulting model complexity prevented the MCMCs from converging, all \( R > 1.05 \) at 800,000 iterations. Because of this, a non-hierarchical estimation approach was adopted that estimates parameters per session and test-half independently for each person. We employed a Bayesian analysis with flat priors (beta distributions with parameters \( \alpha = 1 \) and \( \beta = 1 \)) for each parameter implemented again with the TreeBUGS R package (Heck et al., 2018). The estimations finished normally.

3.2.1 | ReAL model

For each participant, attitude parameters \( A_1 \) (me-positive) and \( A_2 \) (other-positive), recoding parameter Re, and parameters \( L_1 \) through \( L_4 \) (label-based identification of target and attribute stimuli) as well as six technical parameters (Meissner & Rothermund, 2013) were estimated per session and test-half. They were probit-transformed for further analyses with linear models. Since the main goal was to extract an ISE parameter, we will primarily report results concerning the \( A_1 \) parameter in the following. Exploratory analyses involving Re and \( A_2 \) will be amended later. The model fit for each unit was investigated by calculating T1 posterior predictive p-values (PPP), which quantify deviations of the predicted frequencies from the observed frequencies. Median T1 PPP was .50, with 34 units (1%) showing values smaller than .05 (however, for a critique of the usefulness of PPP values see Klauer, 2010). Since the exclusion of these units did not change the results of our analyses, they were not discarded.

3.2.2 | Quad model

For each participant, each of the five parameters of the Quad model (Conrey et al., 2005) was estimated per session and test-half. We will only report results concerning the AC1 and AC2 (activation of the self-positive and other-negative associations) parameters. Model fit was again evaluated by PPP values. Median T1 PPP was .41, with 206 (9%) units showing values smaller than .05. Since the Quad model parameters were not of prime interest in this study, these units were not excluded. In general, the fit of the Quad model was worse than that of the ReAL model.

3.3 | Diffusion modeling

Diffusion model analyses were conducted analogously to Klauer et al. (2007) using fast-dm (Voss & Voss, 2007). Parameter estimates were optimized using the Kolmogorov-Smirnov statistic. Parameter \( z_0 \) (the starting point of the diffusion process) was fixed to .5. \( p \) (guessing) was fixed to 0. Following the recommendations of Röhner and Ewers (2016) on diffusion modeling of IAT data, all parameters of intertrial variability were fixed to 0. Hence, only the drift rate (\( \nu \)), response caution (\( \alpha \)), and non-decision time (\( t_d \)) were freely estimated for compatible and incompatible test blocks respectively. Only results for the \( \nu \) parameter will be reported since it was found to uniquely reflect construct-related variance (Klauer et al., 2007; Rebar et al., 2015). The fit of the model was assessed by the p-value of the Kolmogorov-Smirnov statistic. Median \( p \) was .56 and only four units exhibited a misfit of the model and were not excluded (for a comment on the interpretability of these p-values, see Voss et al., 2013). An effect parameter IAT\(_v\) was calculated by subtracting \( v \) estimates for the incompatible condition from the estimates in the compatible condition. Thus, a positive IAT\(_v\) reflects a higher drift rate in the compatible condition compared to the incompatible condition.

3.4 | Latent state-trait modeling

Tables of all results of the LST analyses including model fit and parameter estimates can be found in the Supporting Information. After detailing the model selection process we will briefly summarize the results and only explicate important findings to maintain a clear report.

3.4.1 | Estimating variance proportions

Model specification

In a first step, mono-construct (i.e., one latent trait) LST models were estimated for all measures to separate stable, occasion-specific, method-specific (i.e., test-half-specific), and error variance components of all measures. Figure 1 shows a schematic depiction of the model. For technical details of model specification and selection see Supporting Information—Appendices A1, A2, and A3. The models were applied to the data using lavaan (Rosseel, 2012). Since the distribution of self-esteem parameters in 69% of all sessions and test-halves deviated from normality as indicated by Shapiro-Wilk tests, a robust maximum likelihood estimator was chosen and the according robust fit indices corrected with the Yuan-Bentler correction (Yuan & Bentler, 2000) will be reported.
Model fit evaluated by $\chi^2$ and RMSEA statistics was ranging from acceptable to good for most measures (see Supporting Information—Appendices A1, A2, and A3 for all details concerning mono-construct LST models). The only exceptions were the AST, $\chi^2(8) = 33.17$, $p < .001$; RMSEA = .10, $p < .01$, the SWLS, $\chi^2(5) = 23.48$, $p < .001$; RMSEA = .11, $p = .02$ and the FSQ, $\chi^2(6) = 96.08$, $p < .001$; RMSEA = .22, $p < .001$. The modeling of the latter also resulted in significantly negative reliability estimates. Therefore, the FSQ was at this point excluded from all further analyses.

Variance proportions and mean trait ISE
Estimates of reliability, consistency, occasion- and method-specificity for all measures can be found in the Supporting Information—Appendices A2 and A3. Technical explanations of these quotients are given in the Supporting Information—Appendix A1. In general, reliability (i.e., the proportion of variance not attributable to measurement error) was highest for ESE and criterion measures (NEU, SWLS, RSES, SSES, and BDI) with estimates across all occasions exceeding .94. Only the AST displayed lower reliability with .82. For the NLT, reliability estimates ranged from .69 to .77 across occasions, for the $D_2$ score from .66 to .74. Reliability was lowest for the $A_1$ parameter, .48 to .54. A similar pattern emerged for the consistency of these measures. The consistency (i.e., the proportion of variance that is attributed to the stable trait component) of the explicit questionnaire measures was highest, ≥.78, moderate for the AST, .71; NLT, .44 to .64, and the $D_2$ score, .42 to .54, and lowest for the $A_1$ parameter, .35 to .40. Concerning the occasion-specificity (i.e., the proportion of variance that is attributed to latent state residuals), no clear pattern emerged. However, in proportion to their reliability, the IAT measures showed the highest occasion-specificity (on average 27% for $A_1$, 29% for $D_2$). The NLT exhibited on average 22% of reliable occasion-specific variance. For the questionnaire measures, this quotient overall was smaller, with the higher value being registered for the SSES (16%) and the lower for the RSES (9%). Interestingly, the NLT scores exhibited comparatively large test-half-specificity (i.e., the proportion of reliable variance that is not shared by the two test-halves) with 20% to 27% of the variance of scores per initial being attributable to $M$. This proportion did not exceed 5% for any other measure.

$Z$-tests indicated that the estimated expectation values of $\theta$ for the ISE parameters were all significantly greater than 0, all $p < .001$, showing the self-positivity bias on a latent trait level (see Supporting Information—Appendix A1).

3.4.2 Latent trait correlations

Model specification and selection
In a second step, dual-construct models were used to estimate latent trait correlations ($r_{\theta}$) between ISE and ESE as well as criterion measures. To this end, the mono-construct models were combined pairwise. Figure 2 depicts a general dual-construct LST model. For the sake of parsimony, the restrictions regarding the equality of error and state residual variances from the
mono-construct models were retained for each measure respectively. Correlations between method factors were set to zero. Latent state residual correlations \( r_\zeta \) were set to zero when allowing them did not significantly improve model fit, as tested with tests of \( \chi^2 \) differences (see Supporting Information—Appendix B). Since the PBI was administered at only one session, no latent trait for PBI-C and PBI-O was modeled. Instead, these scores were entered into the model as single manifest covariates of the respective other latent trait.

**Model fit**

Model fit was good for most models involving \( A_1 \) and the NLT with only one significant \( \chi^2 \) test of model fit per ISE parameter and no RMSEA significantly larger than .05, all \( p \geq .96 \). For all models involving \( D_2 \), however, \( \chi^2 \) tests indicated model misfit, all \( p \leq .01 \). On the other hand, for these models as well, no RMSEA was significantly larger than .05, all \( p \geq .16 \) (see Supporting Information—Appendix B).

**Parameter estimation**

As preregistered, latent trait correlations of ISE parameters and ESE measures as well as criterion measures were tested with one-tailed tests according to the direction of our hypotheses. They were overall small and only partially statistically significant (see Supporting Information—Appendix B). \( A_1 \) was significantly correlated with AST, \( r_\theta = .16, CI_{95} = [.04, .29], p = .01 \), and with PBI-C, \( r_\theta = .19, CI_{95} = [.09, .29], p < .01 \). The direction of most other
correlations was in line with our expectations, but they were statistically insignificant, all \( p \geq .05 \). The NLT was correlated with AST, \( r_\theta = .18, CI_{95} = [.06, .29], p < .01 \), with PBI-C, \( r_\theta = .11, CI_{95} = [.01, .21], p = .04 \), with RSES, \( r_\theta = .21, CI_{95} = [.11, .31], p < .001 \), with SSSE, \( r_\theta = .23, CI_{95} = [.13, .33], p < .001 \), and with SWLS, \( r_\theta = .24, CI_{95} = [.13, .34], p < .001 \). The direction of all other correlations was in line with our expectation but they were statistically insignificant, all \( p \geq .07 \). The D2 score was found to be correlated with BDI, \( r_\theta = -.13, CI_{95} = [-.33, -.02], p = .03 \), with NEU, \( r_\theta = -.11, CI_{95} = [-.33, -.01], p = .04 \), with PBI-O, \( r_\theta = -.22, CI_{95} = [-.5, -.12], p < .001 \), with RSES, \( r_\theta = .13, CI_{95} = [.02, .24], p = .03 \), and with SSSE, \( r_\theta = .13, CI_{95} = [.02, .24], p = .02 \). Most other correlations had the expected sign but were statistically insignificant, all \( p \geq .10 \).

Between the different ISE indicators there was only one significant correlation between A1 and D2, which contrary to our expectations was negative, \( r_\theta = -.20, CI_{95} = [-.37, -.03], p = .02 \), both other \( p \geq .58 \). For comparison, latent trait correlations between and within the criteria and ESE measures were mostly high and always statistically significant, \( \min |r_\theta| = .17 \), all \( p < .01 \), except for correlation of AST and PBI-O, \( r_\theta = -.09, CI_{95} = [-.21, .04], p = .17 \).

### 3.4.3 Incremental validity of ISE measures

#### Model specification and selection

In the last step, triple-construct models were used to test the incremental validity of ISE parameter traits in predicting criterion measure traits over and above ESE traits. Triple-construct LST models were created by combining mono-construct models in triplets, the first component being an ISE measure, the second being an ESE measure and the third being a criterion measure. Figure 3 depicts such a triple-construct LST model.

Since the latent state self-esteem (SSES) and Rosenberg self-esteem scale traits were highly correlated, \( r_\theta = .89, CI_{95} = [.86, .93], p < .001 \), and the SSSES trait estimation was based on four occasions instead of only two, thereby providing higher precision, we decided to implement the SSSES trait as the primary trait ESE parameter. As reported above, correlations of state residuals of ISE measures and SSSES did not improve model fit. Accordingly, all these correlations were fixed to zero and only latent ISE and ESE traits were allowed to correlate. However, as pointed out above, allowing for ESE-criterion correlations of latent state residuals significantly improved model fit. These correlations were therefore estimated freely but fixed to be equal across occasions. In a multiple linear regression, the criterions' latent trait was then regressed onto the latent ISE and ESE traits, with coefficients \( \alpha \) denoting the standardized slope (i.e., the standardized path coefficient) for the ISE trait and \( \beta \) denoting the standardized slope for the ESE trait (see Figure 3). In such a model \( \alpha \) represents the incremental validity of ISE in predicting the criterion over and above ESE. The PBI scales were again introduced as manifest criteria instead of being modeled like the other criterion measures.

#### Model fit

Tests of \( \chi^2 \) indicated model misfit for most models. The RMSEA, however, was acceptable for all models, max RMSEA = .05, min \( p = .36 \) (see Supporting Information—Appendix C for all details concerning triple-construct LST models).

#### Parameter estimation

As preregistered, standardized latent regression coefficients were tested with one-tailed tests against zero, according to the direction of our hypotheses. Standardized \( \beta \) coefficients were showing mostly strong and statistically significant relations between trait ESE and all criterion measures, \( \min |\beta| = .19 \), all \( p < .01 \). Standardized \( \alpha \) coefficients, on the other hand, were mostly small and insignificant. For A1, three out of six alphas showed the expected sign, and only the incremental prediction of PBI-C was significant, \( \alpha = .15, CI_{95} = [.05, .25], p < .001 \), all other \( p \geq .11 \). For the NLT, four out of six alphas showed the expected sign and a significant \( \alpha \) emerged in predicting SWLS, \( \alpha = .10, CI_{95} = [.00, .20], p = .04 \), all other \( p \geq .23 \). For D2, five out of six alphas showed the expected sign and a significant \( \alpha \) was found in the prediction of PBI-O, \( \alpha = -.20, CI_{95} = [-.5, -.09], p < .01 \), all other \( p \geq .28 \).

### 3.4.4 Investigation of other process modeling approaches

The diffusion model and Quad model ISE parameters IAT\( v \) and AC\textsubscript{1} were investigated analogously to the primary ISE parameters. Mono-construct models fit the data acceptably (see Supporting Information—Appendix A2). The estimated expectation values of \( \theta \) for both parameters were significantly larger than zero, both \( p < .001 \). The reliability estimates of AC\textsubscript{1} ranged from .56 to .64 across sessions. Consistencies lay between .42 and .50 and occasion-specificity between .06 and .22. For IAT\( v \), reliability estimates ranged from .60 to .68, with consistencies between .47 and 59 and occasion-specificities between .01 and .21.

Dual construct models of AC\textsubscript{1} and other ISE, ESE, and criterion measures mostly fit reasonably well (see Supporting Information—Appendix B). The dual construct models revealed significant latent trait correlations of AC\textsubscript{1} and AST, \( r_\theta = .13, CI_{95} = [.01, .24], p = .04 \). All other
FIGURE 3 General triple-construct LST model. $Y_i$ = observed measurement $i$ at time $t$, $\eta_i$ = latent state at time $t$, $\theta$ = latent trait common to all occasions, $\zeta_i$ = occasion-specific latent state residual at time $t$, $M$ = stable method factor. Subscripts A, B, and C generically denote different constructs. In our application construct A is ISE, B is ESE and C is a criterion. $\alpha$ and $\beta$ are the latent regression coefficients, with $\alpha$ reflecting the incremental validity of ISE. Correlations of latent state residuals are omitted for clarity of presentation, see text for details.
correlations were pointing in the expected direction but were statistically insignificant, all $p > .05$. No latent state residual correlations emerged between AC$_1$ and ESE as well as criterion measures.

For the latent trait of IAT$_v$, significant latent trait correlations with ESE or criterion measures emerged with BDI, $r_\theta = -.14$, CI$_{95} = [-.5, -.04]$, $p = .01$, with NEU, $r_\theta = -.12$, CI$_{95} = [-.29, -.04]$, $p = .02$, with PBI-O, $r_\theta = -.15$, CI$_{95} = [-.5, -.05]$, $p < .01$, with RSES, $r_\theta = .16$, CI$_{95} = [.06, .2]$, $p < .01$, and with SSES, $r_\theta = .14$, CI$_{95} = [.04, .2]$, $p < .01$. The other correlations had the expected sign but were not statistically significant, all $p \geq .06$.

Triple construct models fit acceptably and revealed a significant incremental prediction of PBI-O by IAT$_v$, $\alpha = -.13$, CI$_{95} = [-.02, .0]$, $p = .03$, all other $p \geq .08$.

### 3.4.5 Exploratory investigation of non-ISE processes in the SE-IAT

Because participants might rely on the valences of target constructs (here: "me" and "other") in recoding the compatible block, the Re parameter might also capture variance that is related to the evaluation of the self. Furthermore, as we argued above, the evaluation of the other-category might be as important in predicting some mental health outcomes as the evaluation of the self. Therefore, we exploratively investigated the Re (recoding) and the A$_2$ (other-esteem) parameters in the same manner as the ISE parameters. These analyses were not preregistered. The results can be found in the respective appendices in the Supporting Information.

**Recoding (Re)**

A mono-construct LST model fit the probit-transformed Re estimates well, $\chi^2(40) = 39.58$, $p = .49$, RMSEA < .01, $p = .99$. The reliability was estimated at .59 with a consistency of .43 and an occasion-specificity of .15 on all occasions. Dual construct models fit the data acceptably and yielded a significant correlation between the Re latent trait and the BDI latent trait, $r_\theta = -.16$, CI$_{95} = [-.29, -.04]$, $p = .01$, and PBI-O, $r_\theta = -.19$, CI$_{95} = [-.32, -.06]$, $p < .01$, all other $p \geq .07$. Triple construct models revealed significant incremental validity of the Re latent trait in predicting BDI, $\alpha = -.09$, CI$_{95} = [-.19, -.00]$, $p < .05$, and PBI-O, $\alpha = -.18$, CI$_{95} = [-.31, -.04]$, $p < .01$, all other $p \geq .09$.

**Other-esteem (A$_2$)**

A mono-construct LST model fit the probit-transformed A$_2$ estimates acceptably well, $\chi^2(40) = 56.94$, $p = .04$; RMSEA = .04, $p = .84$. The expectation value of $\theta$ for A$_2$ was not significantly different from 0, $E(\theta) = -.03$, CI$_{95} = [-.06, .01]$, $p = .10$. The reliability was estimated at .57 with a consistency of .46 and an occasion-specificity of .12 in all occasions. Dual construct modeling yielded no significant correlations of the A$_2$ parameter latent trait and latent traits of criterion measures, all $p \geq .17$. Triple construct models yielded a significant $\alpha$ values for the prediction of SWLS, $\alpha = .11$, CI$_{95} = [.00, .22]$, $p < .05$, all other $p \geq .11$. Model fit of these triple-construct models was acceptable.$^4$

### 3.4.6 Exploratory investigation of ISE-ESE interactions

Since many previous studies investigated ISE-ESE interactions in the prediction of criteria but the implementation of latent interaction variables was out of the scope of this paper, we implemented a less sophisticated analysis to this end. After averaging test scores for each measure (i.e., not the test-half scores) across all available occasions and standardizing the resulting means we ran multiple regressions of the criteria on the ISE parameters, the SSES mean, and their interaction. Among the 21 models (3 ISE parameters $\times$ 7 criteria) only one showed a significant interaction term, which will not be discussed further since encountering a significant result in 21 tests closely matches the type I error probability.

### 4 DISCUSSION

We investigated the incremental validity of the latent stable trait component of multiple supposed ISE parameters in predicting plausible self-esteem correlates over and above ESE. Overall the results provided very little support for our hypotheses concerning the incremental prediction of criterion variables by supposed ISE measures and are therefore in line with fundamental criticism of these measures and the concept of ISE as an association itself.

Applying latent state-trait analyses to a me-positive association parameter (A$_1$) extracted from the SE-IAT by virtue of multinomial process modeling (Meissner & Rothermund, 2013) yielded lower reliability and higher occasion-specificity estimates compared to ESE measures. Nonetheless, the greater part of its reliable variance could still be attributed to a stable latent trait. This finding supports findings reported in previous research with a similar methodology (Dentale et al., 2019) and is consistent both with the comparatively high malleability of SE-IAT and NLT scores (Baccus et al., 2004; Dijkstraerhuis, 2004; Espinosa et al., 2018; Grumm et al., 2009; Jraidi & Frasson, 2010; O’Connor et al., 2011; Quirin et al., 2018; Stewart et al., 2017) and
their moderate retest reliability (Bosson et al., 2000). The $A_1$ parameter overall exhibited lower reliability than the conventional $D_2$ score. Variance proportions of the NLT scores varied strongest across occasions but tended to show a slightly higher consistency than the $A_1$ parameter and $i_1$. The expectation value of all ISE latent trait parameters was significantly positive.

We found no correlation between neither $A_1$ nor the conventional $D_2$ score and the NLT scores on the latent trait level. Hence, the null correlations between NLT scores and SE-IAT scores found repeatedly in the past, could not be explained by measurement error, occasion-specific variance, or confounding processes in the IAT (i.e., recoding or the evaluation of the “other” category) as we initially expected. The only conclusion can be that the NLT and the SE-IAT, in fact, do not measure the same supposed implicit self-evaluation. Even when minimizing multiple potential sources of noise with sophisticated analyses, there is still no sign of a common underlying construct to these tests.

In terms of convergent validity, the NLT and the $D_2$ score outperformed the $A_1$ parameter in that they showed significant correlations with the ESE measures and with overlapping subsets of the criteria. The $A_1$ trait parameter was only significantly correlated with two criterion measures. While the majority of the other correlations had the expected sign, they were not significant and far from reaching the level of ESE-criterion latent trait correlations. The NLT latent trait overall displayed the strongest relations to the criteria out of the potential ISE parameters. Analyses involving other supposed ISE parameters extracted from the SE-IAT by means of the diffusion and the quad model yielded no further insights.

Investigating the incremental validity of supposed ISE measures in predicting criterion measures over and above ESE on a latent trait level, we obtained a pattern of mostly statistically insignificant relations. Among the $A_1$ and the $D_2$ parameter and the NLT score, each measure was able to predict only one unique criterion measure out of six the over and above ESE. In light of this enigmatic and unsystematic pattern, we refrain from detailed interpretations of the few significant coefficients with regard to the involved constructs. Analyses of ISE-ESE discrepancies also produced no further insights.

Exploratory analyses showed that the evaluation of the “other” category in the SE-IAT as modeled by the ReAL models $A_2$ did not correlate with the criteria. This points to the fact that the variance in composite IAT parameters produced by the evaluation of this contrast category—contrary to our argumentation—does not mask relations between self-evaluations and their potential correlates.

An exploratory analysis demonstrated that the Re parameter of the ReAL model, which indicates the extent of recoding processes in the SE-IAT, showed slightly stronger incremental validity than the actual ISE parameters by predicting two criteria instead of one, one of which was identical with the one the $D_2$ score incrementally predicted. This finding casts doubt on the proposition that evidence for the validity of the $D_2$ score hints towards an actual implicit association as conventionally assumed. Rather, it seems that measures of other, non-associative cognitive processes that are elicited during the IAT in order to simplify responding might be more informative.

### 4.1 Implications

The methodologically sophisticated approaches advanced in this study to control for situational factors and confounding factors were not able to establish a robust link between (a) the NLT and SE-IAT parameters and (b) convincing evidence for the incremental validity of those measures in predicting self-esteem correlates over and above an explicit self-esteem questionnaire on a latent trait level. In our view, these results give further weight to the criticism put forward in previous research and reviews of the literature (Bosson et al., 2000; Buhrmester et al., 2011; Falk et al., 2015).

Based on our findings and also on previous attempts at establishing the validity of the NLT and the SE-IAT as ISE measures, two additional important points should be highlighted. First, the NLT and the SE-IAT do not measure the same construct. Bosson et al. (2000) used the metaphor of blind men touching different body parts of an elephant and describing them as if they were touching different animals as an analogue for the different ISE measures capturing different aspects of the same construct. In view of the presented results, we now have to state: There is no elephant there. The blind men are really touching different animals.

Second, and perhaps even more importantly, the SE-IAT and the NLT both do not measure something that is robustly informative beyond explicit self-esteem questionnaires concerning important outcomes like depressive symptoms or neuroticism. While we found small correlations between ISE parameters and ESE, the former were not consistently predictive of our criteria after controlling for ESE. Therefore, we have sufficient reason to question whether the constructs measured by the IAT and the NLT can justifiably be called “implicit self-esteem”. While the answer might largely depend on the definitions used (for an overview and critique of the uses of the term “implicit”, see Corneille & Hüttner, 2020), we would argue that the original promise of implicit measures to assess something that questionnaires are unable to assess cannot be held with regard to what have been considered to be
implicit measures of self-esteem, so the term “implicit self-esteem” is misleading at best. These claims are not new in any way. However, since they have been comprehensively supported by the much-cited review by Buhrmester et al. (2011) there has been no sign of a more careful nomenclature in the field. For another decade, the SE-IAT and the NLT were still widely treated as interchangeable measures of “implicit self-esteem”.

Using this term sparks manifold intuitive associations based on what is known about (directly measured) self-esteem and its relations to other constructs as well connotations of the term implicit (e.g., unconscious, automatic). However, the validity of these intuitions is not robustly supported by the evidence. As the results of the present study demonstrated, the supposed measures of “implicit self-esteem” may be slightly influenced by an individual’s “self-esteem” but they deliver little information beyond questionnaire measures. This conclusion, however, does not provide a convincing explanation for the positive findings that were reported in the literature, with regard to the predictive validity of implicit measures of self-esteem. Of course, we do not want to insinuate that all of these reported findings are unreliable or spurious. Still, many of these studies (A) did not have sufficient power for a robust estimation of relations between constructs (Schönbrodt & Perugini, 2013), (B) did investigate convergent validity instead of incremental, (C) did not control for extraneous processes in implicit measures that might have produced these correlations without actually reflecting automatic associative processes of self-evaluation. Thus, based on the results of the current study, we feel justified in taking a somewhat skeptical stance towards interpretations of these findings in terms of implicit self-esteem.

In the field of evaluative learning, Kurdi and Banaji (2017) proposed using descriptors for learning paradigms that are inspired by their superficial features and not based on theories about the hypothetical underlying processes to avoid confusion and “fuzzy thinking” (p. 195). We propose that this convention would be helpful for the topic of ISE as well. Abstaining from intuitions about implicit self-esteem and its supposed measures and also from associative theories and frameworks of self-esteem might thus lead to clearer and more precise thinking about the cognitive and affective foundations of self-esteem. Conscientiously differentiating measures and constructs (De Houwer & Moors, 2010) means accepting the fact, that the NLT first of all measures a preference for initials and that the SE-IAT first of all measures performance differences in a specific double categorization task. Those, in turn, might be loosely linked to different psychological constructs or behaviors, but the evidence for that is hard to come by as this study and previous studies have shown. Removing the SE-IAT and the NLT from their unjustified privileged position in the ISE researcher’s arsenal would allow for the necessary development and widespread adoption of new and more promising measures of implicit self-evaluative associations (e.g., Krause et al., 2012). Going a step further, we suggest that the focus on evaluative associations as the sole candidate of cognitive representation of self-esteem is an unnecessary constraint (Dentale et al., 2020; Meissner et al., 2019; Rothermund et al., 2020). We also recommend using existing implicit measures of propositional and motivational processes which have been developed recently (Demeyer et al., 2018; Dentale et al., 2020; Koranyi et al., 2017; Müller & Rothermund, 2019; Remue et al., 2014) to do justice to the multifaceted literature that exists on ESE, for example on its relation to self-efficacy (Rosenberg et al., 1995), self-concept (Campbell, 1990), and self-related beliefs (Wigfield & Eccles, 1994).

Systematic basic research using a wide range of implicit measures may allow us to paint a more robust picture of the (implicit) cognitive and affective underpinnings of self-esteem, self-confident behavior, and cognition, as well as physical and psychological well-being than we are able to at this point. Until then researchers should refrain from using the term implicit self-esteem in a way that seems to presuppose that the elephant has already been found.

### 4.2 Limitations and strengths

When discussing the implications of our findings, it is important to emphasize that our results regarding the predictive validity of the SE-IAT and NLT are specific to the outcomes we investigated and should not be generalized to other potential correlates of ISE. We specifically selected criteria which were (A) closely linked to self-esteem either conceptually or in terms of shared processes without the necessity of complex theoretical frameworks or hypotheses to invoke said link, and (B) were not shown to be predicted robustly by supposed ISE measures in previous studies. This combination of properties has been a prime reason for raising doubts regarding the validity of supposed ISE measures and was therefore given priority in this investigation. Testing the incremental validity of supposed ISE measure in predicting the selected criteria, we wanted to (A) prevent ambiguities about whether a relation between supposed ISE measures and the criteria is actually to be expected, and (B) investigate whether sophisticated methodology will overcome the outlined obstacles (situational variability and construct-irrelevant variance in ISE measures). Our choice of criteria in combination with our methodological approach should produce results that are unequivocal where they have been ambiguous in the past. Accordingly, we see a lot of reason for skepticism regarding the validity of supposed ISE
measures in the fact that this undertaking still produced mostly null findings. Nonetheless, supposed ISE measures may of course be robustly linked to other constructs that we did not investigate. It also has to be noted that our set of criteria consisted exclusively of self-report measures, which may be less sensitive to detect potential influences of ISE. Future studies should include criteria that are more objective or indirect in nature in order to decrease the probability that method variance is the primary driver of correlations and to provide indicators that may be more sensitive to detect the influence of ISE.

What is more, our design was exclusively aimed at the investigation of effects on the trait level. We did not employ manipulations (e.g., threat, negative or positive events) nor did we assess any situational moderates that might interact with ISE in predicting the outcomes. Accordingly, our findings are not incompatible with assuming more situation- and context-specific predictive validity of ISE measures. On the other hand, we see the establishment of the presence or absence of robust effects on the trait level as an important part of the investigation of the validity of supposed ISE measures as ESE robustly shows these situation-independent effects. In addition, this kind of investigation can also inform a more fine-grained investigation of context-specific effects, at least from some perspectives. For example, from a health psychologist’s point of view, one might wonder whether the investigation of context-specific effects of ISE on affect, mood, or pathological behavior is indicated or what these findings actually tell us about the long-term implications of ISE, when there is no robust evidence for a link between ISE and mental health outcomes on a trait level.

From a methodological standpoint, our findings are limited in that model fit was barely acceptable for some models. Our sample size also does not yield adequate power to detect very small effects. To ensure a power larger than .80, the critical effect size of the central path coefficients was found to be .15 when testing with one-tailed tests. However, this estimate is based on the model most disadvantageous for the detection of incremental validity and can be seen as a lower bound. Furthermore, we would argue that the presence or absence of even smaller effects is not central to the interpretation of the results we put forward.

Another limitation is the fact that the adaptation of the SE-IAT procedure has to be taken into account when comparing our findings to results produced using the standard IAT procedure. It is possible that a non-speeded IAT relies on different processes than the speeded version we used. On the other hand, the speeded version could be argued to rely more heavily on implicit (in the sense of “fast”) processes. What is more, Calanchini et al. (2021) systematically varied different procedural parameters of the IAT in order to investigate the range of applicability of the ReAL model and found that the ReAL model yields similar results irrespective of the procedural variants of the IAT. This supports the inference that our findings with the adapted IAT variant can be generalized and do not reflect the influence of specific procedural factors.

What is more, it is important to point out that our participants were sampled from the general population. Hence, our findings concerning the relation of supposed measures of implicit self-esteem and depressive symptoms may not be generalizeable to clinical depressions.

The last point refers to the data collection in an online setting, which could potentially decrease measurement reliability due to increased disturbances during the study and lower commitment of the participants. However, comparing our results to similar studies shows that the estimated reliability of the D2 was only slightly lower, which could be explained by the procedural changes, and that of RSES scores is even higher (Dentale et al., 2019).

The online setting also brings with it some benefits for our study. First, data acquisition is close to the everyday life of the participants. Accordingly, the results can be assumed to have some external validity. Second, the sample was more heterogeneous than samples in many psychological studies. This decreases the risk of generalizing a finding that is specific for a certain subpopulation. What is more, we introduced a sophisticated method to analyze IAT data by combining process modeling and latent state-trait modeling. This strategy allowed for detailed insights and the controlling of multiple sources of error.

4.3 Conclusion

The present study gives further weight to criticism of the conventionally used measures of ISE and the construct itself. Combining process modeling and latent state-trait modeling showed that the reliable variance in supposed ISE indicators can largely be attributed to a stable latent trait. However, our analyses yielded little support for the convergent and the incremental validity of these stable traits in predicting interpretation bias, neuroticism, depression, life satisfaction, or remembered parenting behavior over and above ESE. We argue that these results call for a shift in perspective and terminology that is long overdue in most of the research involving ISE. In our opinion, there are no more excuses for ignoring or relativizing results that cast doubt on the validity of the conventional ISE measures at this point. To strengthen the intellectual rigor in the field, our recommendation is to refrain from using the term “implicit self-esteem” without actual justification and instead set out for more systematic basic research on the cognitive and affective
foundations of self-esteem using all implicit measures of associations, propositions, and motivation that are at our disposal.

ACKNOWLEDGMENT
Open access funding enabled and organized by ProjektDEAL.

CONFLICT OF INTEREST
We have no known conflict of interest to disclose.

AUTHOR CONTRIBUTION
None.

ETHICS STATEMENT
The procedure of this study was approved by the Ethical Commission of the Faculty of Social and Behavioral Sciences of the University of Jena (reference number FSV 19/068).

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are openly available in OSF at https://osf.io/ahvge.

ORCID
Adrian Jusepeitis https://orcid.org/0000-0001-5348-5513
Klaus Rothermund https://orcid.org/0000-0002-9350-5272

ENDNOTES
1 Since measure and construct have to be carefully differentiated (De Houwer & Moors, 2010), we will speak of the discussed measures as “supposed ISE measures”. While this might reduce readability, it is crucial to the perspective advanced in this paper.

2 The first wave of data collection started with \( n = 120 \) in January of 2020. After peer review of the present paper a larger sample size was deemed necessary, which is why a second wave of data collection with initially \( n = 240 \) started in June of 2021. The second data collection was preregistered separately: https://osf.io/gpqme.

3 An intermediate power-analysis based on the data of the first wave of data collection was conducted and showed \( N = 250 \) to be a reasonable sample size. Because the final \( N \) was greater than 250 and the characteristics of the full sample were slightly different from those of only the first wave, a post-hoc power estimation based on the final characteristics of the full data is reported below.

4 Applying the same analyses to the quad models other-negative parameter yielded no result of interest. See appendices.

5 However, we speculate that the reliability of the mean of two \( A_1 \) estimates based on two test-halves might be more realistically interpreted as a lower bound of the reliability of an \( A_1 \) estimate based on all trials. Contrary to item scores in a questionnaire test, the information used for multinomial process modelling might add up in a non-linear way: Doubling the information going into the estimation might result in more than twice the information yielded by the estimation. This speculation, however, would have to be investigated using simulation studies.

REFERENCES


Jusepeitis, A., & Rothermund, K. (2022a). Data for “No elephant in the room: The incremental validity of implicit self-esteem measures”. osf.io/x5c6s


SUPPORTING INFORMATION
Additional supporting information may be found in the online version of the article at the publisher’s website.

How to cite this article: Jusepeitis, A., & Rothermund, K. (2022). No elephant in the room: The incremental validity of implicit self-esteem measures. Journal of Personality, 00, 1–21. https://doi.org/10.1111/jopy.12705