Abstract

According to dual-process models, implicit self-esteem (SE) is based on automatic self-associations which can be measured with indirect techniques based on an associative conception of implicit cognition (e.g., Implicit Association Test; IAT). However, alternative theoretical proposals (e.g., Relational Frame Theory; RFT) propose that implicit SE might not be based on automatic self-associations, but on implicit propositional self-evaluations that can be captured only with non-associative implicit measures (e.g., Implicit Relational Assessment Procedure; IRAP). In the present study, both reliability and validity of a new propositional measure of implicit self-esteem (Relational Responding Task; RRT) were assessed, and compared with the SE-IAT and with two self-report scales of self-esteem. In the first study, two alternative self-esteem RRTs (i.e., SE-RRT and RSE-RRT) were administered along with a SE-IAT and other scales, to assess reliability and validity issues. The results showed: 1) acceptable, though not optimal, reliability for both RRTs, 2) an adequate support for convergent validity, with significant correlations between implicit and explicit measures of SE, 3) the criterion validity was supported for the RSE-RRT (with significant correlations with all theoretically-linked scales), while only partially supported for the SE-RRT (with a significant correlation only with depression, 4) RRTs were not significantly correlated with impression-management and self-deception and 5) incremental validity of implicit propositional SE on depression, controlling for automatic SE associations and explicit self-esteem. In a second study, it was experimentally demonstrated that SE-RRT showed levels of “fakeability” similar to a classical implicit-self-esteem measure like the SE-IAT, and considerably lower than SE scales.

Implicit self-esteem; Relational Responding Task; Dual models; Implicit social cognition.

Introduction

In a traditional psycho-social conception, Self-Esteem (SE) was conceived as the distance between the actual and ideal Self, with a specific focus on human expectations of success and failure. More recently, SE was defined as the evaluative component of self-concept (e.g., Markus, 1977), stimulating the birth of SE empirical investigation. Within this research tradition, several self-report scales have been developed with robust psychometric properties in terms of reliability as well as of convergent and criterion validity (e.g., Buhrmester, Blanton, & Swann, 2011). However, notwithstanding these qualities, SE scales showed two important limitations: the proneness to impression management bias (e.g., Cai et al., 2011), and the difficulty of capturing self-related information using introspection, both for self-deception effects (e.g., Hofmann, Gschwendner, Le, & Schmitt, 2005) and for participant' inabilities linked to emotional awareness (e.g., Dentale, San Martini, De Coro, & Di Pomponio, 2010; Dentale, Vecchione, De Coro, & Barbaranelli, 2012).

In recent decades different models, that assume the distinction between implicit and explicit social cognition, were theorized and empirically tested (for a review see Gawronski & Payne, 2010). In line with these models, self-esteem, along with many other constructs (e.g., attitudes, stereotypes, personality traits, etc.), has been reformulated from an implicit social cognition perspective (e.g., Gawronski & Payne, 2010). For instance, from a dual models perspective, implicit self-evaluations depend on simple mnemonic associations between self-concept and positive-negative attributes, which can be automatically activated through specific environmental patterns of stimulation (Gawronski & Bodenhausen, 2006). On the contrary, explicit self-evaluations are conceived as propositional judgments depending on reflective processes (Strack & Deutsch, 2004) that are based on the possibility of ascribing a logical 'truth' value to them. In recent decades, several measures were developed to capture indirectly automatic associations towards the self, such as the Implicit Associations Test (SE-IAT; Greenwald & Farnham, 2000) and its variants.

In the following paragraphs the SE-IAT along with the associative conception of implicit self-esteem are described and discussed in their methodological and theoretical limits. Successively, an alternative theory of implicit SE, which assumes that self-esteem evaluations are based on automatic propositional beliefs, is introduced along with a classical measure consistent with this perspective, the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes et al., 2006).

Finally, notwithstanding IRAP's merits, some limits of this experimental paradigm are described, and a new implicit propositional measure designed to overcome these limits, the Relational Responding Task (RRT; De Houwer, Heider, Spruyt & Roets, 2015) is introduced. The present research aims to apply for the first time the Relational Responding Task (RRT) to measure implicit self-esteem, and to evaluate a series of reliability and validity issues of this new instrument.

The Associative Conception of Implicit Self-esteem: Limits of the SE Implicit Association Test

From a dual-process model perspective, implicit SE is traditionally defined as the association between the self and positive *vs*. negative attributes (Greenwald et al., 2002). The most used and tested measure developed to assess automatic SE associations is the SE-IAT (Greenwald & Farnham, 2000) which is based on self-related word categorization tasks. The SE-IAT is a computer-based task designed to measure automatic associations between two opposing target categories (i.e., self *vs*. others) and two opposing attribute categories (i.e., positive *vs*. negative). For each trial, participants are instructed to classify a stimulus-word (e.g., me, incompetent, ext. *vs*. them, competent, etc.) as quickly and accurately as possible into "self-positive" *vs*. "others-negative" categories. In the first combined block, the target and attribute categories are presented with a specific associative pattern (e.g., self-positive *vs*. others-negative). In a second combined block, the position of target categories is switched (e.g., others-positive *vs*. self-negative). The size of self-related automatic association can be calculated as the difference between the mean latencies of the first and second combined blocks.

As clarified in two recent reviews (Buhrmester, Blanton & Swann, 2011; Falk & Heine, 2014), only weak evidence has been reported supporting the convergent and criterion validity of the SE-IAT, questioning the assumption that it actually measures implicit self-esteem. Even if, from a dual process model perspective, automatic associations and reflective self-evaluations depend on different processing systems, a certain degree of concordance between implicit and explicit SE is theoretically expected. Dual models assume that spontaneous self-evaluations can be normally translated into propositional format (Strack & Deutsch, 2004) leading to considerable implicit-explicit correlations, except when introspective limits (e.g., in psychopathological participants) and presentation bias (e.g., in personnel a selection context) assume critical values. Surprisingly, meta-analytic data (Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005) showed that the SE-IAT is only weakly related to self-report measures (*r* =.13). It is worth noting that this weak correlation does not necessarily represent an evidence for a lack of validity but it opened some relevant questions concerning: 1) the low reliability of the SE-IAT, 2) what exactly the SE-IAT measures and3) possible moderators of the implicit-explicit SE relationship.

A second important point that threaten the validity of the Implicit Association Test as a measure of SE derives from the low correlations found between the SE-IAT and other implicit measures of SE (Buhrmester et al., 2011; Falk & Heine, 2014) based on different functioning principles, such as tasks based on priming effects (e.g., Affective Priming Task; e.g., Spalding & Hardin, 1999) or on preference towards autobiographical information (e.g., the Name Letter Task; Hoorens, 1990). Weak or close to zero correlations were recently found between the SE-IAT and several other measures of implicit SE (Falk, Heine, Takemura, Zhang & Hsu, 2014), failing to provide evidence for its convergent validity. These results could depend both on the low reliability of implicit measures and on the different processes activated by each specific task. In the latter case, this suggests that different measures of implicit SE do not refer to the same construct, stimulating again the question of what the SE-IAT actually measures.

A further important point is the lack of compelling evidence for the criterion validity of the SE-IAT. In this vein, an extensive review (Buhrmester et al., 2011) showed that the SE-IAT exhibit weak correlations with theoretically-linked criteria, such as psychological well-being, depression, physical health problems, and others. Since negative self-views and depressive symptoms are linked by a consolidated theoretical and empirical research tradition (e.g., Williams, 1997; Ingram, Miranda, & Segal, 1998), of particular interest here is the relationship between SE-IAT scores and depression. Unexpectedly, several studies revealed that although depressed and non-depressed individuals showed different levels of explicit SE, their mean scores on the SE-IAT were nearly equal (e.g., De Raedt et al., 2006; Remue, Hughes, De Houwer & De Raedt, 2014). For instance, in a study including both depressed participants and healthy controls, De Raedt and colleagues (2006) showed that SE-IAT scores, as well as scores on two other implicit self-esteem measures, showed similar levels of implicit self-esteem between groups (see Remue, De Houwer, Barnes-Holmes, Vanderhasselt, De Raedt,2013). Questioning the classical assumption of cognitive therapy regarding the relationship between negative automatic self-evaluations and depression, this dramatic contradiction stimulated researchers to more deeply understand what exactly the SE-IAT measures.

A propositional Conception of Implicit Self-esteem: The Implicit Relational Assessment Procedure (IRAP) and its Limits

In order to provide a possible interpretation for these unexpected results, it was hypothesized (e.g., De Raedt et al., 2006) that implicit measures based on associative models of implicit social cognition (e.g., SE-IAT) might confound two different kinds of self-evaluation, what we currently perceive ourselves to be (i.e., actual SE) and what we want to be (i.e., ideal SE). Actual self-evaluations refer to positive or negative judgments about the current-self (e.g., I am good), while ideal self-evaluations refer to judgments about our desirable or undesirable self (e.g., I want to be good). Notably, the distinction between actual and ideal self-evaluations depends on the relational information that links target-categories and attributes (e.g., I am good ≠ I want to be good) and not on mere association between them. Since associative implicit measures (e.g., the SE-IAT) cannot capture this relational information, they are not suited for distinguishing between actual and ideal SE.

Recently, new theoretical proposals based on functional contextualism were advanced, assuming that implicit evaluations depend on an automatic activation of propositions and not on automatic mental associations (e.g., Hughes & Barnes-Homes, 2013; De Houwer, 2014): “A vital difference between propositions and associations is that propositions contain relational information, that is, information about how concepts are related. It is often also assumed that propositions about events can be formed not only on the basis of the repeated experience of those events but also as the result of a single instruction or inference concerning those events" (p. 344).In this vein, De Houwer (2014) provided some evidence demonstrating that propositional evaluations can be formed automatically (e.g., Heider, Spruyt, & De Houwer, 2014), and can also be automatically retrieved (for a model of memory in accordance with this hypothesis, see Hintzmann, 1986). In particular, in Relational Frame Theory (RFT; Hayes, Barnes-Holmes & Roche, 2001), a behavioral analytic approach for the study of language and cognition, it is assumed that: "humans are capable of learning in ways that differ markedly from non-humans. Specifically, our ability and tendency to relate stimuli bi-directionality allows for the emergence of complex untrained relations that cannot be traced to a history of direct training or learning. According to RFT, this form of relational learning emerges early on in our development through interactions with the verbal community and is an important defining element of both human language and cognition" (p.7-8). On the basis of RFT, the Relational Elaboration and Coherence (REC) model was proposed (e.g., Hughes & Barnes-Homes, 2013) that reformulates the dichotomy between implicit vs. explicit cognitions within the functional contextualism epistemology in order to avoid the assumption of mediating mental constructs. More specifically, in the REC model, implicit and automatic cognitions correspond to Brief and Immediate Relational Responses (BIRRs) that occur relatively quickly after a given stimulation. These quick relational responses are followed by additional relational responses called Extended and Elaborated Relational Responding (EERRs) and that correspond to explicit cognitive processes. These additional relational responses may occur toward the stimulus itself or toward the initial response to that stimulus. With sufficient time, these additional relational responses will likely form a coherent relational network. However, as for dual process models of social cognition, also in a functional perspective Immediate Relational Responses and Extended and Elaborated Relational Responding can diverge as a function of moderators linked to impression management and introspective limits of participants. Moreover, whilst in a cognitive psychology perspective proposition-based evaluations are conceived as reflective judgments, and thus not linked to implicit processes that are conceived as associative in nature (e.g. Strack & Deutsch, 2004), Hughes & Barnes-Holmes (2013) clarified that "when implicit cognition is defined functionally (in terms of relational responding) a non-associative indirect procedure (*for measuring*) is not only possible but necessary" (p.12).

Notably, in accordance with these theoretical assumptions, to capture implicit propositional evaluations, are necessary instruments that used stimuli specific statements that include relational information about how target categories and attributes are linked with each other. In this view, Remue et al. (2013) applied the Implicit Relational Assessment Procedure (IRAP; Barnes-Holmes et al., 2006) to separately measure actual and ideal implicit SE. Based on the Relational Frame Theory (RFT; Hayes, Barnes-Holmes, Roche 2001), the IRAP is able to measure not mere mnemonic associations, but true beliefs with their intrinsic relational information. The IRAP is a reaction time task requiring respondents to categorize stimuli following a stimulation pattern consistent with their actual beliefs (e.g., high actual SE statements), and another one inconsistent with their beliefs (e.g., low actual SE statements). For example, respondents are instructed to categorize high SE statements, such as "I am successful" or "I am not incompetent" as true on half of all trials (consistent trials), and to categorize them as false on the other half (inconsistent trials; see Figure 1).

INSERT FIGURE 1 ABOUT HERE

At the same time they are instructed to categorize low SE statements, such as "I am not successful" or "I am incompetent" as true on half of all trials (inconsistent trials) and as false on the other half (consistent trials). Computing latency differences between consistent and inconsistent trials makes it possible to indirectly assess the target self-evaluations. To separately measure actual and ideal SE, Remue et al. (2014) administered, to depressed and healthy participants, an actual SE-IRAP and an ideal SE-IRAP along with an SE-IAT. In line with precedent studies (De Raedt et al., 2006), the results indicated that, surprisingly, depressed and healthy participants revealed similar (positive) mean scores on the SE-IAT, suggesting that all participants more strongly associated the self with positive rather than with negative attributes. IRAP mean scores indicated, in line with cognitive therapy hypotheses (e.g., Williams, 1997; Ingram, Miranda, & Segal, 1998), that depressed participants showed a pattern of lower actual SE and higher ideal SE if compared with healthy participants. However, notwithstanding these interesting results, not all of Remue et al.'s (2014) expectations were confirmed. For instance, according to the authors, the lack of a significant mean difference between depressed and healthy participants on the SE-IAT can be ascribed to the fact that the IAT tends to capture primarily ideal-self in the former group and actual-self in the latter. However, unexpectedly, no significant correlations were found neither between SE-IAT and ideal SE-IRAP in depressed participants, nor between SE-IAT and actual SE-IRAP in healthy ones. These results might depend on the low reliability of actual and ideal SE-IRAP, which showed rather low split-half correlations (respectively r_{tt} =.53 and r_{tt} =.22). In accordance with this view, De Houwer, Heider, Spruyt and Roets (2015) noted that several participants (about 20%) were not able to complete the task due to intrinsic performance difficulties linked to this experimental paradigm. This high rate of incomplete tasks would be a consequence of trial to trial variation of “true” and “false” response buttons (e.g., press “E” for true and “I” for false on certain trials, and press “E” for false and “I” for true on the other trials) that characterizes this task.

A New Relational Measure of Implicit Beliefs: The Relational Responding Task (RRT)

To overcome the IRAP's limits, De Houwer et al. (2015) developed the Relational Responding Task (RRT), a new implicit measure designed to assess human beliefs. In performing the RRT, participants were invited to respond consistently with certain beliefs (e.g., “I am a valid person”).A series of whole propositions (specifically developed for the RRT) were presented on the monitor, and participants are required to respond “as if” they were true, or “as if” they were false (see Figures 2 and 3).

INSERT FIGURE 2a and 2b ABOUT HERE

In the first block, respondents were instructed to press the “true” button when statements indicating high SE (e.g., “I am a valid person”) were shown, and to press the “false” button when low SE statements (e.g., “I am not good at all”) were presented. Conversely, in a second block, participants were invited to press the “true” button for statements indicating low SE (e.g., “I am not good at all”), and the “false” button for statements indicating high SE (e.g., “I am a good person”). As for the SE-IAT and SE-IRAP, final scores can be obtained by computing mean latency differences between these two blocks for each participant (see the “Measures” section for D scores computation). Unlike the IRAP, in the RRT, the location of true and false buttons is the same for all trials. The risk of recoding strategies through positional information is minimized using inducer trials with words or sentences necessarily “true” or “false” (e.g., “real” vs. “unreal”; see the Supplementary Materials for a complete list of these stimuli).

Since RRT and IAT experimental paradigms are rather similar, the former should enjoy the practical advantages linked to IAT’s simplicity and robustness. Notwithstanding these similarities, the RRT shows various important differences with respect to the IAT. In particular, while participants in the SE-IAT participants are instructed to categorize a series of stimuli (e.g., “Me”) into the correct category (e.g., “Self”), participants in the RRT have to understand the relational information embodied in the statements (“I am good” or “I want to be good”) and then use the propositions’ meaning to categorize them “as if” they are true or false. Therefore, participants in the RRT must necessarily use relational information to respond, so it is able to capture not mere associations between concepts but true beliefs (e.g., “I am good” or “I want to be good”).On a theoretical point of view, implicit and explicit propositional beliefs pertaining to self-esteem can be deeply different with each other as a consequence of their different proneness to moderators linked to impression management and introspective ability. In particular, it is expected that confounding factors, like social desirability, self-deception and lack of introspective awareness, can considerably influence response to self-report items but are not able to affect neither implicit associative nor implicit propositional measures. To date, various studies have applied the RRT to measure implicit beliefs in different research domains, such as parenting style towards adolescent alcohol consumption (Koning, Spruyt, Doornwaard, Turrisi, Heider & De Houwer, 2016), smoking desires

(Tibboel, De Houwer, Dirix, & Spruyt, 2017) and actual vs. ideal sexual identity (Dewitte, De Schryver, Heider & De Houwer, 2017). There is initial evidence for its reliability ($r_{tt} \geq .64$) and construct validity.

Aims of the Research

To overcome both the theoretical limits of the SE-IAT (i.e., its mere associative nature), and the methodological limits of the SE-IRAP (i.e., too difficult for several participants), in the present paper the RRT was applied to measure implicit SE with the aim to provide evidence for its reliability and validity.

In the first study, two self-esteem RRTs were developed in order to assess: 1) internal consistency, convergent validity (in terms of correlations with the SE-IAT and with two self-report measures of SE),and criterion validity with respect to theoretically-related scales (i.e., optimism, life satisfaction, depression scales and two measures of self-enhancement),2) correlations with self-deception and impression management scales, and3) incremental validity of implicit propositional SE (as measured with the RRT)on depression with respect to automatic self-associations (as measured with the SE-IAT)and explicit SE (as measured with RSES and a semantic differential SE scale).

In a second experimental study, the proneness to faking of the SE-RRT was investigated and compared with that of SE-IAT and self-report measures of SE.

Study 1

Two RRTs were developed which differed exclusively in the sentences used as stimuli. A first Self-Esteem RRT (SE-RRT)included high and low SE sentences balanced for three strict criteria: 1) the number of characters; 2) the presence of positive and negative words(e.g., I am *satisfied* with myself for I rarely *devaluate* myself, for high SE sentences vs. I am *unsatisfied* with myself or I rarely *appreciate* myself, for low SE sentences); 3) the presence of affirmative or negative sentence formulation (e.g., people like me *can be* satisfied with themselves or confidence *does not lack* in me, for high SE sentences vs. people like me *cannot be* satisfied with themselves or confidence *lacks* in me, for low SE sentences)[1]. Respecting these strict criteria, the target statements prevent participants from using recoding strategies, based on construct-unrelated stimulus properties (i.e., different number of characters, of positive/negative words, and of affirmative/negative formulations for high and low SE sentences) in choosing, trial to trial, the right response button ("E" vs. "I" keys).However, notwithstanding the importance of this methodological balancing between high and low SE sentences, one undesirable consequence is a high complexity in sentence formulation that could lead to high difficulty in performing the task. In this vein, considering that some statements used for this first SE-RRT are not of immediate understanding, a second RRT (with a perfect balancing of positive/negative words between high vs. low SE sentences) was developed, using easier sentences inspired by Rosenberg Self-Esteem Scale (RSE-RRT; see the on line Supplementary Materials for a complete list of stimuli).

The first aim of the present study was to investigate the psychometric properties and internal consistency of RRT measures. A second aim was to evaluate RRTs' convergent and criterion validity, analyzing their correlations with the SE-IAT, the RSES and a semantic differential scale of SE (SE-DS), as well as with a series of SE related-constructs such as optimism, life satisfaction, depressive symptoms and two indices of self-enhancement. A third aim was to compare the correlations of implicit and explicit SE measures with impression management and self-deception enhancement. A last aim was to evaluate the incremental validity of Implicit Propositional SE (IPSE) on depression, controlling for Implicit Associative Self-Evaluations (IASE) and explicit SE (ESE).

Methods

Participants and Procedure

A total of 130Italian students (106 females), mean age = 22.40, (SD= 3.28), were recruited for volunteer participation and contacted to schedule an experimental session in the laboratory. At the beginning of the session, an informed consent form was given to each participant to read and sign.

[1] See online supplementary materials for a complete list of the stimuli used for the implicit measures.

A SE-IAT and two RRTs (composed of different statements) were administered followed by a booklet of self-report scales: the RSES (Rosenberg, 1965), the SE-SD, the Satisfaction with Life scale (SWL; Diener, Emmons, Larsen & Griffin, 1985),the Life Orientation scale (LOT; Scheier, Carver & Bridges, 1994), the Impression Management and Self-Deception Enhancement scales (IM and SDE; Paulhus, 1991),a scale aimed to measure depressive symptoms in the general population (the Center for Epidemiologic Studies Depression Scale, CES-D; Radloff, 1977), and two self-enhancement indices, that are a scale of Self-Assessed Intelligence (SAI; Chamorro-Premuzic & Furnham, 2006), and a measure of the Better Than Average effect (BTAV; Brown, 2012).The order of implicit and explicit measures (as well as the order among implicit measures) were not randomized between participants, as the present study was focused on statistical analyses (i.e., correlations and/or regressions) that are based on the investigation of true individual differences and not on the investigation of true mean scores. In this case the randomization of measures' order is not recommended (Teige-Mocigemba, Klauer, & Sherman, 2010), as it can increase random error in participants' scores and, as a consequence, decrease the relationships among variable under investigation. No problems emerged during the administration of RRTs, and all participants completed every task proposed.

Measures

Self-Esteem implicit association test (SE-IAT). As in the original SE-IAT (Greenwald & Farnham, 2000), participants were instructed to categorize a sequence of stimuli-words into "Me" vs. "Other" target-categories, and "Positive" vs. "Negative" target-attributes. Five stimuli-words were included for each category (see the online Supplementary Materials for a complete list of stimuli), with a random presentation for each block of trials. Seven blocks of trials were shown: two initial training-blocks of 20 trials (blocks 1: Me vs. Other; block 2: Positive vs. Negative) with a single-categorization task, a further training-block of 40 trials (Other vs. Me) with switched single target-categories location, four test-blocks of 40 trials (3–4 and 6–7: Me or Positive vs. Other or Negative; Me or Negative vs. Other or Positive)] with double-combined categorization tasks. The order of test-blocks (3-4 and 6-7) was randomized between participants. Participants were instructed to perform all categorization tasks as quickly and accurately as possible. To compute SE-IAT scores, the D2 algorithm (built-in error penalty procedure) was applied (Greenwald, Nosek, & Banaji, 2003). Higher D2 scores represent higher implicit self-esteem. Two test-halves were calculated applying the D2 algorithm to blocks 3–6 and 4–7 separately to estimate split-half reliability.

Relational Responding Task (RRT). The SE-RRT and RSE-RRT, based on different statements (see online Supplementary Materials), were developed to measure Implicit Propositional Self-Esteem (IPSE). Each RRT included seven blocks of categorization tasks. In the first single-categorization block (20 trials), participants were instructed to categorize 10 stimuli (inducer-words) synonymous with "true" (5 items, e.g., 'correct') or "False" (5 items, e.g., 'incorrect'). Inducer-words were presented in orange font with a randomized order. In the second single-categorization training-block (20 trials), 10 sentences concerning self-evaluation were randomly presented. Half of them were high SE statements (e.g., "I am a valid person"), and the remaining were low SE statements (e.g., "I am no good at all"). Each of these target sentences was presented in a blue font and with a randomized order. participants were invited to categorize each stimulus(as quickly and accurately as possible) as if they were high SE individuals (i.e., using the true response key "E" for high SE sentences and the false response key "I" for low SE sentences). The third and fourth were combined blocks of 40 trials that included both inducer words and target sentences randomly presented. Participants were instructed to categorize inducer words (i.e., synonyms of true and false) following their correct meaning and target sentences as if they were high SE individuals. The fifth block was a single-categorization block of 20 trials that included target sentences randomly presented with an inversion of the categorization key. Participants were instructed to categorize sentences as if they were low SE individuals (i.e., using the true response key "E" for low SE sentences and the false response key "I" for high SE sentences). Finally, the sixth and seventh were combined blocks of 40 trials including both inducer words and target sentences randomly presented. Participants were instructed to categorize inducer words in accordance with their correct meaning, and target sentences as if they were low SE individuals. The order of test-blocks (3-4 and 6-7) were randomized between participants.

During the entire task the response labels “TRUE” and “FALSE” were shown at the top left and top right corner of the monitor, respectively. All sentences appeared in the middle of the monitor until a response was performed. A red cross appeared under the stimulus when participants responded incorrectly and remained on the screen until they provided the right response. Inter-trial intervals were fixed at 750 ms. The D2 algorithm was applied to compute final scores of the RRTs, with the exclusion of both practice trials and inducer trials (see De Houwer et al., 2015). Moreover, latencies exceeding the cutoff of 10,000ms were excluded, and participants with more than 10% of latencies faster than 300 ms were removed. RRT scores were computed so that higher scores reflected a higher SE. Two test-halves were calculated for both SE-RRT and RSE-RRT applying the D2 algorithm to blocks 3–6 and 4–7 separately to estimate split-half reliability.

Rosenberg self-Esteem scale (RSES). To measure ESE we used the RSES (Rosenberg, 1965; for the Italian version see Prezza, Trombaccia, & Armento, 1997), a 10 items instrument with 5 point Likert scale (0–4). Various studies (e.g., Schmitt & Allik, 2005) revealed appropriate levels of internal consistency (Cronbach’s α = .81) and test–retest reliability (.85 < r_{tt} <.88) for this scale, along with clear evidence of construct and criterion validity. An example item is “I take a positive attitude toward myself”. An overall index of SE was computed summarizing all items scores.

Satisfaction with Life Scale (SWSL). The SWLS (Diener et al., 1985; for the Italian version see Di Fabio & Ghizzani, 2007) consists of 5 items designed to rate how participants perceived to be satisfied with their lives on a scale ranging from 1 (strongly disagree) to 7 (strongly agree). The SWLS showed adequate levels of reliability (Cronbach’s α= .87; *rtt* = .82) and validity (Diener et al., 1985). An example item is “In most ways, my life is close to my ideal.” An overall index of life satisfaction was computed summarizing the scores of all items.

Life Orientation Test (LOT). Optimism was assessed with the revised LOT (LOT-R), composed of 10 items (Scheier et al., 1994; for the Italian version see Giannini & Di Fabio, 2008) with a 5 point scale ranging from 1 (strongly disagree) to 5 (strongly agree). The scale included four fillers that were not used for computing the final score. The LOT-R revealed an adequate level of internal consistency (Cronbach's *α*= .82; see Scheier et al., 1994) and test–retest correlations (*rtt* = .79; see Smith, Pope, Rhodewalt, & Poulton, 1989), as well as various evidence of construct and predictive validity. An example item is "I'm always optimistic about my future." A general score of optimism was computed summarizing the items' scores (excluding fillers).

Center for Epidemiologic Studies Depression scale (CES-D). Depressive symptoms (e.g., despondency, hopelessness, loss of appetite and interest in pleasurable activities, sleep disturbance, crying bouts, loss of initiative and self-deprecation, and so forth) were evaluated with the CES-D, which revealed high estimates of reliability and validity (Radloff, 1977; for the Italian version, see Fava, 1983). Participants were requested to rate the occurrence of each symptom in the last week with a 4-point liker-type scale. A sample item is "I was bothered by things that usually don't bother me."

Better Than Average (BTAV). In order to assess the tendency to evaluate the self as better than most of other people (i.e., better than average effect), a questionnaire formed with two parts was administered (Brown, 2012), requesting participants to make a series of judgments regarding 10 personal attributes (i.e., agreeable, conscientious, imaginative, secure, sociable, competent, honest, intelligent, kind and responsible) using a 5 point likert scale. In the first part, they were instructed to rate how well each attribute describes them, while in the second they were invited to evaluate how well each attribute describes most other people. To provide an estimate of the BTAV effect, difference scores between the first and the second part was computed for each attribute. A high internal consistency emerged among these difference scores (Cronbach's α = .79; Brown, 2012). A total score was computed by summing up each of the difference scores. Previous studies demonstrated that BTAV effects were correlated with SE and depression scores.

Self-Assessment of Intelligence (SAI). In accordance with Furnham and Buchanan's (2005) procedure, participants were instructed to evaluate their intelligence, providing an estimate of their IQ score in 14 different content areas (i.e., general, verbal, logical-mathematical, spatial, musical, kinesthetic, interpersonal, intrapersonal, naturalistic, creative, existential, spiritual, emotive, and practical). Previous studies (Costantini et al., 2016) showed that these 14 items have a high level of internal consistency (Cronbach's α= .84), and one-factor dimensionality. Moreover, global score of the SAI showed a significant correlation with SE (as measured with the RSES), and with a series of measures linked to Positive Orientation (i.e., a Positivity IAT, a Positivity self-report scale, the Positive Orientation Adjectives scale, and the SWL), confirming that SAI scores depend, at least partially, on self-enhancement tendencies.

Results

Descriptive Statistics and Reliability

Both SE-RRT and RSE-RRT showed mean latencies ($ML_{\text{SE-RRT}}$ = 1773 ms and $ML_{\text{RSE-RRT}}$ = 1184 ms respectively) and error percentages ($EP_{\text{SE-RRT}}$ = 11.76% and $EP_{\text{RSE-RRT}}$ = 8.62% respectively) higher than SE-IAT ones ($ML_{\text{SE-IAT}}$ = 918 ms; $EP_{\text{SE-IAT}}$ = 6.73%). This is in line with results reported in the literature about RRT (e.g., De Houwer et al., 2015; ML = 1551; EP = 11.1%). Moreover, SE-RRT showed higher levels of mean latencies and error percentages if compared with the RSE-RRT, indicating that the former may be more difficult than the latter. Consistently with this hypothesis, as mentioned before, sentence-stimuli used for the SE-RRT are a bit more complex than those of RSE-RRT. At the same time, it is worth noting that the SE-RRT was administered after the RSE-RRT for all participants, and therefore we cannot exclude that the higher mean latency of the former may be due, at least partially, to fatigue effects and not to the different stimuli used for the two measures. However, in the literature on this kind of experimental tasks (e.g., the IAT),it was showed that learning effects, in repeated measures designs, tend to be rather large, with empirical results going in the opposite direction of those expected by fatigue effects, and a substantial reduction of mean latencies by practice (Schmukle & Egloff, 2004; Dentale, Vecchione, Ghezzi, & Barbaranelli, 2019). Of the initial 130 participants, only four were excluded for a violation of the D scoring algorithm criteria, confirming that all implicit experimental paradigms were performed appropriately by the most of people. As regards the internal validity of the implicit measures, none of them (i.e., SE-RRT, RSE-RRT and SE-IAT) showed significant correlations with the corresponding mean latency and error percentage (all r's ≤ .06 in absolute value, all p's ≥ .51), except for the significant small relationship found between SE-IAT scores and mean latency (r = .20, p <.05). These results suggest that implicit measures were not influenced (or were only weakly influenced, in the case of the SE-IAT) by possible confounds linked to both speed and accuracy of responses (e.g., cognitive factors, responding styles etc.). Moreover, no significant correlation was found between the order of test-blocks' and implicit measures' scores (all r's ≤ .05 in absolute value, all p's ≥ .63), confirming that they were not influenced by this potential confound.

Table 1 reports descriptive statistics for both implicit and explicit measures included in the study. All measures showed a close to normal distribution, except for Self-Assessed Intelligence (SAI), which exhibited a negatively skewed distribution with a very high level of kurtosis.

An estimation of reliability in terms of internal consistency was provided for all measures. For each implicit measure a Spearman-Brown corrected split-half correlation was computed. In all cases, 95% bootstrapped confidence intervals (based on 5,000 replications) were provided. Values for both RRTs and SE-IAT measures were of acceptable, even if not optimal level (Nunnaly & Bernstein, 1994), and in line with those observed in similar studies (e.g., for the RRT case, see De Houwer et al., 2015; for the SE-IAT case, see Buhrmester, 2011). Finally, reliability coefficients for all self-report measures ranged between .67 and .95.

INSERT TABLE 1 ABOUT HERE

Convergent and criterion validity

In Table 2 (below the main diagonal) were reported zero-order correlations with 95% bootstrapped confidence intervals, that are marked if r coefficient were significantly different from zero, that is when alpha critical level adjusted with false discovery rate procedure (Benjamini & Hochberg, 1995) was exceeded. Results showed that whilst no significant relationships emerged between SE-IAT and RRT measures, the correlation between SE-RRT and RSE-RRT was significant and of small size in terms of Cohen's standards (Cohen, 1988).Moreover, both Relational Responding Tasks showed significant small correlations with self-report measures of self-esteem (RSES and SE-SD), while the SE-IAT did not. In Table 2 (above the main diagonal) were also reported the same correlations disattenuated for the unreliability of the measures. Results showed correlations of moderate size between SE-RRT and RSE-RRT, and also between RRTs and self-report measures of self-esteem (SE-SD and RSES), while the SE-IAT showed weak correlations with all other measures. Overall, these results provided first evidence for RRTs convergent validity.

INSERT TABLE 2 ABOUT HERE

As regards the criterion validity, as illustrated in Table 3 (in even columns were reported zero-order correlations with 95% bootstrapped confidence intervals) the RSE-RRT showed significant and small correlations with theoretically-linked scales measuring life satisfaction (SWLS), optimism(LOT) and depression(CES_D), and also with self-enhancement indicators (i.e., BTAV and SAI), supporting its criterion validity. The SE-RRT showed: a significant and small negative correlation with depression (CES_D), small but not significant correlations with life satisfaction(SWLS) and optimism(LOT), and not significant correlations with self-enhancement indicators(BTAV and SAI). These results support only partially the criterion validity of the SE-RRT. The SE-IAT did not show significant correlations with criteria. Moreover, in Table 3 (odd columns) were also reported the same correlations disattenuated for the unreliability of the measures. Results showed correlations of small/moderate size between RRTs and life satisfaction, optimism and depression, while only RSE-RRT showed correlations of moderate size with self-enhancement indicators (BTAV and SAI). The SE-IAT showed small disattenuated correlations with optimism and depression, and a small correlation with self-assessment of intelligence (SAI).

INSERT TABLE 3 ABOUT HERE

Regarding the proneness to possible confounding factors such as Impression Management (IM) and Self-Deception Enhancement (SDE), as expected, whilst self-report measures of self-esteem (SE-SD and RSES)showed significant and small/moderate correlations with IM ($r = .31$, $p = .001$, bootstrapped CI = .15 – .44 and $r = .21$, $p = .017$, bootstrapped CI = .05 – .36, respectively) and SDE ($r = .56$, $p < .001$, bootstrapped CI = .40 – .68 and $r = .52$, $p < .001$, bootstrapped CI = .38 – .65, respectively),SE-IAT and RRT measures did not. These results suggest that, differently from self-report measures of SE, implicit measures are not affected by these confounding factors.

Incremental validity of Implicit Propositional SE on depression

To investigate the incremental validity of implicit self-esteem beliefs as measured with the RRT, a structural equation modeling approach was used (SEM; see Figure 3) including Implicit Associative Self-Evaluations (IASE; two SE-IAT halves as indicators), Implicit Propositional Self-Esteem (IPSE; based on two first-order factors, IPSE1 and IPSE2, estimated using two test halves for the SE-RRT and two for the RSE-RRT), and Explicit Self-Esteem (ESE; RSES and SE-SD as indicators) as latent predictors and depression as a latent criterion (using three test parcels of the CES_D as indicators). Considering that both residual variances of IPSE1 and IPSE2 and their factor loadings were very similar, we decided to constrain them to be equal. Moreover, considering that residual variances CES_D parcels (i.e., CES_D1, CES_D2, and CES_D3) were very similar, we decided to fix them to be equal. The results showed excellent fit indices for this model ($\chi^2(40) = 24.94$, $p = .97$; CFI = 1.00; TLI = 1.04; RMSEA = .00, 90% CI (.00 –.00); SRMR = .04) with 47% of depression variance explained by predictors. While significant and unique contributions were found for both ESE (Standardized Structural Coefficient = -.37, $p < .05$) and IPSE (Standardized Structural Coefficient = -.38, $p < .05$), a non-significant impact on depression was found for IASE (Standardized Structural Coefficient = -.22, $p < .07$). Interestingly, the IPSE latent factor was significantly and moderately correlated with ESE ($r = .47$, $p < .001$) but was not correlated with IASE. A non-significant correlation also emerged between ESE and IASE. These results provide initial evidence for the incremental validity of IPSE on depression when IASE and ESE are controlled.

INSERT FIGURE 3 ABOUT HERE

Discussion

Study 1 provided initial evidence for the reliability of self-esteem RRTs, which showed acceptable, even if not non-optimal, levels of internal consistency (ranging between .60 and .70), with values close to those usually found for the SE-IAT. The results also showed that RRTs were significantly correlated with each other and with self-report scales of self-esteem (RSES and SE-SD), even if with low effect size, supporting their convergent validity. The size of r coefficients became moderate when the correlations were disattenuated for the unreliability of the measures. In contrast, the SE-IAT did not correlate significantly neither with RRTs nor with explicit SE measures. Moreover, RSE-RRT was significantly correlated with a series of scales theoretically-linked to self-esteem, such as life satisfaction (SWLS), optimism (LOT), depression (CES_D), and two indicators of self-enhancement, like the Better than Average Effect (BTAV) and the Self-Assessment of Intelligence (SAI), providing evidence in favor of its criterion validity. Conversely, the SE-RRT showed a significant negative correlation of small size with depression, small but not significant correlation with life satisfaction and optimism, and not significant correlations with self-enhancement indicators, with an only partial support for its criterion validity. The SE-IAT did not show significant correlations with criteria included in the study. Interestingly, whilst self-report measures of self-esteem showed significant and small/moderate correlations with Impression Management (IM) and Self-Deception Enhancement (SDE), neither RRTs nor the SE-IAT revealed any significant correlation with these confounding factors, confirming the robustness of implicit measures to them.

Finally, using SEM, it was demonstrated that implicit self-esteem beliefs (as measured with RRT measures)showed a unique contribution in the prediction of depression when implicit associative self-evaluations (as measured with the SE-IAT) and explicit self-esteem (as measured with RSES and a semantic-differential scale) were controlled for, supporting their incremental validity.

Study 2

Several studies demonstrated that IAT measures were generally less prone to faking effects than self-report measures, and more predictive of behavioral criteria in situations at risk of impression management biases (Greenwald et al., 2009). More specifically, self-esteem and self-concept IATs were found to be rather robust to faking attempts, if not immune to them (Vecchione, Dentale, Alessandri & Barbaranelli, 2014). For instance, previous experiences with the IAT and/or a detailed description of how to fake it (i.e., providing information about the latency-based nature of this technique) allowed participants to bias their scores (Röhner, Schröder-Abè, Schültz, 2011; Steffens, 2004).

What do we know about the proneness to impression management and faking effects of the RRT? The results of Study 1 showed that RRTs were not significantly correlated with impression management and self-deception enhancement. This seems to support the claim that these measures are not affected by socially desirable responding style. However, the use of full sentences (representing high and low SE beliefs), and the instructions to categorize them as true or false, make the RRT more similar to self-report measures than SE-IAT, facilitating participants to intuitively understand the aims of the task. For these reasons, an experimental study was designed to directly compare the proneness to faking of SE-RRT, SE-IAT and self-report SE measures. A sample of students was recruited and randomly assigned to three different experimental conditions. In the first, participants were invited to perform two implicit measures (i.e., SE-IAT and SE-RRT) and two self-report measures of SE (i.e., RSES and SE-SD) with the instruction to respond as honestly as possible. In the second, both implicit and explicit measures were administered with the instruction to look like people with a high SE, but without technical explanations about the adequate procedure to fake results. In the last condition, participants were invited to appear as high SE participants, with specific instructions on how to do it (see the "Participants and Procedure" section for a detailed description of experimental instructions).

Methods

Participants and Procedure

A sample of 91 undergraduate students (63 females), mean age = 26.03 (SD = 5.18), was recruited for a voluntary participation in an experimental study aimed to compare the effects of three faking conditions (i.e., “no faking”, “faking without instructions”, and “faking with instructions”) on SE-IAT, SE-RRT, RSES and SE-SD scores. Each participant was invited to the laboratory and randomly assigned to one of the three groups. At the beginning of the experimental session, an informed consent form was given to participants to read and sign. In the “no faking conditions”, participants were invited to perform the implicit and explicit SE measures as honestly as possible. By contrast, in both faking conditions participants were invited to perform these measures “as if” they had high SE. Notably, in the “faking without instructions” condition, no instructions were provided to help participants appear to have high SE participants. Differently, in the “faking with instructions” condition, participants were instructed to respond to each self-report item to look like a confident person, and to categorize stimuli as quickly and accurately as possible in compatible blocks (i.e., Self-Positive vs. Others-Negative, for the SE-IAT; High SE sentences as true vs. Low SE sentences as false, for the SE-RRT), and slower and more inaccurately in the incompatible blocks (i.e., Self-Negative vs. Others-Positive, for the IAT; High SE sentences as false vs. Low SE sentences as true, for the SE-RRT).Pre-test measurements were not included because previous studies have shown that implicit latency-based scores can be influenced by precedent experiences with the instrument. The IAT effect tends to decrease from the first to the second administration (Greenwald, Nosek & Banaji, 2003), probably due to learning effects. Moreover, as mentioned before, participants increased their capacity to fake the IAT when they were not naïve to the task (Steffens, 2004).Implicit measures were presented before explicit measures for all participants, as we preferred to observe implicit measures’ mean scores not influenced by performance on precedent instruments. Moreover, the SE-IAT was presented as first for all participants as the SE-RRT is the most difficult task to be performed and, in our opinion, it can work as a training for the Relational Responding Task (reducing the number of invalid results).

Measures

Implicit self-esteem. To measure implicit self-esteem both SE-IAT and SE-RRT were administered in all aspects equivalent to those used in Study 1.

Explicit self-esteem. As in Study 1, RSES and SE-SD were used to measure explicit self-esteem.

Results

Descriptive Statistics

Table 5reports the means and standard deviations of implicit and explicit SE measures for each group ("no faking", "faking without instructions", and "faking with instructions"). No relevant deviations from normality were found for any measure in each condition (skewness and kurtosis values were all between ±1). As shown in Table 4, significant correlations emerged between SE-IAT and SE-RRT (moderate effect size) and between RSES and SE-SD (large effect size). Moreover, both implicit measures correlated significantly and moderately with self-report measures of SE.

INSERT TABLE 4 ABOUT HERE

Testing the Fakeability of SE-RRT compared to SE-IAT and Self-Report Measures of SE

To compare the fakeability of implicit and explicit SE measures, a MANCOVA was conducted, including the faking condition ('no faking', 'faking without instructions', and 'faking with instructions') as an independent variable, the implicit and explicit measures of SE as dependent variables, and sex and age as controls. The results showed a significant multivariate effect of gender (Wilks' lambda = .88, F(8, 166) = 2.86, p <.05, ηp2 = .12) but no significant multivariate effect of age. The multivariate effect of gender is specified by a significant univariate effect on RSES (F(1, 86) = 5.44, p<.05), with lower mean values for females than for males (a result that is consistent with those usually found in the literature). No other significant univariate effects of gender were found. Importantly for the present study, a significant multivariate effect of the faking condition (Wilks' lambda = .31, F(8, 166) = 16.42, p<.001, ηp2 = .44) was found, along with significant univariate effects for both implicit and explicit measures (see Table 5). However, as suggested by the Bayes Factors (BF_{10}), the alternative hypothesis for such effects (i.e., significant differences between faking conditions) was consistently more evident than the null hypothesis for self-report measures of SE (SE-SD and RSES), while for both SE-RRT and SE-IAT such evidence was weak. Moreover, effects' size found for self-report scales were 7/8 times larger (ηp2 =.56 for the SE-SD andηp2 = .64 for the RSES) than those for implicit measures which were identical with each other (ηp2 = .08 for both SE-IAT and SE-RRT). Sidak post-hoc tests suggested that SE-IAT mean scores were lower in the "no faking" condition rather than in the "faking without instructions" (even if not significantly) and in the "faking with instructions" conditions (significantly, p = .03).No considerable and significant mean differences were found between faking with or without instructions groups. In a similar way, SE-RRT mean scores were lower in the "no faking" condition compared to both the "faking without instructions" (p = .05) and the "faking with instructions" conditions (p = .08), even if with not significant mean differences. No considerable and significant mean differences emerged between faking with or without instructions conditions. Finally, both self-report self-esteem (RSES and SE-DS)mean scores were significantly lower in the "no faking condition" compared to the "faking without instructions" and the "faking with instructions" conditions(p < .001), while no significant differences were found between the last two groups.

INSERT TABLE 5 ABOUT HERE

Discussion

Overall, even if SE-IAT and SE-RRT appear not to be immune to faking with significant effects that are identical in terms of effect size, both measures revealed a lower vulnerability to faking effects with respect to self-report measures of SE (the effects sizes found for implicit measures were 7/8 times lower than those for self-report measures). Thus, even if SE-RRT stimuli and instructions are more similar to self-report measures than SE-IAT ones (making it easier to guess how to manipulate test scores), they showed a similar robustness to faking attempts that appear to be remarkably higher than self-report measures.

General Discussion

The present set of studies aimed to apply the RRT to measure SE in a theoretical framework that distinguish implicit associative self-evaluations, implicit SE beliefs and explicit self-esteem. In the first study two self-esteem RRTs, that differed for the complexity of SE statements used as stimuli, were administered along with several other measures revealing the following results: 1) in line with the literature on latency-based implicit measures, internal consistency for both RRTs was acceptable, even if not optimal (.60 < r < .70); 2) initial evidence for the convergent of RRTs were found both in terms of correlations between them and with self-report measures of self-esteem(RSES and SE-DS); 3)criterion validity was supported for the RSE-RRT in terms of correlations with life satisfaction (SWLS), optimism (LOT), depression(CES-D), and also with two indicators of self-enhancement (BTAV and SAI). Criterion validity of the SE-RRT was only partially supported as we found a negative significant correlation with depression, but not with the other criteria; 4) lack of correlations with SDE and IM were found, suggesting robustness of RRTs to these confounds; 5) implicit SE beliefs showed a unique contribution in the prediction of depression even when implicit associative self-evaluations and explicit self-esteem were controlled for, supporting their incremental validity.

In a second experimental study, the faking effect size found for SE-RRT was identical to that of SE-IAT (eta squared = .08 for both measures), and 7/8 times lower than those found for self-report measures of SE, which revealed remarkably large effects (eta squared> .50). These results suggest that the proneness to faking of SE-RRT is similar to that of SE-IAT, and strongly lower than that of both explicit measures of SE.

Notwithstanding these encouraging evidence, some important critical comments regarding all results reported, along with the limitations of the studies can be summarized as follows: 1) As regards the reliability of RRTs, first of all, future studies with a longitudinal design may apply Latent State-Trait models (LST; Steyer, Ferring & Schmitt, 1992) to decompose state, trait and error components of RRTs scores, improving in this way reliability estimates. Moreover, the acceptable (.60 < rtt < .70) but not optimal internal consistency found may be improved by increasing the number of trials for each test block or, in alternative, by adding further test-blocks in the task (Nunnally & Bernstein, 1994). Notably, in doing this, it would be important to avoid effects linked to fatigue that may introduce further inter-individual differences not linked to the attribute to be measured. In this view, it may be interesting the idea to increase the number of blocks (and not simply increase the number of trials) in order to give the subjects more opportunities to recover mental resources during the task.

2) As regards the convergent validity, even if significant correlations emerged between RRTs and between RRTs and self-report measures of SE (SE-SD and RSES), the relationships found were low in terms of effect size. The low size of these inter-correlations were probably due (at least partially) to the suboptimal reliability that emerged for the new implicit measures. Indeed, as reported in Table 2, when these correlations were disattenuated for measures of reliability, the r coefficients become moderate in size, confirming the role of measurement error in decreasing the inter-correlations emerged. In line with the theoretical distinction between implicit propositional beliefs of SE and automatic associations linked to SE (see the introduction for a large discussion of this point), the non-significant correlations found between RRTs and SE-IAT confirm that these measures do not refer to the same construct. In future studies, it may be important to test the convergent validity of RRTs with respect to other implicit measures not based on latency-based tasks, such as the Name Letter Task (Hoorens, 1990).

3) As regards the criterion validity, it is worth noting that, even if the RSE-RRT is less refined than the SE-RRT from a methodological point of view (i.e., high vs. low self-esteem sentences are not balanced for the number of positive vs. negative words), it showed a larger number of

significant correlations with theoretically-linked criteria. This is probably due to the higher simplicity of RSE-RRT sentence-stimuli with respect to SE-RRT ones, demonstrated by both lower mean latencies ($ML_{RSE\text{-}RRT}$ = 1184 ms and $ML_{SE\text{-}RRT}$ = 1773 ms respectively) and lower error percentages ($EP_{RSE\text{-}RRT}$ = 8.62% and $EP_{SE\text{-}RRT}$ = 11.76% respectively). Indeed, the higher easiness in performing the RSE-RRT may have decreased the probability that measures' scores were influenced by potential confounding factors, such as cognitive abilities (e.g., speed in recognizing stimuli words, in task switching capacity, in producing motor responses, etc.), emotional and motivational reactions to the task (e.g., fatigue, emotional reaction to error, etc.) and responding styles (e.g., a more conservative or a more liberal attitude in responding to a latency-based task as the RRT, etc.). Regarding the latter factor in particular, the higher complexity of SE-RRT sentence-stimuli may have led participants to be excessively conservative in performing both compatible and incompatible blocks, introducing a source of systematic variance in SE-RRT scores not due to the attribute to be measured. The weak and non-significant correlations between the SE-IAT and the other theoretically-linked constructs are in line with results of the literature, and may depend on both the associative nature of the Implicit Association Test (not suited for measuring self-esteem beliefs) and on its acceptable but not optimal reliability.

4) From a theoretical point of view, the distinction among automatic SE associations, implicit SE beliefs and explicit SE beliefs, proposed in the literature (e.g., De Houwer, 2014), appeared to be useful and deserving of further investigation. Indeed, implicit SE beliefs showed unique contributions in the prediction of depression even if implicit associative self-evaluations and explicit self-esteem were controlled for. Interestingly, results of the present studies are not in accordance with traditional dual process models (e.g., Strack & Deutsch, 2004), that conceive implicit self-esteem as strictly linked to automatic SE associations, and provide initial evidence for a relational conception of implicit SE in accordance with a behavioral analytic (i.e., functional contextualism) approach to language and cognition like the RFT (Hayes, Barnes-Holmes, & Roche, 2001).In order to provide further evidence for a relational conception of implicit SE, other studies

may be conducted, applying a multi-trait multi-method approach (Nosek & Smyth, 2007)on different research domains (e.g., self-esteem, personality traits, cultural stereotypes, politician attitudes, etc.), to provide more general evidence for the distinction among automatic associations, implicit beliefs and explicit beliefs, separating method (i.e., IATs, RRTs and self-report scales) and substantive factors (i.e., automatic associations, implicit beliefs and explicit beliefs).

5) Finally, the lack of correlations between RRTs and Impression Management(IM) and self-deception enhancement (SDE) scales, along with the robustness to faking effects of the SE-RRT showed in study 2, provide further support for the validity of the new measures also when personally or socially sensitive issues are investigated. In particular, even if the instruction to categorize a series of SE statements as true or false, make the SE-RRT more similar to traditional self-report measures than the SE-IAT, the results of study 2 showed similar vulnerability to faking effects for SE-RRT and SE-IAT.A limitation of this research issue is that, even if no differences are expected between SE-RRT and RSE-RRT in terms of proneness to faking effects, the vulnerability of the latter to this kind of distortion remains untested.

Overall, results of the present research supported the use of the RRT to measure SE, indicating that the RSE-RRT, even if less methodologically refined in terms of balance between high vs. low self-esteem sentences, is better than the SE-RRT at least for the criterion validity. In this view, we recommend to use preferentially the RSE-RRT as it was easier to be performed, and thus probably less prone to confounding factors, such as cognitive ability, emotional/motivational reactions in performing the task and responding styles. The evidence found encourage to conduct further studies to confirm the distinction among automatic SE associations, implicit SE beliefs and explicit SE beliefs. In this vein, it may be important to design: 1) theoretical studies that deeply discuss the differences among implicit associative, implicit propositional and explicit self-esteem, comparing cognitive psychology and functional contextualism approach in order to provide an adequate definition of automatic self-evaluations; 2) multi-trait multi-method studies on different research domains to support also empirically the distinction among automatic associations, implicit beliefs and explicit beliefs;3) experimental studies designed to test whether levels of implicit SE beliefs vary as a function of stimuli which are expected to influence SE (e.g., stimuli that threaten SE; see Rudman, Dohn, & Fairchild, 2007); 4) longitudinal studies devoted to decompose stability, occasion specificity and error variability of the RRT scores (LST models; Steyer, Ferring & Schmitt, 1992), and also to test the effect of IPSE on depression controlling for baseline levels of depression (e.g., RI-CLPM; Hamaker, Kuiper & Grasman, 2015); 5) studies that include clinical populations (e.g., depressed patients) in order to test the capacity of the RRT to discriminate between healthy and psychopathological individuals, as well as among different groups of psychopathologies; 6) studies testing the vulnerability to faking effects also for the RSE-RRT; 7) As previously investigated with the SE-IRAP (Remue et al., 2014), studies aimed to develop RRT measures of ideal self-esteem, with the possibility to test the predictive validity of actual vs. ideal SE discrepancy on depression; 8) As previously investigated, for instance with the Task Switching Ability IAT (TSA-IAT; Back, Schmukle & Egloff, 2005), studies aimed to test divergent validity of the RRT using neutral sentence-stimuli (unrelated to self-esteem) in order to exclude that the correlations with indices of self-esteem and depression would be explained by general intellectual and learning abilities related to RRT performance.

Compliance with Ethical Standards

Conflict of interests: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical approval: “All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee (Ethical Committee of the Department of Dynamic and Clinical Psychology, Sapienza University of Rome, n. 8/2018) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.”

Informed consent: “Informed consent was obtained from all individual participants included in the study.”

Availability of Data and Materials

The data of both study 1 and study 2 were uploaded as electronic supplementary materials along with the other files.

References

Back, M. D., Schmukle, S. C., & Egloff, B. (2005). Measuring Task-Switching Ability in the Implicit Association Test. *Experimental Psychology, 52*(3), 167-179.

Barnes-Holmes, D., Barnes-Holmes, Y., Power, P., Hayden, E., Milne, R., et al. (2006). Do you really know what you believe? Developing the Implicit Relational Assessment Procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist*, 32, 169–177.

Brown, J.D. (2012). Understanding the better than average effect: Motives (still) matter Personality and Social. *Psychology Bulletin*, 38 (2), 209-219. DOI: 10.1177/0146167211432763.

Benjamini, Y. and Hochberg Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal *of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, No. 1, pp. 289-300.

Buhrmester, M. D., Blanton, H. & Swann, W. (2011). Implicit self-esteem: Nature, measurement, and a new way forward. *Journal of Personality and Social Psychology*, *100*, 365–385. doi: 10.1037/a0021341.

Cai, H., Sedikides, C., Gaertner, L., Wang, C., Carvallo, M., Xu, Y., … Eckstein-Jackson, L. (2011). Tactical self-enhancement in China: Is modesty at the service of self-enhancement in East-Asian culture? *Social Psychological and Personality Science, 2*, 59–64. DOI: 10.1177/1948550610376599.

Chamorro ‐ Premuzic, T. & Furnham, A. (2006). Self ‐ Assessed Intelligence and Academic Performance.

Educational Psychology: An International Journal of Experimental Educational Psychology, 26:6, 769-779. http://dx.doi.org/10.1080/01443410500390921.

Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic

Costantini, G., Perugini, M., Dentale, F., Barbaranelli, C., Alessandri, G., Vecchione, M., & Caprara, G. V. (2016). Assessing Positive Orientation With the Implicit Association Test. *European Journal of*

Psychological Assessment. Advance online publication. http://dx.doi.org/10.1027/1015-5759/a000362.

De Houwer, J (2014). A Propositional Model of Implicit Evaluation. *Social and Personality Psychology Compass*. 8 (7), 342-353. https://doi.org/10.1111/spc3.12111.

De Houwer, J., Heider, N., Spruyt, A., Roets, A. and Hughes, S. (2015). The relational responding task: toward a new implicit measure of beliefs. *Frontiers in Psychology. 6: 319.*https://doi.org/10.3389/fpsyg.2015.00319.

De Raedt R., Schacht R., Franck E., De Houwer J. (2006) Self-esteem and depression revisited: Implicit positive self-esteem in depressed patients? Behaviour Research and Therapy, 44(7), 1017–1028. doi: 10.1016/j.brat.2005.08.003.

Dentale, F., San Martini, P., De Coro, A. & Di Pomponio, I. (2010). Alexithymia increases the discordance between implicit and explicit self-esteem. *Personality and Individual Differences*, *49*, 762–767. doi: 10.1016/j.paid.2010.06.022.

Dentale, F., Vecchione, M., De Coro, A. & Barbaranelli, C. (2012). On the relationship between implicit and explicit self-esteem: The moderating role of dismissing attachment. *Personality and Individual differences*, *52*,173–177. doi: 10.1016/j.paid.2011.10.009.

Dentale, F., Vecchione, M., Ghezzi, V., & Barbaranelli, C. (2019). Applying the latent state-trait analysis to decompose state, trait, and error components of the Self-Esteem Implicit Association Test. *European Journal of Psychological Assessment*, *35*(1), 78-85.

Dewitte, M., De Schryver, M., Heider, N., De Houwer, J. (2017). The Actual and Ideal Sexual Self Concept in the Context of Genital Pain Using Implicit and Explicit Measures. Journal of Sexual Medicine May, 14(5):702-714. doi: 10.1016/j.jsxm.2017.03.246.

Di Fabio, A. & Ghizzani F. (2007). La soddisfazione di vita in un campione di apprendisti maggiorenni: alcuni correlati e predittori. *Gipo - Giornale Italiano Di Psicologia Dell'Orientamento*, 8, 3-11.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, *49*(1), 71-75. http://dx.doi.org/10.1207/s15327752jpa4901_13.

Falk, C. F., Heine, S. J., Takemura, K., Zhang, C. X. J., Hsu, C. (2013). Are implicit self-esteem measures valid for assessing individual and cultural differences? *Journal of Personality*. Advance online publication. doi:10.1111/jopy.12082

Falk, C. F., & Heine, S. J. (2015). What is implicit self-esteem, and does it vary across cultures?*Personality and Social Psychological Review, 19*, 177-198.

Fava, G.A. (1983). Assessing depressive symptoms across cultures: Italian validation of the CES-D self-rating scale. *Journal of Clinical Psychology, 39*, 249–251.

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. https://doi.org/10.1037/0033-2909.132.5.692.

Gawronski, B., & Payne, B. K. (2010). *Handbook of implicit social cognition: Measurement, theory, and applications*. New York, NY: Guilford Press.

Giannini, M., & Di Fabio, A.M. (2008). Misurare l’ottimismo: proprietà psicometriche della versione italiana del Life Orientation Test-Revised (LOT-R). *Counseling,* 1, 73-83.

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. doi: 10.1037/0022-3514.79.6.1022.

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

Greenwald, A. G., 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*(2), 197-216. http://dx.doi.org/10.1037/0022-3514.85.2.197.

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, 20(1), 102-116. doi: 10.1037/a0038889.

Hayes, S.C., Barnes-Holmes D., Roche B. (2001). Relational frame theory: A post-Skinnerian account of human language and cognition. New York, NY: Plenum Press.

Hintzmann, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review,* 93, 411– 428.

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 measure. *Personality and Social Psychology Bulletin*, 31, 1369–1385. doi: 10.1177/0146167205275613.

Hoorens, V. (1990). Nuttin's affective self-particles hypothesis and the name letter effect: A review. *Psychologica Belgica*, 30, 23–48.

Ingram, R.E., Miranda, J., Segal, Z.V. (1998). Cognitive Vulnerability to Depression. New York: Guilford Press (330pp). Clinical Psychology & Psychotherapy, 6(1), 69–69. doi: 10.1002/(sici)1099-0879(199902)6:1,69::aid-cpp177.3.0.co;2-w.

Koning I.M., Spruyt A., Doornwaard S.M., Turrisi R., Heider N. & De Houwer J. (2016): A different view on parenting: automatic and explicit parenting cognitions in adolescents' drinking behavior, Journal of Substance Use, doi: 10.1080/14659891.2016.1217088.

Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology, 35*(2), 63-78. http://dx.doi.org/10.1037/0022-3514.35.2.63 .

Nosek, B. A., & Smyth, F. L. (2007). A multitrait-multimethod validation of the Implicit Association Test: Implicit and explicit attitudes are related but distinct constructs. *Experimental Psychology*, 54, 14–29. https://doi.org/10.1027/1618-3169.54.1.14.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGrawHill.

Paulhus, D. L. (1991). *Measurement and control of response bias*. In J. P. Robinson, P. R. Shaver & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.

Prezza, M., Trombaccia, F.R., & Armento, L. (1997). La scala dell'autostima di Rosenberg. Traduzione e validazione italiana. *Bollettino di Psicologia applicata,223,* 35-44.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385–401. doi:10.1177/014662167700100306.

Remue, J., De Houwer, J., Barnes-Holmes, D., Vanderhasselt, M.A., De Raedt, R. (2013). Self-esteem revisited: performance on the implicit relational assessment procedure as a measure of self- versus ideal self-related cognitions in dysphoria. *Cognition and Emotion*, 27, 1441–1449. doi: 10.1080/02699931.2013.786681.

Remue J, Hughes S, De Houwer J, De Raedt R (2014). To Be or Want to Be: Disentangling the Role of Actual versus Ideal Self in Implicit Self-Esteem. *PLOSE ONE,* 9(9): e108837.https://doi.org/10.1371/journal.pone.0108837

Röhner, J., Schröder-Abé, M., & Schütz, A. (2011). Exaggeration is harder than understatement, but practice makes perfect! Faking success in the IAT. *Experimental Psychology*, 58, 464-472. doi:10.1027/1618-3169/a000114.

Rudman, L. A., Dohn, M. C., & Fairchild, K. (2007). Implicit self-esteem compensation: Automatic threat defense. *Journal of Personality and Social Psychology, 93*(5), 798-813. http://dx.doi.org/10.1037/0022-3514.93.5.798.

Scheier, M. F., Carver, C. S. & Bridges, M. W. (1994). Distinguishing optimism from nevroticism (and trait anxiety, self-mastery, and self-esteem): A re-evaluation of the Life Orientation Test. *Journal of Personality and Social Psychology,* 67, 1063-1078.doi: 10.1037//0022-3514.67.6.1063.

Schmitt, D. P., & Allik, J. (2005). Simultaneous Administration of the Rosenberg Self-Esteem Scale in 53 Nations: Exploring the Universal and Culture-Specific Features of Global Self-Esteem. *Journal of Personality and Social Psychology, 89*(4), 623-642. Http://dx.doi.org/10.1037/0022-3514.89.4.623.

Smith, T. W., Pope, M. K., Rhodewalt, F., & Poulton, J. L. (1989). Optimism, neuroticism, coping, and symptom reports: An alternative interpretation of the Life Orientation Test. *Journal of Personality and Social Psychology, 56*(4), 640-648. http://dx.doi.org/10.1037/0022-3514.56.4.640.

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

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. doi/10.1111/1467-9280.00202.

Steffens, M. C. (2004). Is the Implicit Association Test immune to faking? *Experimental Psychology,* 51, 165–179. doi:10.1027/ 1618-3169.51.3.165.

Steyer, R., Ferring, D. & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment, 8*, 79–98.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, 8, 220–247. https://doi.org/10.1207/s15327957pspr0803_1.

Teige-Mocigemba, S., Klauer, K. C., & Sherman, J. W. (2010). A practical guide to implicit association tests and related tasks. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 117-139). New York, NY, US: The Guilford Press.

Tibboel, H., De Houwer, J., Dirix, N., Spruyt, A. (2017). Beyond associations: Do implicit beliefs play a role in smoking addiction? Journal of Psychopharmacology, Jan;31(1):43-53. doi: 10.1177/0269881116665327.

Vecchione, M., Dentale, F., Alessandri, G. & Barbaranelli, C. (2014). Fakability of implicit and explicit measures of the big five: Research findings from organizational settings. *International Journal of Selection and Assessment*, 22, 211–218.https://doi.org/10.1111/ijsa.12070.

Williams, J.M.G. (1997). Depression. In DM Clark, & CG. Fairburn (Eds.), Science and practice of cognitive behaviour therapy (pp. 259–283). Oxford: Oxford University Press.

Figure captions

Fig. 1 caption: Example of actual an SE-IRAP (Adapted from Remue, Hughes, De Houwer, De Raedt, 2014)

Fig. 2a caption: Self-Esteem RRT: compatible combined block

Fig. 2b caption: Self-Esteem RRT: incompatible combined block

Fig. 3 caption: SEM model including Implicit Associative Self-Evaluations (IASE), Implicit Propositional Self-Esteem (IPSE), and Explicit Self-Esteem (ESE) as latent predictors and Depression as latent criterion.

Note. SE-RRT1: Self-Esteem RRT first half; SE-RRT2: Self-Esteem RRT second half; RSE-RRT1: Rosenberg Self-Esteem RRT first half; RSE-RRT2: Rosenberg Self-Esteem RRT second half; IAT1: Self-Esteem IAT first half; IAT2: Self-Esteem IAT second half; SE-SD: Self-Esteem Semantic Differential; RSES: Rosenberg Self-Esteem Scale