

Indirectly measuring evaluations of several attitude objects in relation to a neutral reference point[☆]

C. Miguel Brendl^{a,*}, Arthur B. Markman^b, Claude Messner^c

^a INSEAD, Boulevard de Constance, 77305 Fontainebleau Cedex, France

^b Department of Psychology, University of Texas at Austin, 1 University Station, A8000, Austin, TX 78712-0187, USA

^c Fakultät für Psychologie, Universität Basel, Missionsstrasse 60/62, 4055 Basel, Schweiz, Switzerland

Received 23 June 2003; revised 24 June 2004

Available online 6 November 2004

Abstract

This article presents a new response time measure of evaluations, the Evaluative Movement Assessment (EMA). Two properties are verified for the first time in a response time measure: (a) mapping of multiple attitude objects to a single scale, and (b) centering that scale around a neutral point. Property (a) has implications when self-report and response time measures of attitudes have a low correlation. A study using EMA as an indirect measure revealed a low correlation with self-reported attitudes when the correlation reflected between-subjects differences in preferences for one attitude object to a second. Previously this result may have been interpreted as dissociation between two measures. However, when correlations from the same data reflected within-subject preference rank orders between multiple attitude objects, they were substantial (average $r = .64$). This result suggests that the presence of low correlations between self-report and response time measures in previous studies may be a reflection of methodological aspects of the response time measurement techniques. Property (b) has implications for exploring theoretical questions that require assessment of whether an evaluation is positive or negative (e.g., prejudice), because it allows such classifications in response time measurement to be made for the first time.

© 2004 Elsevier Inc. All rights reserved.

Introduction

Imagine you are an astronomer who uses a new telescope that can see much further into space than traditional telescopes, and, in addition, that it sees through the Earth's atmosphere with little distortion. The signal is acoustic rather than optical. The technology is promising because by looking at new regions of space new discoveries can be made. The catch is that you have to work out what the sounds you are hearing signal and whether there are new kinds of distortions to be dealt

with. With response time measures of attitudes we face a similar problem.

Attitudes have traditionally been measured in two different ways, either by means of self-report, that is, by asking respondents directly for them, or by *indirect measures*, such as response time procedures. Indirect measures infer attitudes from behaviors without directly asking respondents for their attitude. The situation is similar to that of the astronomer because, first, indirect measures of attitudes may be less contaminated by a known distortion (social demand) than self-report measures, but we need to learn about new distortions. Second, we may be able to look at attitudinal constructs that we could not look at with self-report measures. In this respect there has been debate about whether indirect measures tap attitudes that respondents are not aware of, raising the question of what it is they are signaling. One way to find this out is to deter-

[☆] We thank Joachim Vosgerau for his comments on an earlier draft of this article. This research was partly funded by Grant DFG BR1722/1-2 from the German Science Foundation and an INSEAD Research and Development Award, both to C. Miguel Brendl.

* Corresponding author. Fax: +33 1 60745537.

E-mail address: miguel.brendl@insead.edu (C.M. Brendl).

mine whether the measurement signal detects properties that the potentially measured theoretical construct ought to have. A property that unconscious attitudes ought to have is functional independence from conscious attitudes (Kihlstrom, 2004).

Objectives of our research

We extend previous developments of response time measures by introducing a new measure, the Evaluative Movement Assessment (EMA). EMA assesses two properties that attitudes ought to have, but which have not been discussed for response time measures of attitude. First, EMA can *assign multiple attitude objects to a single scale*, retaining their rank order on a good–bad dimension. Second, EMA can *assign a neutral point to that scale* indicating a separation between positive and negative evaluations. Finally, we show that these properties are useful for investigating whether response time measures tap unconscious attitudes.

We assumed that self-reported ratings of words are adequate measures of how respondents evaluate these words, as long as there are no response biases. To demonstrate assignment of multiple attitude objects to a single scale, we tested whether EMA would rank order more than two attitude objects (i.e., words) on a single scale and whether it would place these objects in the same order as self-reported evaluations. To demonstrate neutral point assignment we tested whether EMA would assign the neutral point of that scale in the same region as self-reported evaluations.

The principles underlying EMA

EMA embodies the idea that it is difficult to actively evaluate a stimulus as positive without having an approach response to it or to evaluate a stimulus as negative without having an avoidance response to it. In EMA, the participant's first name is displayed on a computer screen. The name represents the participant. A row of X's appears randomly either to the right or left of the name. The X's are then replaced by a positive or negative attitude word (e.g., sun). By moving a joystick to the right or left, respondents move the attitude word to the right or left, thus creating a perception of moving it toward or away from the first name (see Fig. 1).¹ Based on previous findings (Chen & Bargh, 1999; Solarz, 1960) we assumed that moving the attitude

word toward one's name would be an approach response, while moving away from it would be an avoidance response. If so, then toward movements should be faster than away movements in response to positive words, and away movements should be faster than toward movements in response to negative words. The difference in response time between a toward and an away movement should allow us to infer indirectly the valence of a word: the faster a toward movement (compared to an away movement), the more positive the word; the faster an away movement (compared to a toward movement), the more negative the word.

Relevance of assigning a neutral point to an attitudinal scale

Response time measures are sometimes interpreted as indicating negative evaluations of out-groups or self-positivity biases. Although this practice implies that the measure is anchored around a neutral point, we are not aware of a study having verified this assumption. For example, imagine we are interested in prejudice defined as a negative evaluation of some group. A single attitude score showing a preference for group A over group B does not necessarily mean a negative evaluation of group B (Brendl, Markman, & Messner, 2001), or, as Kihlstrom (2004) put it, if I prefer tiramisu to zabaglione it does not mean that I dislike zabaglione. Even though my liking of these desserts will be relative to other foods serving as reference points, neutrality is a very special reference point (cf. Kahneman, 1999). First, approach responses change into avoidance responses. Second, it seems difficult for context to push objects across to the other side of neutrality. At least for some experiences it is impossible. As Parducci (1995, cited in Kahneman, 1999) put it, there is no context in which cutting oneself shaving will be pleasant.

Relevance of assigning multiple attitude objects to a single attitude scale

Obviously, measures that can assign more than two attitude objects to a single scale are useful when more than two attitude objects are under investigation, for example when Livingston and Brewer (2002) compared White faces, prototypical Black faces, and non-prototypical Black faces. There are two less obvious advantages of this property that we discuss in greater detail. First, it is a prerequisite for calibrating a single measure around a neutral point. Second, it improves our ability to test for functional independence of indirectly measured and self-reported attitudes.

Calibration around a neutral reference point

Any kind of measurement relates an unknown quantity to a known quantity; for example, when measures

¹ Based on previous findings that arm movements are affected by valence (Chen & Bargh, 1995—cited in Bargh, 1997; Chen & Bargh, 1999; Münsterberg, 1892; Solarz, 1960), we first presented a paradigm using forward–backward movements without presenting the participant's name on the screen (Brendl, 1997; June). We are most grateful to Tony Greenwald (personal communication, February 1998), who suggested that we use left–right movements and that we do so by relating the arm movements to a self-related object in space.

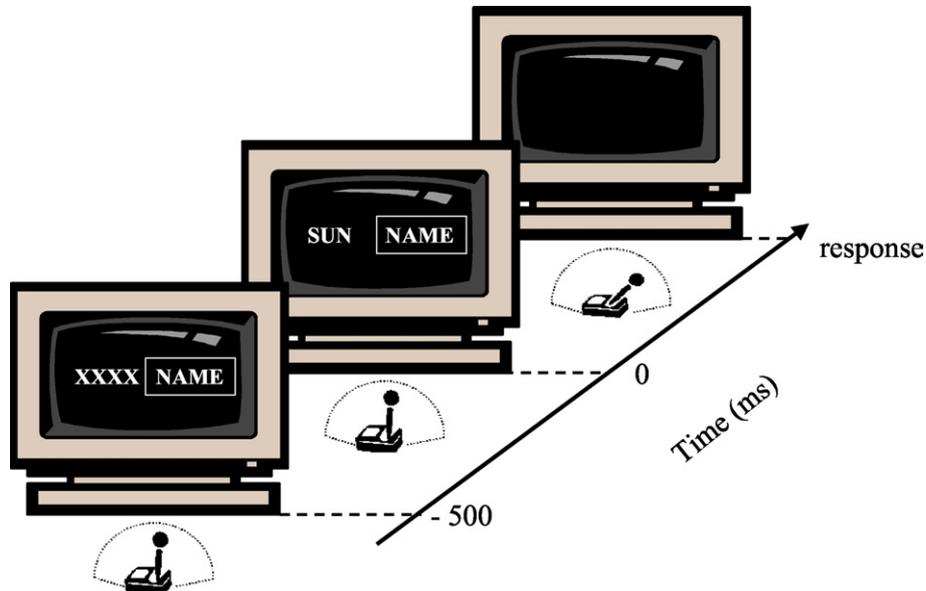


Fig. 1. Sequence of events during one response time trial. In the third picture the joystick is moved to the right.

of length are related to the original meter. Thus, any measurement instrument needs to be calibrated around a reference frame. When respondents self-report attitudes they are expected to do this calibration themselves, but in indirect measurement this is not possible. If an indirect measure of attitudes can assign multiple attitude objects to a single scale, then a subset of attitude objects can be used as a reference frame that is centered around a neutral point and other objects of interest can be related to this frame. Thus, in order to calibrate an indirect measure around a neutral reference point, the measure has to be able to assign multiple attitude objects to one scale.

The relationship between indirectly measured and self-reported attitudes

Let us first review the current understanding of the relationship between attitudes that are indirectly measured by response times and those that are self-reported. Assumptions differ on whether respondents are aware of the attitude itself and whether they can control it. There is agreement that respondents are aware of and have control of self-reported attitudes. However, for indirectly measured attitudes, three possibilities have been discussed. Some articles have argued that response time measures may tap attitudes that people are not aware of (and hence have no control of). We term these attitudes *implicit attitudes* in line with common practice and the meaning of “implicit” in the domain of memory (Fazio & Olson, 2003; Kihlstrom, 2004). Other articles have argued that response time measures tap attitudes that respondents cannot control but may well be aware of (Fazio & Olson, 2003). We call these attitudes *uncontrollable attitudes*. Note that people may be unconscious of

the source of these uncontrollable but conscious attitudes. Finally, a third possibility that has been discussed is that response time measures tap the same controllable attitudes as self-report, but that they are less susceptible to social response biases. We feel that it is important to reserve the term “implicit” attitudes for those attitudes that people are not aware of, first in order to be able to distinguish them from uncontrollable but conscious attitudes, and second to use the term *implicit* consistent with its meaning in cognitive psychology. We believe that Greenwald and Banaji’s (1995) seminal definition of implicit attitudes can be interpreted to subsume both of these types. Hence, our use of the term implicit may deviate from their use.

Implicit attitudes are—by definition—functionally independent of conscious attitudes. Thus, in order to show that a method measures implicit attitudes it is necessary to show that it does not correlate with (or at best shows low correlations with) measures of conscious attitudes (Kihlstrom, 2004). In recent years, the view that indirect measures of attitudes may measure implicit attitudes has gained popularity (e.g., Banaji, Bazerman, & Chugh, 2003). While this view has been applied to various indirect measures, such as response compatibility measures (e.g., the Implicit Association Test or IAT, Greenwald, McGhee, & Schwartz, 1998) and evaluative priming procedures (e.g., the bona fide pipeline technique or BFP, Fazio, Jackson, Dunton, & Williams, 1995; Fazio, Sanbonmatsu, Powell, & Kardes, 1986), word-stem completion, name-letter evaluations, or physiological measures (Fazio & Olson, 2003; for an overview), the evidence that has led to it arises primarily from IAT studies. In these studies, the IAT often shows low correlations with self-report at the level of the individual respondent.

In other words, an individual's IAT-based attitude toward an object is often a very weak predictor of his or her self-reported attitude toward the same object. However, it is important to note that medium to high within-subject correlations between self-report and the IAT are also regularly found.

Our question is not about the construct of implicit attitudes itself, nor about the possibility that certain response time measures may tap these. Instead, we ask whether in the presence of high correlations, the low correlations between response time measures and self-report justify an assumption that response time measures tap implicit attitudes. To support that assumption, low correlations need to exist and these must not be due to purely methodological reasons (e.g., high measurement noise, response biases). When there is social demand, low correlations could always be the result of participants more effectively biasing their responses in self-report than in a response time task, leading to low correlations. This suggests studying low correlations in domains with low social demand.

However, according to a widely held view, correlations are not low in these domains (Greenwald et al., 2002). Yet, the few published studies with low social demand show low and high correlations too, just as domains with high social demand (see Table 1). What is of interest to us is the large variability in correlations rather than the mean of the correlations. Further, the data pattern suggests that certain attitude domains consistently show low correlations (e.g., preferences of flowers to insects), whereas others may show higher correlations (e.g., preferences of voting for candidate A than B). Of the nine correlations in Table 1 tapping attitudes toward flowers and insects, eight range between $-.19$ and $.16$. In sum, in domains with low social demand we find variability in correlations, and we can rule out that this variability is due to social demand. We therefore conducted our studies in this domain.

Variability in correlations between self-reported and indirectly measured attitudes: Insensitivity to within-subject variability as a potential methodological reason

Studies of implicit attitudes have typically assessed individual differences between attitudes toward one object and another (e.g., male vs. female, or Coke vs. Pepsi). In this section we explain that this research design may not detect associations between two attitude measures under conditions when attitudes vary substantially within-subjects but vary much less between-subjects. Thus, under these conditions the structure of this research design may actually cause the low correlations between measures.

To illustrate this methodological problem we use an IAT study with vegetarians and omnivores as participants (Swanson, Rudman, & Greenwald, 2001). This

study measured preferences between white meat and other protein foods with the IAT and self-report tools (e.g., thermometer scales), and found a high correlation ($r = .54$) across both samples but a much smaller correlation for the vegetarians alone ($r = .28$) (Table 1). According to an implicit attitudes interpretation of these findings, some vegetarians—who presumably have a conscious preference for vegetables over meat—have an unconscious preference for meat, that is, preferences on these measures are reversed.

Our interpretation suggests precisely the opposite—that the low correlation for vegetarians is the result of a high level of agreement among vegetarians on both measures that vegetables are preferable to meat. The IAT effect expresses the relative preference of one attitude object (e.g., vegetables) to a second (e.g., meat). It does not reflect attitudes toward each object independently. Hence, a low correlation between IAT effects and ratings means that the *degree* to which a person prefers, say, vegetables to meat in an IAT does not predict the *degree* to which he or she prefers vegetables to meat in self-report. Finding a correlation between two measures presupposes that the data contain systematic variability. Among vegetarians there may be very little systematic variability in the degree to which they prefer vegetables to meat, leading to the low correlation. Most importantly, we conclude that low correlations do not necessarily signify preference reversals.

When the authors included omnivores in the analysis sample, there was likely to be considerably more systematic variability than when they included vegetarians alone, as some subjects preferred meat to vegetables and others preferred vegetables to meat. When the authors included omnivores, the systematic between-subjects variability was increased because preference orderings between meat and vegetables were reversed across subjects.

All the systematic variability that we have considered so far is between-subjects. Within-subject variability in this data set consists of a single person preferring vegetables to meat or vice versa. If most vegetarians preferred vegetables to meat on an indirect measure and on a self-report measure, then the two measures agree based on within-subject variability, even if the degrees of preferences do not correlate. A disadvantage of this type of analysis is that it has low measurement power because it only captures the rank ordering of two levels of magnitude. An attitude domain may contain more than two levels of meaningful within-subject variation, for example, a vegetarian may prefer broccoli to soy sausage to cheese to steak. Research designs that assess only the preference of one attitude object over another will omit much of this systematic variation in the attitude domain and are therefore low power designs for estimating systematic within-subject variability.

Table 1
Published correlations between IAT effects and rating scales

Study	Description	N	IAT—self-report correlations (<i>r</i>)
Brunel et al. (in press, Study 1)	IAT with IBM vs. Apple computer brands and semantic differentials	54	.50
Greenwald et al. (1998, Study 1)	IAT with flowers vs. insects; ratings were:		
	• thermometer scales	32	.13
	• semantic differentials	32	.12
	IAT with instruments vs. weapons; ratings were:		
• thermometer scales	32	.29	
• semantic differentials	32	.19	
Karpinski and Hilton (2001, Study 1a)	IAT with flowers vs. insects; semantic differentials of:		
	• category labels (e.g., flowers)	42	.11
Karpinski and Hilton (2001, Study 1b)	IAT with flowers vs. insects; thermometer scales of:		
	• category labels (e.g., flower)	28	–.31
Karpinski and Hilton (2001, Study 2)	IAT with apple vs. candy bar; ratings were:		
	• thermometer scales	40	.16
	• semantic differentials of “apples/candy bars”	40	–.09
	• semantic differentials of “like eating apples/candy bars”	40	–.10
Maison et al. (2001, Study 1)	IAT with sodas vs. juices and semantic differentials	71	.38
Maison et al. (2001, Study 2)	IAT with low vs. high calorie foods and semantic differentials	51	.72
Maison et al. (in press, Study 1)	IAT with Danone vs. Bakoma yoghurt brands and semantic differentials	32	.47
Maison et al. (in press, Study 2)	IAT with McDonald’s and Milk Bar fast food brands and semantic differentials	20	.35
Nosek et al. (2002)	IAT with voting preferences in US presidential elections and rating scale	36840	.52
Olson and Fazio (2001, Study 2)	IAT of conditioned Pokemon characters with semantic differentials as rating scales. A correlation of $r = .07$ was obtained when the ratings preceded the IAT, but one of $r = .54$ when this order was reversed. In both cases conditioning affected means of ratings and of the IAT (personal communication, Michael Olson, 20/12/01)	26	.07
Olson and Fazio (2004, Study 3) ^a	IAT with items representing apples vs. candy bars and		
	• thermometer scales	NA	.11 ^b /.57 ^c
Olson and Fazio (2004, Study 4) ^a	• semantic differentials	NA	.01 ^b /.42 ^c
	IAT with voting preferences in US presidential elections and:		
• thermometer scales	NA	.56 ^b /.75 ^c	
• semantic differentials	NA	.50 ^b /.77 ^c	
Swanson et al. (2001, Study 2)	IAT with white meat vs. other proteins		
	All subjects (vegetarians and omnivores); ratings were:		
	• thermometer scales	101–107	.54
	• semantic differentials	101–107	.51
Only vegetarians; ratings were:			
• thermometer scales	32–34	.28	
• semantic differentials	32–34	.31	

Note. Studies that measure attitudes toward attitude objects (i.e., IATs measuring associations with positive valence and negative valence) in domains that have been or can be considered to involve minimal social demand. IATs measuring associations with the self-concept are not included because these are not considered to measure evaluations of attitude objects. Also, only correlation coefficients are included based on self-report preceding IATs. If IATs precede self-report, correlations could be artificially high due to self-perception during the IAT affecting subsequent ratings, as has been observed (cf. table entry for Olson and Fazio, 2001). There could also be an influence of self-report on IATs; however, we felt that the opposite influence is theoretically more problematic and also practically more likely. The standard practice is to have IATs precede self-report.

^a IAT scores are based on a modified scoring algorithm (cf. Olson and Fazio, 2004).

^b IAT with traditional instructions.

^c IAT with modified instructions that resulted in increased correlations.

More generally, we argue that before a low correlation between an indirectly measured and self-reported attitude is interpreted as evidence that the indirectly measured attitude is implicit, we must be confident that the low correlation does not reflect the research design

failing to detect systematic variability (either between-subjects or within-subject). The two kinds of variability contain psychologically different information and either one may be important depending on the research question. If the claim is, however, that there is no (or little)

shared variability between two measures, all potential variability needs to be considered.

Using EMA we will demonstrate empirically that research designs that assign more than two attitude objects to a scale rather than exactly two are more powerful for capturing within-subject variability. We will elaborate in the General Discussion on whether the most widely used response time measures could be employed in such a research design.

Study 1: Assigning multiple attitude objects to one scale

To create a response time measure that can assign multiple attitude objects to a single scale, the measure requires a measurement property that we call *single item sensitivity*. A single item is any type of single stimulus (e.g., the word “lentil”) that is presented to the respondent during a trial. The attitude score that the measure assigns to a single item must capture unique properties of that item. Attitude scores must meaningfully reflect evaluations of more than two single words presented within one task, even if these words represent different attitude objects. It follows that single item sensitivity goes beyond being sensitive to a single attitude object.

Study 1 introduces the novel response competition task described earlier, a precursor of EMA. Making right-left movements with a joystick, respondents moved each attitude word toward or away from their names depending on the valence of the word. In Study 2, the words actually moved on the screen. In Study 1, they remained stationary, but we use “move words toward/away from their names” as shorthand for “moved the joystick toward/away from their names.” One purpose of Study 1 was to validate that these movements (even if only the hand moved) would represent approach and avoidance behaviors respectively. The main purpose was to test whether this type of task would be sufficiently sensitive to measure the valence of single attitude objects and hence exhibit single item sensitivity. To that end we assessed whether this task could capture evaluations of 92 different attitude objects within a single task.

Attitude objects were words pre-tested to range from very negative to very positive. In one block of trials, participants moved the attitude word toward their name if they considered it positive (e.g., summer) and away if they considered it negative (e.g., cancer). In another block, they moved the attitude word away from their name if positive and towards it if negative. We assumed that because respondents had the task of categorizing attitude words according to their valence, these words would represent attitude objects. We do not make assumptions about the degree to which the task measures enduring attitudes versus spontaneously constructed evaluations of attitude objects. Respondents

also self-reported, that is, rated the valence of the 92 attitude words on thermometer scales.

We predicted that attitude words that respondents rated as positive would facilitate toward movements (word toward name) and that attitude words rated as negative would facilitate away movements (word away from name).² This prediction is based on our hypothesis that the response time task measures uncontrollable rather than implicit evaluations. This hypothesis is falsifiable by finding no association between movement direction and self-report. If people are faster to make toward movements than away movements for positively rated attitude words and faster to make away movements than toward movements for negatively rated attitude words, this task is sensitive to the “conscious” valence of the attitude words. However, it may lump all attitude words into two groups, one positive and one negative, either lacking single item sensitivity, or sensitivity to more than two levels of valence.

We can also explore this issue. The response time task can estimate the magnitude of the valence for each word by the size of the response time difference for toward and away movements. If this response time difference correlates with the level of valence inferred from ratings for each respondent individually, then the response time task is sensitive to magnitude of valence as well as to single items (i.e., here single attitude words).

Note that this response time measure is indirect because we infer the respondent’s evaluation of a word as positive whenever a toward movement is faster than an away movement and as negative whenever an away movement is faster than a toward movement, regardless of whether the respondents explicitly meant their particular movements to express positivity or negativity.

Method

Overview and design

Each participant first completed the response time task that indirectly measured evaluations of a set of positive and negative attitude words. Using a joystick, half the participants were first instructed to move each word they saw as quickly as possible toward their first name if they considered it positive, and away if they considered it negative. Subsequently, they saw the same set of

² The prediction implies that participants represent their name as stationary and the stimulus word (target or distractor) as moving, even though, as pointed out above, in Study 1 words did not move on the screen. Our task instructions were consistent with this assumption. If participants represented their name as moving and the stimulus word as stationary, our predictions would reverse. Thus, by presenting all words as stationary in Study 1, we made it an empirical question which mental representation would be the more natural in this task. In Study 2 stimulus words actually moved on the screen, forcing all participants to mentally represent stimulus words as moving and their name as stationary.

relevance for the German student population (see Fig. 2—bottom panel—for the full set of words).

Participants

University of Constance students were recruited to attend two laboratory sessions, one week apart. Because of the verbal stimulus materials, we tried to recruit only native speakers of German. Further, due to the reaction time task we recruited only participants who had either normal vision or could bring their corrective glasses. Those who did not meet these conditions but were nonetheless recruited were excluded a priori from the data analyses as were participants with dyslexia. Of the 79 participants meeting the above criteria, a number were excluded due to technical equipment errors ($N = 5$), experimenter error ($N = 5$), attending only one of the two sessions ($N = 6$), or failing to follow instructions ($N = 2$). Of these 61, 11 had to be eliminated for reasons described below, leaving 50 (20 females and 30 males) whose data were analyzed.

Procedure

Overview. Each participant was seated alone in a room in front of a computer with a joystick. Session 1 started with the reaction time task followed by a rating of 46 of the 92 attitude words. A week later, in Session 2, participants conducted a different reaction time task followed by the rating of the remaining 46 attitude words, the order of ratings being counterbalanced between-subjects.

Reaction time task. There were six blocks of trials: four practice blocks (each with 24 trials) and two main blocks (each with 96 trials). Each trial began by displaying the participant's first name in a rectangular frame (see Fig. 1) in a white font centered vertically either slightly to the right or the left of the center of the screen. Whether the name appeared on the left or the right was determined randomly in each trial. Simultaneously, on the same line on the opposite side of the center of the screen, a string of four red X's was displayed. After 500 milliseconds (ms), the X's were replaced by an attitude word in red (e.g., "summer"), selected randomly from the set of 92 words.

The participant's task was to move the joystick quickly either to the right or the left. When the joystick was moved about three-fourths of its possible distance sideways, the screen was erased (see Fig. 1). Recall that in this experiment the attitude words did not actually move on the screen (see Footnote ²). Response times were measured from the time the attitude word appeared on the screen until the joystick had been moved a small distance. A trial ended with an error message when no response was made within 3000 ms. The inter-trial interval was 600 ms.

Below, we describe the sequence of events for participants who began the task by "moving" positive attitude

words toward their name and negative attitude words away from their name. In the *first practice block*, participants were asked to move the joystick as quickly as possible (without mistakes) toward their name when they saw the word "good" and away from their name when they saw the word "bad." Each word was presented 12 times. In the *second practice block*, the words "good" and "bad" were replaced by words pre-tested as unambiguously positive ($N = 12$) or negative ($N = 12$). Participants decided whether they found a given word positive or negative and moved the joystick accordingly (again, positive toward, negative away). In the *first main block*, the 92 attitude words from Fazio et al. (1986) were each presented once in random order. Following Greenwald et al. (1998), to reduce the typically high variability at the beginning of a block, the main block began with two additional filler words. These were not included in the analyses. After half of the trials, participants were offered a break, followed by two filler words and the remaining attitude words. This concludes blocks 1–3.

Next were blocks 4–6. These were the same as blocks 1–3 but with reverse instructions for "moving the words": negative words toward the respondent's name and positive words away from it. Block 6 constituted the second main block. We were interested in the response times to the 92 attitude words in each of the two main blocks (i.e., 184 trials). Recall that half of the participants started with blocks 4–6 followed by blocks 1–3.

Rating task. Participants rated each of the 92 attitude words on thermometer scales. Half of the participants first rated each word on unipolar thermometer scales ranging from "not warm (+10)" to "very warm (+30)." They then rated the same words in the same order on unipolar thermometer scales ranging from "cold (−10)" to "not cold (+10)." These numbers correspond to the Celsius temperature scale. The other half of the participants completed the ratings in the reverse order. Participants were instructed to rate their feelings about the words and to concentrate either on their warm feelings only, ignoring their cold feelings, or vice versa. People are surprisingly comfortable with this standard procedure used in research on attitude ambivalence (Kaplan, 1972, p. 365 f.). Our use of different numeric anchors for the two scales helped to avoid unwanted comparisons between the two scales. Word order was counter-balanced between-subjects.

Results

A posteriori data exclusions

Because in each of the two main blocks of the reaction time task participants evaluated each of the 92 attitude objects once, each attitude word should have been moved once toward and once away. This would allow us to compute a difference score in response time for the

two movements. However, sometimes a word was moved twice in the same direction. We had to discard these trials because we could not take difference scores for them. This data loss is undesirable and will be fixed by the EMA procedure introduced in Study 2. We excluded all participants for whom 40% or more attitude objects had to be discarded ($N = 5$). Our reasoning was that because they evidently did not perform the task according to instructions, these participants were mentally performing another task, putting into question the meaning of their responses. Presumably they were either not motivated or able to follow the instructions. Following common practice, we also excluded trials with response times of less than 300 ms and more than 3000 ms, trials in which the subject initiated a joystick movement to one side and then changed it to the other, and trials with a technical data logging error. All participants for whom eight or more trials (i.e., above 4.3%) of the 184 target trials were discarded were excluded from the analyses ($N = 6$). Of the 9200 trials of the 50 analyzed participants, 9053 trials were valid. Of the 4600 possible ratings, 4566 were completed.

Transformation of ratings

An overall scale from -20 to $+20$ with negative versus positive numbers representing negative versus positive evaluations respectively was created by means of the following transformation: overall rating = $w - (21 - c)$, where w stands for the rating of feeling warm, and c for that of feeling cold. Subjects' ratings ranged from -17.8 to 12.9 , with a mean of -3.0 , that being significantly different from zero, $t(49) = 9.72$, $p < .01$.

Movement direction

To confirm whether toward movements represent approach responses and away movements represent avoidance responses, we split the response times of these movements according to whether they had occurred for positively or negatively rated attitude words. To also see whether indirectly inferred valence can be valid for subject samples and not only for individuals, we used the sample's average ratings to determine whether an attitude word is positive or negative. (As one would expect, the results also hold when we determine the valence of attitude words for each respondent individually based on his or her ratings.)

For each subject, we computed median response times in milliseconds (ms) for each cell of this design and conducted a 2×2 random effects ANOVA of rated valence of attitude word (negative vs. positive) \times movement direction (toward versus away). We computed medians instead of means because response time data are skewed. Participants responded faster when moving positively rated words toward their name than away from their name ($M = 812$ ms vs. $M = 1006$ ms, respectively). In contrast, they responded faster when moving

negatively rated words away from their name than toward their name ($M = 863$ ms vs. $M = 1004$ ms, respectively), $F(1, 49) = 58.42$, $p < .01$. This interaction effect shows that toward movements represent approach and that away movements represent avoidance.

Both main effects were also significant. Participants evaluated positive attitude words faster ($M = 909$ ms) than negative attitude words ($M = 934$ ms), $F(1, 49) = 5.78$, $p < .05$. This effect is not germane to our hypothesis. Moreover, because all our analyses are based on difference scores for each attitude word, this effect will not influence our interpretation of the results.

On average, participants executed toward movements faster ($M = 908$) than away movements ($M = 935$), $F(1, 49) = 6.51$, $p < .05$. Note that this effect is unlikely to have been caused by stimulus valence being on average positive, because according to ratings that was, if anything, negative. In fact, it is a standard finding that in response time tasks people execute positive responses (e.g., good, yes) faster than negative responses (e.g., bad, no). As Gordon Logan (personal communication, 2001) put it: "It's so generally true that there aren't really any references for it." Note that this main effect introduces a positivity bias, that is, all attitude words appear slightly too positive. To the degree that we are interested in using these response times to determine whether an attitude is positive or negative, such a bias could lead to incorrect classification of attitude objects and would need to be corrected.

Validation of reaction-time-based valence scores and assignment of multiple attitude objects to one scale

We have just confirmed that response time differences for toward and away movements meaningfully score whether an attitude object is positive or negative. It is also of interest whether these *valence scores* were related to the magnitude of the valence according to rating scales. This would show that valence scores are sensitive to the magnitude of valence, rather than simply to dichotomous positivity versus negativity. It would also show that valence scores have single item sensitivity because each level of valence was represented in the stimulus set by a different item. Finally, this would demonstrate that we can place multiple attitude objects on a single scale of valence scores and preserve their rank order of valence.

As some of the analyses focus on single responses, we could not use medians; we therefore log-transformed all the reaction times instead. We then computed valence scores from the response time task by subtracting the response time for toward movements from the response time for away movements for each participant and attitude word. This procedure yielded 92 valence scores per participant, one for each attitude word, indicating negative versus positive valence when the score was negative versus positive respectively.

We regressed the ratings of each individual subject onto his or her response time valence scores. On average, the more positively participants evaluated an attitude object according to the response time task, the more positively they rated the attitude object, $B = 10.14$, $t(49) = 8.33$, $p < .01$. The mean of the corresponding correlation coefficients was $r = .43$, that of the coefficients of determination was $r^2 = .26$, both based on Fisher- z transformations (see the top panel of Fig. 2 for a regression at the group level and the bottom panel for the actual attitude words).

In a recent article, we speculated that the more unfamiliar words are, the more negative they might be evaluated in response time tasks (Brendl et al., 2001). To test this possibility, we regressed the reaction-time-based valence scores both onto ratings and onto word frequency estimates taken from the CELEX database (INFORMATION, 1995). While more positive ratings of word valence predicted more positive reaction time valence scores, $B = .015$, $t(49) = 7.88$, $p < .01$, word frequencies did not predict reaction time valence scores, $t < 1$. However, word frequencies did significantly predict reaction time valence scores when entered without ratings into the regression equation, $B = 5.7 \times 10^{-5}$, $t(49) = 2.68$, $p < .05$. Although the effect was small, higher word frequencies predicted more positive evaluations based on response times. Ratings and word frequencies were collinear, $B = .0034$, $t(49) = 5.5$, $p < .01$, and thus word frequencies did not have a unique contribution beyond that of ratings in predicting reaction-time-based valence scores. Note that the number of syllables was collinear with word frequency and also did not predict uniquely. (We revisit the issue of stimulus familiarity in the General discussion).

Grammatical gender may be a potential moderating variable in this task. German nouns in their singular form have a gender, and so it is possible that response times may be affected by whether noun gender and the gender of the respondent match. The singular form gender of 35 words was female, of 21 words it was neuter, and of 36 words male. We computed a three-level variable reflecting whether a word matched the subject gender, was neutral in respect to it, or mismatched it. In individual subject regression analyses this variable predicted neither response time valence scores, nor did it explain additional variance beyond ratings in predicting response time valence scores, largest $t(49) = 1.13$, $p = .27$. Interestingly, it also did not predict word ratings, $t < 1$. Furthermore, response time valence scores did not differ as a function of the respondents' gender, $t < 1$.

Discussion

Participants were able to “move” positive words more rapidly toward than away from their names and “move” negative words more rapidly away from than

toward their names, showing that toward versus away movements represent approach versus avoidance, respectively. The degree to which respondents initiated one movement faster than the other correlated with the degree of positivity or negativity expressed via rating scales for individual respondents separately. This finding demonstrates that the reaction time task distinguishes more than two levels of valence of objects for individual subjects, has single item sensitivity, and can assign multiple attitude objects to a single scale.

However, this task does not permit calibration of the scale around a neutral point. One reason for this problem derived from signal detection theory is that the task does not allow us to assess whether participants shift their response criterion when asked to work on blocks with different movement instructions, which is a well-known task-inherent confound in response time tasks (Brendl et al., 2001). Given single item sensitivity, the correlation we observed will not be affected by such criterion shifts, but the neutral point of a scale will. We modified the task to address this issue.

The data suggest that the neutral point of the reaction-time-based valence scale is shifted. While most attitude objects are assigned the same valence by response times and ratings (cf. Fig. 2), a bias of the neutral point is reflected in the x-axis intercept of the regression line indicating that a neutrally rated object ($y = 0$) would receive a positive reaction-time-based valence score ($x > 0$). Recall that the reaction time scores have a positivity bias because respondents execute toward movements faster than away movements, irrespective of word valence. This is one factor that can shift the neutral point. If there are other factors as well, and if they affect respondents differently, then not only would the neutral point shift but correlations between response time valence scores and ratings based on between-subjects variability (not ones based on within-subject variability) would be lowered.

Study 2: Calibrating the attitude scale around a neutral point—the evaluative movement assessment

Study 2 introduces the *Evaluative Movement Assessment* (EMA), a variation of the procedure of Study 1. EMA involves two sets of attitude words. *Distractor words* are both positive and negative, like those in Study 1, but here the task is always to move positive distractor words toward the name and negative distractor words away from it. So far, then, the procedure is identical to the block in Study 1 where participants were instructed to move positive words toward their name and negative words away from it. One change we introduced was that upon responding, the words actually moved horizontally on the screen, either toward or away from the participant's name. The participant's name always remained

in a fixed position, but moving the joystick to the left or right moved the words to the left or right, respectively.

The second set of attitude words were *target words*, which are those whose valence EMA assesses. In one block of trials the respondent had to move all the target words toward his or her name, no matter what their valence, while in another block the same respondent had to move all target words away from his or her name. At the beginning of the task the respondent needs to learn the target words' identity because we did not provide any other cue for identification. In this experiment we asked participants to memorize five target words. In other experiments, we identified target words by category membership (e.g., all vegetables). Target and distractor words were interspersed within a block so that respondents were required to hold in memory both a response rule based on the valence of attitude words and a rule based on a word's category membership in the target word set. For example, if "cavities" is part of the target word set and all target words have to be moved toward the respondent's name, then upon encountering "cavities" a respondent must decide whether it is a member of the target word set, because it must then be moved toward the name; but if it is not a member, it must be a distractor, and then due to its negative valence it must be moved away from the respondent's name.

Because each target word has to be moved in each direction at least once, EMA can estimate the valence of a target word as a difference score between the response times for moving it away from and toward the respondent's name. We call this difference score an *EMA score*. Thus, more positive EMA scores express more positive (less negative) valence. The target words were "movie theater," "clown," "taxes," "politician," and "cavities." These five words were purposely chosen from Study 1 (see rectangles in Fig. 2, bottom panel), because they are spread across the range of possible valences on multiple levels both according to a response time measure and a self-report measure.

EMA should measure uncontrollable evaluative responses because participants are asked to move target words as quickly as they can in a particular direction (e.g., toward, as if they evaluated them positively), ignoring their valence. If participants cannot ignore (i.e., control) their own evaluation of the words, as evidenced by different response times for away and toward movements, then this effect of valence on response times is involuntary and presumably uncontrollable. In this sense, EMA is similar to Jacoby's process dissociation procedure for implicit memory as it pits controllable responses (e.g., moving cavities toward as if they were positive) against uncontrollable responses (e.g., an automatic tendency to move cavities away) (Jacoby, Yonelinas, & Jennings, 1997).

An advantage of EMA over the procedure in Study 1 is that it can detect response criterion shifts that may al-

ter the neutral point of the scale. The concept of response criterion shifts derives from signal detection theory. When respondents set their response criterion at a different level for blocks with different instructions for target words, they are said to have shifted their response criterion. For example, they may respond more cautiously in all trials of blocks where they are required to move target words away from their name than in blocks where they have to move target words toward their name. Unlike Study 1, in EMA we can use distractor trials to estimate any response criterion shifts. Instructions for distractor words are identical in all blocks of EMA, but for target words they change because in some blocks of trials respondents move all target words away from their name while in others they move all target words towards it. At first sight, responses to distractor words should be independent of instructions for target words because, by definition, these do not apply to distractor words. If changing instructions for target words affects response times to distractor words, this could be caused by a response criterion shift. We cannot only detect response criterion shifts but can also statistically correct for them.

We had a number of predictions. First, correlating self-reported attitudes toward these words with their EMA scores should yield sizeable correlations at the level of the individual, as the stimuli were chosen to contain within-subject variability and we hypothesize that EMA is sensitive to this.

Second, computing pair-wise difference scores of ratings and EMA scores for each of these objects and correlating these should yield lower (possibly non-significant) correlations. Our reasoning is that: (a) pair-wise difference scores are much less sensitive to within-subject variability while being sensitive to between-subjects variability, and (b) there is little systematic between-subjects variability in this data set. For example, most people prefer movies to cavities and the degree of this preference varies little. Recall our hypothesis that assessing correlations between self-reported and indirectly measured attitudes of vegetarians toward types of food would yield substantially higher correlations if variability in preferences was assessed within-subjects rather than between-subjects. Our second prediction allows us to demonstrate empirically that low correlations between self-reported and indirectly measured attitudes that are based on between-subjects variability can increase to substantial levels if within-subject variability is assessed. Of course, this applies to the particular situation of an attitudinal domain that has little between-subjects variability but much within-subject variability.

As our third prediction we hoped that the level of valence of at least three of the five target words would be distinguishable by being assigned significantly different EMA scores. This would provide additional support for the ability of EMA to measure more than two levels

of valence, and because these levels of valence were represented by different stimulus items it would also be further support for single item sensitivity.

Finally, we expected that the neutral point of an EMA scale would correspond to that of the ratings scale, after having statistically corrected the EMA scale for any response criterion shifts.

Method

Participants

Paris pedestrians, typically students, were recruited to the INSEAD lab in return for a small payment. Fifty-two met the criteria set in Study 1 (native speakers, vision corrected to normal, and non-dyslexic). The following were excluded from the analyses: a group of three participants did not complete the study, there was an experimenter error with another participant, and four participants made too many errors (see below), leaving $N = 44$ participants.

Ratings

Before the EMA task, each of the five target words was rated on a nine-point scale anchored “negative (−4)” to “positive (+4)” with 0 as a neutral point.³

EMA procedure⁴

Each trial began with a blank screen displayed for 600 ms, followed by presentation for 700 ms of the participant’s name in a white frame with four red X’s to the right or left of the name. The row of X’s was then replaced by a red attitude word (i.e., a filler, distractor, or target word). When a participant responded by moving the joystick to the right or left, the word also moved to the right or left, creating the impression that the joystick physically moved the word.⁵ Error feedback was given in each error trial, defined as an early response (between onset of the first name and up to 100 ms after stimulus onset), a late response (3000 ms after stimulus onset), or a response in the wrong direction.

There were six blocks of trials. The first block presenting only distractor words was introduced to respondents as practice. Respondents had to move a distractor word toward or away from their names according to whether most people would consider the word positive ($n = 12$) or negative ($n = 12$), respectively. We did not ask for the participant’s own evaluation of the distractor words because we wanted to give error feedback to par-

ticipants: previously a few participants had protested that it was inappropriate to give error feedback on subjective evaluations. Note that these instructions apply only to distractor words.

Five main blocks then followed, each consisting of 83 trials: eight filler trials (repeating four words twice), 40 distractor word trials (composed of 40 different words), and 35 target word trials (repeating each of five target words seven times). At this point, respondents were told to continue moving words (i.e., distractor and filler words) toward or away from their name according to whether most people would consider them positive or negative respectively, with the exception of a list of words they were now to memorize (i.e., the target words). These were “movie theater,” “clown,” “taxes,” “politician,” and “cavities.” Depending on the block (alternative instructions in parentheses) they were told: “Please do not try to decide whether you find these words positive or negative. Instead, each time you see one of these five words, evaluate it as negative (as positive). Pretend that you find each word negative by moving it away from your name (positive by moving it toward your name).”

Recall that we are measuring valence only for target words. Asking respondents to pretend that they have certain evaluations of target words has two implications. First, their task is to *not* let their evaluations influence their response, enabling us to infer that any such influence was uncontrollable. Second, we now refer to their own evaluation, which is different from instructions for distractor words. These referred to most people’s evaluations. We hoped what would be even more influential in activating respondent’s own evaluations was that attitude words are always moved in respect to their own name. We made it an empirical question whether these instructions would result in EMA scores that measure participants’ own evaluations. Our test was whether they predict respondents’ own self-reported evaluations. We acknowledge that we cannot tell whether respondents’ representations of other people’s evaluations also affect EMA scores. If this were the case, it would be unproblematic to change the instructions for distractor and filler words to reflect respondents’ own evaluations.

Half of the filler and distractor trials were positive, the other half negative. The main block started with four filler trials with a break after half of the trials, followed by the remaining four filler trials. Observing this restriction, the order of stimuli was determined randomly for each participant. Whether a stimulus was presented to the left or right of the name was determined randomly in each trial, with the restriction that each target was presented three times on one side and four times on the other and that the more frequent side was the same for all blocks. Whether in the first main block participants had to move all target words toward or away from their name was counterbalanced between-subjects.

³ A study in which EMA preceded ratings led to the same results (Study 4 in Appendix A).

⁴ The EMA software as well as the files that implement Studies 1 and 2 can be downloaded from the first author’s www-web page at INSEAD (<http://www.insead.edu/facultyresearch/marketing/brendl>).

⁵ Because the words move, the joystick can be replaced by two keys of the keyboard, as done in Study 3 reported in Appendix A.

The first main block was followed by four additional main blocks, set up in an ABBA design. The second main block used the same instructions as the first. For the third and fourth main blocks instructions to move target words were reversed, and the fifth main block used the same instructions as the first two. Unknown to participants, the first main block was for practice only. This block is important for overcoming initial non-linear learning effects.

Results

Data exclusions

Error trials were defined as above, except that responses shorter than 300 ms were considered early responses. All error trials were excluded from the analyses. We excluded all participants from the analyses who made errors on more than 7% of the 300 distractor and target trials in the four main blocks. Prior to these exclusions the mean error rate was 3.21%, after the exclusions it was 2.1%.

Correlations based on five attitude objects per respondent

EMA scores are computed as the response times of moving a word away from the participant's name minus moving it toward the participant's name plus a constant, the computation of which is described further below. Although we typically aggregate an individual's response latencies using medians, because of a better approximation of a normal distribution we averaged log-transformed response times. For each respondent we regressed his or her ratings of the attitude objects onto his or her EMA scores for these objects and *t* tested the slope coefficients against zero. The more positive an object's EMA score, the more positively it was rated, $B = 18.6$, $t(43) = 6.17$, $p < .01$. The mean of the corresponding correlation coefficients was $r = .64$, that of the coefficients of determination was $r^2 = .48$, all based on Fisher-*z* transformations. These results clearly show that the two types of measures share a substantial amount of variance.

Correlations based on two attitude objects per respondent

By reducing the number of attitude objects from five to two, our data set allows us to simulate the analyses performed in research designs that investigate the relative preference of one attitude object to another. We computed correlations using all possible combinations of two attitude objects. For instance, for the objects "cavities" and "movie theater" we computed one difference score between their ratings and another one between their EMA scores. This is the design typically used in those studies whose low correlations between ratings and response time measures have led to the assumption that implicit measures tap implicit attitudes. Just as in those studies, some correlations do not differ significantly from zero and others are low (cf. Table 2—bottom triangle). Two correlations reached conventional levels of significance ($p < .05$).

Our correlations based on five and two attitude objects mirror those reported in the literature on "implicit-explicit correlations," as a glance at Table 1 confirms. Sometimes they are quite high, sometimes low, and quite frequently not statistically different from zero. In our case, however, it is obvious that this enormous range is a result of increasing versus decreasing systematic variability, here within-subject variability, by correlating rank orders versus degree differences between attitude objects. High and low correlations, here, do not reflect explicit and implicit attitudes.

Preferences between two attitude objects

With two attitude objects per respondent it is also possible to capture within-subject variability, although at a crude level. To do so, we compared the number of respondents for whom the preference rank order between two objects was the same according to both measures to the number of those for whom it was opposite. Respondents who rated both objects equally were dropped. Out of ten possible pair-wise comparisons, in seven cases the two measures agreed significantly (cf. Table 2 top triangle). In other words, in these cases EMA predicts reliably which of two attitude objects a

Table 2
Pair-wise comparisons of attitude objects in Study 2

	Cavities	Taxes	Politician	Clown	Movie theater
Cavities		18/20	17/20	33/8**	35/7**
Taxes	.13		21/13	34/9**	41/2**
Politician	-.24 ⁺	-.06		29/13*	32/9*
Clown	.31*	.27 ⁺⁺	.01		28/12*
Movie theater	.20 ⁺	.12	-.07	.35*	

Note. The top triangle shows before the slash the number of participants for whom the rating scales and EMA agreed as to which of the two attitude objects are preferred. The number of participants for whom these measures disagreed is shown behind the slash. The number of participants who rated the two attitude objects as equally attractive are excluded from this analysis but can be computed by subtracting the sum of the numbers before and after the slash from all participants ($N = 44$). The stars indicate whether significantly more participants agreed than disagreed on both measures according to binomial tests: * $p < .05$, ** $p < .01$. The bottom triangle shows correlation coefficients between difference scores of EMA scores and difference scores of rating scales for each of two attitude objects (* $p < .05$, ++ $p < .10$, + $p < .20$, $N = 44$).

person will choose. This is a remarkable contrast to the only two out of ten cases in which difference scores derived from each type of measure correlated significantly with one another. One analysis suggests that the two measures are associated, the other suggests they are dissociated, yet the raw data are identical.

There is another striking pattern in these data. The two types of measures always agree significantly when two attitude objects of opposite valence are contrasted, but not always when they have equal valence. The greater the difference in valence between two attitude objects, the more subjects will prefer one attitude object to the other, which is equivalent to greater within-subject variability.

In sum, the question raised is how to interpret low correlations between response time and self-report measures. One interpretation is that they demonstrate psychologically dissociated constructs. This interpretation implies that the two measures share minimal variability, which has to extend to both between-subjects and within-subject variability. Our data demonstrate that this conclusion may be erroneous when between-subjects variability in an attitude domain is low. To the best of our knowledge, this potential confound has not previously been discussed in the literature.

Sensitivity to multiple levels of valence

Table 3 presents the median EMA scores for each attitude object.⁶ EMA significantly distinguished three levels of valence. EMA scores always classified “cavities” and “taxes” as more negative than “politician,” and “politician” as more negative than “clown” and “movie theater” (cf. Table 3). Out of 10 possible pair-wise comparisons of target words, eight were significantly different, all p 's < .002. The two non-significant differences were for “cavities–taxes” and “movie theater–clown.”

Why could only three levels of valence be distinguished? We suspect that running EMA on a small number of participants is insufficiently powerful to detect small differences in valence. Using data from Study 2 and its two replications (cf. appendices), we plotted the differences between two EMA scores against the differences between the two corresponding ratings, both as effect size estimates (Fig. 3). The six non-significant differences between two EMA scores are visible below the broken line in Fig. 3. The plot suggests that if the self-reported valence of two attitude words differs by one scale point or more on Cohen's d -scale, EMA scores will likewise differ significantly.

Table 3
Study 2

	EMA scores	Ratings
Cavities	–51	–0.91
Taxes	–44	–1.77
Politician	(6)	(–0.50)
Clown	40	3.00
Movie theater	50	3.55

Note. EMA scores are based on medians. They reflect differences in ms between moving target words away from and toward the first name plus a constant for scale calibration. More positive scores reflect more positive valence. Ratings were collected on 9-point scales from –4 to +4. Numbers in parentheses do not differ significantly from zero. All pair-wise comparisons are statistically significant at the $p = .05$ (two-tailed) level with the following exceptions: EMA scores: cavities–taxes, clown–movie theater; Ratings: cavities–politician, clown–movie theater.

Effect Sizes of Differences between Evaluations of Two Target Words

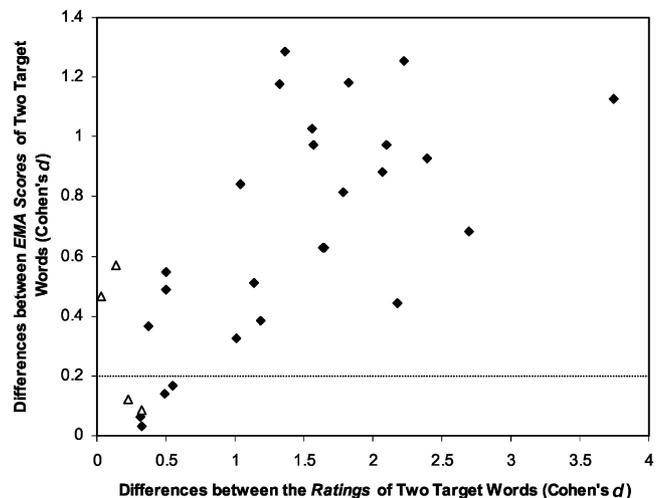


Fig. 3. Differences (Cohen's d) between ratings of two attitude objects as a function of the difference (Cohen's d) between the two EMA scores of the same attitude objects in Experiment 2 and two additional experiments that used the same target word set (cf. Appendix A, Experiments 3 and 4). All coordinates above (below) the dotted line reflect significant (non-significant) differences between two EMA scores. The four hollow triangles represent the only non-significant differences between two ratings.

Single item sensitivity

Because the three levels of valence distinguished by EMA were associated with different items, EMA has to be sensitive to these three items associated with the three levels of valence. But there were additional indications for single item sensitivity. With one exception, the five items of each target word set were rank ordered identically according to EMA and ratings (cf. Table 3). The exception is that in Study 2 participants rated “taxes” as significantly more negative than “cavities,” a result we find quite counterintuitive, but their EMA scores classified “cavities” as non-significantly more

⁶ In Study 2, we conducted all the analyses with logarithms of response times. However, because results are easier to understand when they are on a millisecond scale than on a log scale, we present medians of response times.

negative than “taxes.” In two replications of this study (Appendix A: Studies 3 and 4) EMA and ratings ordered all of these attitude objects identically. Thus, EMA ranks individual items in the same order as explicit ratings, suggesting sensitivity to these items.

Calibrating the EMA scale around a neutral reference point

So far we have focused on cases where one attitude object is preferred to another, irrespective of whether the objects are viewed as positive or negative. A measure that supports the evaluation of multiple attitude objects can potentially be calibrated around a neutral point, because a set of attitude objects can be used as a reference frame that is centered around a neutral point, and other objects of interest can be related to this frame. We are not aware of any response time measure of attitudes for which calibration around a neutral point has been attempted.

As mentioned above, Study 1 provides evidence that, independent of stimulus valence, in our task toward movements are executed more rapidly than away movements. This implies that no matter what the reason for this main effect of movement direction, it biases EMA scores such that they are too positive.⁷ Because of the high correlation between EMA scores and the explicit ratings, for this analysis we validate the neutral point of the EMA scores with the neutral point on the explicit rating scale. As “politician” was rated as neutral, it is an excellent validation criterion for the neutral point of the EMA scale. As would be expected from Study 1, for the same attitude object difference scores between response times of away and toward movements are more positive than ratings. The whole response time scale seems to be shifted toward the positive end. “Politician” has a positive response time difference score ($M = 32.9$, $t(43) = 2.8$, $p < .01$) but should have a neutral one. “Cavities” and “taxes” were rated negatively, but their response time difference scores are non-significantly negative ($M = -23.8$; $M = -16.6$), $t(43) = 1.85$, $p = .07$; $t(43) = 1.72$, $p = .09$. Finally, “clown” and “movie theater” were rated positively and showed very reliably positive response time difference scores ($M = 66.9$; $M = 76.9$), $t(43) = 6.48$, $t(43) = 6.37$, p 's $< .01$.

The advantage of single item sensitivity is that we can estimate the degree of an individual respondent's positivity bias in response times by including a set of items whose valence is pre-tested. The positivity bias contained in other items can then be removed by simply subtracting the bias estimate from all response times. In Study 2, the five target words had been chosen a pri-

ori such that their average evaluation should be roughly neutral. Thus, if toward movements are again executed faster than away movements, a main effect of movement direction, response times include a positivity bias. (The size of this bias is half the difference between the response time means for the toward and away movement conditions). We can estimate the size and direction of this bias individually for each subject and subtract it from the raw response times (see Appendix B for details). Response time difference scores that have been corrected this way are labeled *EMA scores*. The neutral point of the scale of EMA scores corresponds well to that of ratings. Corresponding to its rating, “politician” now has an EMA score that is almost equal to zero, ($M = 6$), $t < 1$. The two negatively rated attitude objects (“cavities” and “taxes”) now have significantly negative EMA scores ($M = -51$; $M = -44$), $t(43) = 6.3$, $t(43) = 5.8$, p 's $< .01$, and the two positively rated attitude objects (“clown” and “movie theater”) significantly positive EMA scores ($M = 40$; $M = 50$), $t(43) = 5.3$, $t(43) = 6.7$, p 's $< .01$ (Table 3).

Although the neutral point of the calibrated EMA scale is determined solely on the basis of response times, thus not involving any ratings, it corresponds very well with the neutral point of the rating scale. This finding validates the calibration procedure. The match between the calibrated EMA scale and the rating scale is impressively close: not only are rank order relations of valence between multiple attitude objects the same on both scales, the point at which both scales switch from positive to negative valence is the same, suggesting that away versus toward movements, after correcting for their positivity bias, capture the valence of attitude objects well. We tested this calibration procedure across six studies reported in Appendix B (see also the second table of Appendix) and the results are consistent with those of Study 2.

Response criterion shifts and learning effects

Recall that response criterion shifts would displace the neutral point of the EMA scale. Appendices C and D describe a meta-analysis of ten very similar EMA studies in which we detected criterion shifts. We show there that the calibration procedure estimates the size of an individual's criterion shift and statistically corrects for it. These shifts leave the rank order of EMA scores in place and therefore do not affect the correlations of EMA scores and ratings, making statistical correction unnecessary if only correlations are important.

There are also learning effects in (uncalibrated) response time difference scores that vary in size depending on whether the first required movement of target words is toward or away, that is, there are *order effects of instructions*. The calibration procedure estimates the overall bias introduced by criterion shifts and by these order effects and removes it successfully. This is indi-

⁷ We speculate that respondents may require a stricter response criterion for responses that mean “negative/away” than for ones meaning “positive/toward.” Consistent evidence is presented in Appendix C.

cated by a close match of the neutral point of calibrated EMA scores with that of ratings and the absence of order effects of instructions in (calibrated) EMA scores.

Internal consistency

For each respondent and target word we computed two EMA scores by splitting all trials in an alternating fashion into two groups. Cronbach- α coefficients were: .73 (movie theatre), .54 (clown), .47 (cavities), .37 (taxes), and .32 (politician). The more target words one includes in EMA, the fewer trials will be assigned to each, reducing reliability. Recall that the most popular research design employed in implicit attitudes research computes a relative preference of one target category to the other, in this way collapsing across all trials and increasing internal consistency. To allow comparison to this research design we computed the relative preference of the two positive target words to the two negative ones, resulting in a Cronbach- α coefficient of .79. Perhaps the most informative analysis regresses for each subject individually the five EMA scores of one half of the data onto the five EMA scores of the other half. The mean of these correlations coefficients is $r = .51$. Of course, it is based on splitting the data in half. Using the Spearman-Brown formula, we can estimate that this coefficient would have been $r = .68$ with twice the number of trials, which corresponds to the number of trials in a test-retest design. A test-retest correlation should hence have been less or equal to $r = .68$.

Discussion

We introduced EMA as a measure suited to capture within-subject variability and validated certain aspects of it using self-reported evaluations. Specifically, we showed that EMA scores reflect evaluations of attitude objects, that they are sensitive to single items and therefore capable of distinguishing ordinal relations of single attitude objects for individual respondents, and that they can be calibrated around a neutral point resulting in assignment of positive versus negative valence to single attitude objects.

Studies on implicit attitudes have typically used a research design that measures attitudes toward one or two attitude objects. Study 2 also demonstrates that this design is not powerful for assessing within-subject variability. This has implications for the question whether low correlations between attitude measures indicate psychological dissociations. When one or two attitude objects are measured, sufficient between-subjects variability is a necessary condition for inferring dissociations from low correlations. If the research question is whether indirectly measured attitudes are implicit attitudes, then the conservative approach is to maximize systematic variability, both between-subjects and within-subject

variability. Study 2 provides an empirical demonstration of this view.

General discussion

The main contribution of this research is the introduction of a new response time measure of evaluation that can: (a) place a number of individual attitude objects onto a single scale and (b) meaningfully center that scale around a neutral point. The first property makes it possible to assess more powerfully the correlation between indirect and self-report measures of attitudes in attitude domains with little between-subjects variability. The second property supports inferences about whether an evaluation is positive or negative. As a secondary contribution, we demonstrated empirically that low correlations between two measures can be the result of a lack of measurement sensitivity to within-subject variability. Finally, we introduced the concept of single item sensitivity and discussed its implications for indirect measurement.

Low correlations of difference scores between two attitude objects need not reflect reversed preference rank orders on two measures

Many discussions in the implicit attitudes literature imply that low correlations between two measures mean that preferences for them are reversed, that we prefer object A to B on one measure but B to A on the other. Study 2 shows that low correlations can occur when subjects have the same preference rank order for both measures. The previously unidentified reason is methodological: it applies to a research design that measures once via self-report and once indirectly the degree to which respondents prefer attitude object A over attitude object B. Even if both measures agree that most respondents rank A before B (i.e., the measures agree on preference rank order), the correlations between the degrees of preference of A over B will be very low if most respondents have the same degree of preference. Thus, low correlations of degree of preference measures need not reflect preference reversals.

Degree of systematic variability moderates correlations between self-reported and indirectly measured attitudes

Study 2 showed that *within the same data set* the correlation between an indirect measure and a self-report measure was low and non-significant when assessing between-subjects variability, but substantial and significant when assessing within-subject variability. The attitude domain exhibited low between-subjects variability and high within-subject variability because most people prefer, for example, movies to cavities and do

so to a similar degree. Although we do not have evidence, we suspect that those IAT studies in Table 1 that show low correlations are from attitude domains that exhibit low between-subjects variability (e.g., flowers vs. insects). As these correlations are based on a research design that measures mostly between-subjects variability, if in fact that variability is lacking in the data, the studies do not test the hypothesis that the IAT measures something different than the self-report. Further, we suspect that those studies showing high correlations are from attitude domains that exhibit between-subjects variability (e.g., voting preferences). More generally, low correlations should only be interpreted as support that a method measures implicit attitudes if it can be ruled out that these low correlations are not due to a lack of both between-subjects and within-subject variability.

Some readers may wonder whether the low correlations between EMA and ratings that we have observed in Study 2 when analyzing two attitude objects could result from a low reliability of EMA and/or ratings. Note that this type of explanation cannot account for the dramatic increase of correlations when we analyze five instead of two attitude objects.

Psychometric evaluation

Reliability

When assessing the relative preference of one attitude object to another, the internal consistency of EMA is comparable to that of response time measures using that research design. Then internal consistency is reasonably high. When EMA scores for individual attitude objects are computed, internal consistency was low for some attitude objects and reasonably high for others. Looking at several single attitude objects comes at a price, namely reduced reliability for each individual respondent's data. Thus, the more single attitude objects a researcher is interested in, the larger the necessary sample size. A look at the second table of Appendix A suggests that at the group level stability is quite high. Most attitude objects have fairly similar standardized EMA scores across the various studies we conducted.

Potential confounds

One contribution of this research is to offer a methodology that removes the influence of criterion shifts as well as learning effects when these could be confounds. Further, we failed to find evidence that gender of nouns confounds EMA scores. However, there are other potential confounds that we did not investigate. (a) It is conceivable that greater *self-esteem* leads to more positive EMA scores for positive attitude objects but to more negative EMA scores for negative attitude objects. (b) *Familiarity* of attitude objects may affect EMA scores. Although in Study 1 word frequency did not

have a unique contribution to predicting EMA scores, this is weak evidence against such an influence. In pilot research we found that non-words, that is, extremely unfamiliar stimuli, received negative EMA scores. Future research will need to determine the influence of familiarity on EMA scores as well as whether such an influence is a confound or a legitimate cause of evaluations measured by EMA. (c) All response competition procedures (e.g., Stroop task, Simon task, IAT, and EMA) are probably affected by respondents' *ability to inhibit distractions* (see McFarland & Crouch, 2002; for the IAT). For example, color Stroop interference has been shown to increase in adults over 60 (MacLeod, 1991). This ability to inhibit should affect between-subjects comparisons but not within-subject rank orders of attitude objects. (d) The *composition of the distractor word set* as well as the *relative frequency of toward and away movements* may affect EMA scores. However, these factors should not affect the rank order of EMA scores, but shift their neutral point. Further, the calibration procedure should reduce such influences. (e) The *composition and type of the target stimuli* in Study 2 involved five target words from five different categories. Imagine an EMA with: (i) picture stimuli, (ii) all from the same category, and (iii) all being perceptually similar, for example, five pictures of very similar chairs as targets. Because one or more of these factors could make the chairs difficult to discriminate and instructions do not require discrimination, single item sensitivity might be lost.⁸ Respondents might be able to respond to target stimuli solely on the basis of category membership without further identifying them. One solution could be to include a sufficiently high number of equally similar chairs among the distractor pictures, forcing respondents to identify individual chairs rather than identifying them only as an instance of the category chair.

For two other potential confounds we can report some preliminary data. (f) There are likely to be *response strategies* that respondents can use. These could constitute confounds, so EMA ought not to be used as a lie detector. Asendorpf, Banse, and Schnabel (2003) investigated whether participants instructed to fake their responses could spontaneously do so both for the IAT and for a mixture of the IAT and EMA they termed the *IAP* (Implicit Association Procedure). Like EMA, the IAP creates response competition via a representation of the respondent's name on the monitor and joystick movements, which are however directed forward-backward rather than left-right (see Markman & Brendl, in press; for a forward-backward movement task more similar to our Study 1). There was some evi-

⁸ We thank Maria Galli for discovering and bringing up this issue during her dissertation research.

dence for successfully faking IAT effects ($p < .05$, one-tailed) but not for IAP effects. These data do not, however, exclude the possibility that individual participants may have successfully faked responses or that more experience with the task could lead to more successful faking.

(g) There is a possibility that EMA scores are affected by a respondent's representations of other people's evaluations (see Karpinski & Hilton, 2001), particularly in the light of our task instructions to move distractor words based on most people's evaluations of them rather than on the respondent's own evaluations. We can report evidence that EMA scores are affected by respondents' own evaluations. In a conditioning study by Galli, Chattopadhyay, and Brendl (in preparation), positively conditioned shampoo bottles were preferred to neutral ones according to EMA. However, these data do not rule out that due to our distractor word instructions respondents' representations of other people's evaluations also influence their EMA-scores. Olson and Fazio's (2004) research with the IAT suggests that: (1) giving error feedback, (2) the type of distractor words, and (3) instructing respondents to categorize words as "I like"/"I dislike" rather than as "positive"/"negative" could have similar influences. We need to explore these possibilities also for EMA.

Predictive validity

In an unpublished study Brendl, Markman, Heller, and Chattopadhyay (2002) explored when EMA predicts deliberate decisions. Participants were split into two groups, "decided" and "torn," concerning their self-reported attitude toward the right of homosexuals to adopt children. "Decided" participants were either for or against this right. "Torn" participants were in conflict. Later all participants made a forced choice regarding whether they would support or oppose in a referendum the granting of legal adoption rights to homosexuals. Finally, EMA scores reflecting their relative evaluation of heterosexuality to homosexuality were obtained. EMA scores did not predict referendum decisions for the decided participants. However, for torn participants, the more positive participants' evaluation of heterosexuality compared to homosexuality according to EMA scores, the more likely they were to decide against the new law. We assume that EMA measures an uncontrollable but conscious good–bad response that makes up one part of an evaluative judgment. In the present study it was hypothesized that decision-makers would rely on this crude information if other information does not allow them to make a decision, as was the case for torn participants. This first construct validation study suggests a condition when EMA scores predict deliberate behavior. We harbor no illusion that much more in terms of unique prediction of behavior must follow if indirect measurement is to have a long-term impact. Part of understand-

ing what construct EMA is measuring will also have to be studies of convergent and discriminant validity with other indirect measures. The current research suggests that these studies will need to ensure adequate systematic variability.

Implications

We have deliberately avoided using the terms "attitude" or "test" regarding EMA, because we did not want to claim that it measures attitudes (as opposed to some other kind of evaluation), nor did we wish to suggest that it is a diagnostic test. The above discussion shows that a great deal of work remains before EMA could become a test, if ever. However, we can start to use EMA if we interpret the results with its limitations in mind. In most experiments, researchers successfully use dependent measures without knowing much about their psychometric properties, but they carefully limit their interpretation of them. We suggest the same strategy for EMA.

Comparison with other response time measures

Evaluative priming (BFP)

Fazio et al. (1986, 1995) introduced the "bona fide pipeline" technique (BFP), an evaluative priming procedure. In multiple regressions involving various predictors and multiple attitude objects, self-report predicted BFP attitude scores (Bargh et al., 1992; Fazio et al., 1986). Other studies have shown that within the category of Black faces, attitude scores separate different types of faces (Livingston & Brewer, 2002). While these studies suggest that multiple attitude objects could be assigned to one BFP scale, they were not designed to address this question. Future studies should assess within-subject correlations of self-report and BFP scores without inclusion of other predictors in the regression and investigate whether a BFP scale would distinguish multiple levels of valence.

It is possible that BFP attitude scores can be calibrated around a neutral point, but this would require that there are no criterion shifts. Because all blocks of trials use the same instructions, the problem of criterion shifts between blocks with different instructions does not arise. However, participants need to respond by categorizing stimuli as good or bad, and they may have different response criteria for these categorizations. One way to compute BFP scores is as an interaction of this categorization (positive vs. negative) with trials in which a prime was present versus absent. Any effect of different response criteria for positive versus negative categorizations would be removed from the interaction term because it is a main effect (Rosnow & Rosenthal, 1991). Thus, such criterion shifts could be removed statistically possibly allowing neutral point calibration of the BFP scale.

IAT

In the IAT, respondents sort instances of categories (e.g., spinach, salad, steak) into two categories (e.g., vegetables vs. meats). The IAT is constructed to measure the relative preference of one category to another, ignoring attitudes toward the individual instances. Hence, by definition the IAT does not assign more than two attitude objects to a single scale and does not have single-item sensitivity. Could it be modified to do so? For example, could trial-by-trial response times to the various instances be analyzed separately? The developers of the IAT have argued against such analyses and recent evidence supports their view. Research shows that such decomposition attenuates the predictive validity of IAT effects (Nosek, Greenwald, & Banaji, 2004). One argument is that the decomposition does not work well because every single response in an IAT is the expression of a relative preference (Nosek & Banaji, 2001). There is another possible reason. If one accepts single-trial analyses of the IAT, then the resulting data pattern suggests that respondents shift their response criterion between different IAT blocks (Brendl et al., 2001). These criterion shifts would add considerable noise to a decomposed IAT effect. If in a vegetable-meat IAT a decomposed IAT-effect for vegetables (i.e., a vegetable IAT-score) is influenced by respondents' attitudes toward meats because of criterion shifts, this influence adds noise to the vegetable IAT-scores.

Another issue with single item sensitivity is that the IAT seems to be sensitive to the valence of the category that an instance represents (e.g., vegetables) in addition to or possibly instead of to the valence of the instance, that is, the item itself (e.g., spinach, salad) (De Houwer, 2001). Single item sensitivity would require that, for example, in a meat vs. vegetable IAT, trial-by-trial data would pick up a rank order of individual instances such as salad being preferred to broccoli being preferred to spinach.

Even if the IAT is not sensitive to single items, by conducting multiple IATs where participants sort two category instances (e.g., broccoli vs. steak, milk vs. steak, milk vs. broccoli) for each IAT, it may be possible to rank order multiple attitude objects on one scale by means of pair-wise comparisons (Brunel, Tietje, & Greenwald, *in press*). However, because all attitude objects have to be paired with each other, the number of attitude objects that can be compared is quite limited.

As IAT effects are an expression of relative preference, it is meaningless to ask whether they are calibrated around a neutral point, and it is hence meaningless to infer liking versus disliking from them. However, conceivably with the strategy of pair-wise comparison among IATs, one could include a pre-tested neutral attitude object as a neutral scale anchor.

GNAT

The go/no-go association task (GNAT) (Nosek & Banaji, 2001) is a modification of the IAT that uses one category instead of two (e.g., vegetables only). Here we discuss a version of the GNAT in which participants are forced to respond very rapidly and their error rates are analyzed as dependent variables using signal-detection analyses. These analyses presuppose that single trials can be analyzed. Hence, multiple attitude objects could potentially be rank ordered within one GNAT. If the GNAT is sensitive to single items, attitude objects need not be limited to one category. The signal-detection analysis can be regarded as a calibration procedure, because it subtracts the estimated influence of response criteria from response times, leaving a "calibrated" attitude score. Thus, this version of the GNAT would permit to remove the influence of criterion shifts from the data. It is conceivable that a response deadline can be set so tightly that respondents have no chance to control their response threshold, a possibility awaiting further research. This could control criterion shifts experimentally. These properties suggest that future studies could show the GNAT to have single item sensitivity. Using pre-tested neutral attitude objects as anchors is conceivable but does presuppose either single item sensitivity or multiple GNATs.

EAST

Recently De Houwer (2003) presented the Extrinsic Affective Simon Task (EAST). Participants sort words displayed in white text into positive versus negative categories by pressing one of two keys. Intermixed with these "white" words, they sort another set of words depending on their font color; for example, they press the "good" key for green colored words and the "bad" key for blue colored words. By changing its font color, each attitude word is sorted once together with positive words and once with negative words. Based on response times and error rates the author reports two sets of EAST scores indicating the valence of these colored words.

It is principally conceivable that the EAST could have single item sensitivity and assign multiple attitude objects to one scale. While De Houwer's Experiment 2 did present five different attitude objects to each respondent, the crucial test of these task properties would be whether EAST scores distinguish at least three levels of valence of three different attitude objects. Because this test is not reported we cannot draw conclusions about these task properties. Currently we also lack the data to assess whether response factors displace the neutral point of the EAST scale, but given single item sensitivity, a calibration procedure could be developed.

EMA compared to other measures

Our research shows that if investigators seek to determine evaluations of several individual attitude objects, or whether an attitude object is positive or negative, they can use EMA. While other measures may be used for these questions too, they have not been systematically tested in this respect. We hope that this article draws attention to the properties necessary for investigation of these questions and that it will foster further development along these lines.

We do not mean to say that measuring attitudes toward single attitude objects is always desirable. Depending on the research question, there can be advantages to measuring relative preferences of one attitude object to another.

The calibration procedure we introduced can reduce biases due to response criterion shifts, but it may not fully correct them. The preferred route would be to undermine response criterion shifts experimentally. One possibility may be the use of a response deadline procedure (Greenwald, Draine, & Abrams, 1995; Ratcliff & McKoon, 1989), as has been done with the GNAT and the IAT (Cunningham, Preacher, & Banaji, 2001). The calibration procedure also introduces a practical limitation because it requires the inclusion of attitude objects among the target words that allow calibration of EMA scores. The target word set needs to include at least one neutral target word or, alternatively, at least one negative and one positive target word of equally extreme valence. This comes at the cost of reducing the number of trials for those target words that the investigator is actually interested in. Of course, calibration is only necessary if researchers wish to estimate the EMA scale's neutral point. The criterion shifts that are removed by the calibration procedure should not affect differences between two EMA scores (i.e., the prototypical research design in implicit measurement), nor within-subject correlations between EMA-scores and other scores.

What is the relation between self-reported and indirectly measured attitudes?

While speculative, we believe that EMA taps an initial evaluative response that is uncontrollable and in that sense automatic. It may measure what is colloquially termed a "gut reaction." The sources of that response may well be unconscious, but this does not mean that the reaction itself is unconscious (for similar views: Fazio & Olson, 2003; Wittenbrink, Judd, & Park, 1997). This view can account for sizeable correlations between EMA and ratings, but it also suggests that EMA may give a purer measure of this initial automatic evaluative response than self-report.

Appendix A. Description of the meta-analyses

In the appendices we summarize meta-analyses of all ten EMA experiments, where Study 2 is the same as that described in the main body of the article. The 10 experiments presented different target words, identified in the left column of the first table of Appendix. Each specific experiment presented those target words for which EMA scores are listed in the relevant columns—for example, Experiment 5 presented the target words "death" and "tarantula." Another difference between experiments was that only in Experiments 2–4 did the target and distractor words move on the computer screen. In Experiments 5–11 they remained in a fixed position. Finally, Experiment 3 used the keyboard in place of a joystick as input device.

The means used in the meta-analyses were weighted by the number of participants in each particular experiment. In order to preserve the millisecond scale and facilitate comprehension, the means used for the table were not standardized. Note, however, that these means are not equivalent to the meta-analyses, which also takes different variability in different experiments into account.

When we describe (uncalibrated) response time difference scores, we mean the difference between the response time means for the toward and away movements for one target word (see first table of the Appendix). (Calibrated) EMA scores (see the second table of Appendix) result from subjecting response time difference scores to the calibration procedure described in Appendix B.

Appendix B. Calibration procedure

It is an empirical fact that in EMA toward movements are executed with shorter response times than away movements, independent of the stimulus valence. In seven out of eight studies that hold stimulus valence constant by having both positive and negative target words, movements *toward* target words were executed faster ($M = 629$ ms) than movements *away* ($M = 664$ ms), $z = 8.67$, $p \rightarrow 0$. No matter what the source of this effect, it will bias EMA scores, and hence needs to be removed if a researcher wishes to infer whether an EMA score reflects a positive or negative evaluation. Calibration is therefore also a de-biasing procedure.

We estimated the positivity bias in terms of the difference between an individual participant's time to move the same target words toward and away from his or her name. To calibrate EMA scores this difference was divided by two, and the resulting score was subtracted from the response time for away movements for the target words and added to the response time for toward movements for the target words. This estimate of the positivity bias is meaningful only if a target word set can be assumed to have equally positive and negative target words, which applies in six of our studies. In practice, a set of pre-tested words can be included in a target word set to estimate the degree (and direction) of bias.

Across the 10 studies (uncalibrated) response time difference scores classified all but one of the positively rated target words as significantly positive, but did not classify any negatively rated target words as significantly negative (first table of Appendix). Excluding the word "politician," all but one of the (calibrated) EMA scores and ratings agreed on whether an attitude object was significantly positive or negative. The one exception was the word "death," whose negative EMA score was not significant in one case. The word "politician" was not consistently rated as negative, a pattern mirrored by the EMA scores (second table of Appendix). The calibration procedure works very well.

Uncalibrated response time difference scores

Study	2	3	4	5	6	7	8	9	10	11
<i>Negative words</i>										
Insects ^a									-18 ⁺	-13
Tarantula				3		8	-9	-25 ⁺		
Bedbug								-6		
Death				13		10	28 ⁺			
Cavities	-24 ⁺	-22	-18 ⁺							
Taxes	-17 ⁺	26 ⁺	19 ⁺							
Politician	33 ^{**}	-39	23							
<i>Positive words</i>										
Favorite meals									65 ^{**}	
16 flowers										56 ^{**}
Sunshine					39 ^{**}	85 ^{**}	53 ^{**}	67 ^{**}		
Butterfly					22	104 ^{**}	50 ^{**}	64 ^{**}		
Clown	67 ^{**}	108 ^{**}	90 ^{**}							
Movie theater	77 ^{**}	118 ^{**}	73 ^{**}							

Note. Uncalibrated response time difference scores are the median response latencies (ms) of moving a word away from the participant’s name minus moving it toward the name.

^a In Study 10 each participant chose two disgusting insects. In Study 11 each participant saw the same 16 negative insect words.

* $p < .05$.

** $p < .01$.

+ $p < .10$.

(Calibrated) EMA scores

Study	2	3	4	5	6	7	8	9	10	11
<i>Negative words</i>										
Insects									NA	NA
Tarantula				NA		-44 [*]	-40 ^{**}	-50 ^{**}		
Bedbug								-31 ^{**}		
Death				NA		-42 ^{**}	-3			
Cavities	-51 ^{**}	-60 ^{**}	-55 ^{**}							
Taxes	-44 ^{**}	-77 ^{**}	-18 [*]							
Politician	6	-12	-15							
<i>Positive words</i>										
Favorite meals									NA	
16 flowers										NA
Sunshine					NA	33 [*]	23 [*]	42 ^{**}		
Butterfly					NA	52 ^{**}	19 [*]	39 ^{**}		
Clown	40 ^{**}	70 ^{**}	35 ^{**}							
Movie theater	50 ^{**}	80 ^{**}	53 ^{**}							

Note. NA: Scores are not available because they cannot be calibrated. Specifically, in Studies 5, 6, and 11 all target words had the same valence and in Study 10 each respondent saw other target words.

* $p < .05$.

** $p < .01$.

+ $p < .10$.

Appendix C. Are response criteria different in blocks with different movement instructions for target words?

There is a possibility that respondents set varying conservative response criteria in blocks with different movement instructions for target words. Because EMA scores involve computing difference scores between these two types of blocks, a change in response criterion would constitute a confound. We would not know whether an EMA score is the result of an item’s valence, or of a shift in response criterion, or both. To exclude this possibility, we analyzed response times to distractor words. We found evidence for criterion shifts in uncalibrated response time difference scores, but also found evidence that calibration of these (i.e., yielding EMA scores), removes these effects. Essentially, the calibration procedure estimates the degree of criterion

shift. A criterion shift adds a constant to each EMA score, that is, it shifts all EMA-scores to the same degree in the same direction. Taking difference scores between EMA-scores will also remove this influence.

How to diagnose criterion shifts

What happens when a participant is instructed to move all target words toward his or her name? If a stimulus is positive—whether a target or distractor word—it must be moved toward the participant’s name; thus a toward movement can be executed without deciding whether the stimulus is a target or distractor word. Responses should be fairly fast because only the valence of the stimulus needs to be identified. If a stimulus is negative, it must be moved toward the participant’s name if it is a negative target word, but away if it is a

negative distractor word. Responses should be slower because, in addition to its valence, the respondent must identify whether the stimulus is a target or distractor word. These predictions are reversed when movement instructions for target words are reversed. In sum, positive stimuli (target and distractor words) should be responded to faster when target words must be moved toward rather than away from the participant's name. In contrast, negative stimuli (target and distractor words) should be responded to faster when target words must be moved away rather than toward the participant's name.

This prediction implies that changing movement instructions for target words (e.g., from moving target words towards the participant's name to moving them away) affects both directions of movement (e.g., speeding up away movement of negative distractor words but slowing down toward movement of positive distractor words). Because both movement directions are affected in opposite directions, such a data pattern cannot be explained by a shift of response criterion between blocks with different movement instructions for target words. If respondents shifted their response criterion when movement instructions for target words were changed (e.g., from moving all target words toward the participant's name to moving them away), responses to positive and negative distractor words would change in the same direction, yielding a main effect of movement instructions for target words on response times to distractor words.

In sum, when analyzing response times to *distractor words*, an interaction of movement instructions for target words with valence of distractor words does not indicate a criterion shift, but a main effect of movement instructions for target words does. To validate this reasoning we first report analyses of these interactions, followed by critical analyses of the main effects.

Interaction of movement instructions for target words with valence of distractor words on response times to distractor words

Positive distractor words were responded to faster when participants were instructed to move target words toward ($M = 642$ ms) rather than away ($M = 708$ ms) from their name, whereas negative distractor words were responded to faster when participants were instructed to move target words away ($M = 680$ ms) rather than toward ($M = 699$ ms) their name. This two-way interaction of movement instructions for target words with valence of distractor words was significant and in the same direction in each single study. Across studies: $z = 16.89$, $p \rightarrow 0$.

If a target word set is *homogeneous* in valence, for example if there are only positive target words, the above model predicts that only positive distractor words should be affected by movement instructions for target words. These predictions are reversed for a target word set composed only of negative words. For example, if there are only positive target words, it is sufficient to identify that a word is negative to know that it needs to be moved away, no matter what the movement instructions for target words are because there are no negative words that would ever be moved toward the name. As only two of the ten studies used homogeneously valenced target word sets, we also analyzed the data from three previous studies that used a different research design. For reasons of space the studies are not reported in detail here. Essentially, the data pattern in these experiments corresponded well to the predictions, with one noteworthy exception: one experiment with a positive target word set did show a stronger effect of movement instructions for target words on positive than negative distractor words, but contrary to expectation the direction of this effect was the same for both valences of distractor words and it was significant even for the negative distractors.

In sum, our reasoning about how respondents process distractor words is very well supported by these data. However, without having confirmed whether there is a main effect of movement instructions for target words on response times to distractor words, we cannot rule out the presence of criterion shifts.

Evidence for criterion shifts: Main effect of movement instructions for target words on response times to distractor words

Participants responded significantly more slowly to distractor words when they had to move target words away from ($M = 718$ ms) rather than toward their names ($M = 645$ ms), $z = 8.59$, $p \rightarrow 0$, which was significant in seven of eight studies. This is a clear indication that respondents shift the response criterion that they apply to all stimuli to more conservative levels in blocks where they are required to move target words away compared to those where they move target words toward their name. This finding is consistent with the idea that respondents may have a more conservative response criterion for away movements than for toward movements. The effect of this criterion shift is that uncalibrated response time difference scores are more positive than they should be. However, because EMA has single item sensitivity, the rank order of EMA scores is preserved. The error concerns only the neutral point of the scale. This is, of course, the positivity bias we observed earlier. Fortunately, the calibration procedure corrects for this bias as indicated by the high accuracy of the procedure. This discussion shows that the calibration procedure is essentially a de-biasing procedure that first estimates each individual participant's bias and then removes it.

Appendix D. Biasing effects of relearning movement instructions for target words

The design of the movement instructions for target words across blocks is ABBA for each individual respondent, but between-subjects there were two orders: $A_{\text{toward}}B_{\text{away}}B_{\text{away}}A_{\text{toward}}$ versus $A_{\text{away}}B_{\text{toward}}B_{\text{toward}}A_{\text{away}}$. It is conceivable that switching instructions the first time (i.e., from A-blocks to B-blocks) is more difficult than switching them the second time (i.e., from B-blocks back to A-blocks). Indeed, across all ten studies the response times to target words were significantly slower in B-blocks ($M = 637$ ms) than in A-blocks ($M = 621$ ms), $z = 4.54$, $p \rightarrow 0$. Even though this effect is relatively small, it does bias uncalibrated response time difference scores. Specifically, these scores will differ depending on whether target words in B-blocks need to be moved toward or away from the participant's name. Thus, these switching costs should bring about order effects of movement instructions for target words on uncalibrated response time difference scores. Specifically, when the first re-learning of movement instructions for target words is from toward to away movements ($A_{\text{toward}}B_{\text{away}}B_{\text{away}}A_{\text{toward}}$), rather than vice versa ($A_{\text{away}}B_{\text{toward}}B_{\text{toward}}A_{\text{away}}$), away movements will take longer, thus making uncalibrated response time difference scores more positive. This was indeed the case for 11 out of 33 uncalibrated response time difference scores. Consistent with our reasoning, significant order effects were present only in those studies where response times were significantly slower in B-blocks than in A-blocks. This effect introduces a slight positivity bias in one order condition and a slight negativity bias in the other, but it leaves the rank ordering of response time difference scores in place. Thus it again shifts the neutral point of the scale. Because the calibration procedure works by estimating these shifts in the neutral point, the order effects should disappear after calibration. After calibrating response time difference scores, that is, computing EMA scores, order effects should vanish. Indeed, after calibration only 2 of the 33 EMA scores were affected by order of movement instructions for target words. In sum, the cost of relearning movement instructions for target words can result in effects of the order in which these instructions are implemented. However, calibration removes these order effects. Thus, when such order effects are relevant, we recommend calibrating EMA scores or holding the order of movement instructions for target words constant.

References

- Asendorpf, J. B., Banse, R., & Schnabel, K. (2003). Fakability of an implicit association test (IAT) and a new implicit association procedure (IAP) for shyness. Unpublished manuscript.
- Banaji, M. R., Bazerman, M. H., & Chugh, D. (2003). How (un)ethical are you? *Harvard Business Review*, 81(12), 56–64.
- Bargh, J. A. (1997). The automaticity of everyday life. In R. S. Wyer, Jr. (Ed.), *Advances in social cognition* (Vol. 10, pp. 1–61). Mahwah, NJ: Erlbaum.
- Bargh, J. A., Chaiken, S., Govender, R., & Pratto, F. (1992). The generality of the automatic attitude activation effect. *Journal of Personality and Social Psychology*, 62, 893–912.
- Brendl, C. M. (1997, June). Annäherungs-Vermeidungsmotivation und Einstellungen implizit messen? Geschwindigkeit von Armbstreckung versus Armbeugung [Measuring approach-avoidance motivation and attitudes implicitly? Speed of arm extension and arm flexion]. *Talk at the 6th Meeting of the Social Psychology Division of the German Psychological Society (DGPs) in Konstanz, Germany*.
- Brendl, C. M., Markman, A. B., & Messner, C. (2001). How do indirect measures of evaluation work? Evaluating the inference of prejudice in the Implicit Association Test. *Journal of Personality and Social Psychology*, 81, 760–773.
- Brunel, F. F., Tietje, B. C., & Greenwald, A. G. (in press). Is the implicit association test a valid and valuable measure of implicit consumer social cognition. *Journal of Consumer Psychology*.
- Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin*, 25, 215–224.
- Cunningham, W. A., Preacher, C. J., & Banaji, M. R. (2001). Implicit attitude measures: consistency, stability, and convergent validity. *Psychological Science*, 12(1), 163–170.
- De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology*, 37, 443–451.
- De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental Psychology*, 50, 77–85.
- Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline?. *Journal of Personality and Social Psychology*, 69, 1013–1027.
- Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research. Their meaning and use. *Annual Review of Psychology*, 54, 297–328.
- Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology*, 50, 229–238.
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102, 4–27.
- Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, 109, 3–25.
- Greenwald, A. G., Draine, S. C., & Abrams, R. L. (1995). Three cognitive markers of unconscious semantic activation. *Science*, 273, 1699–1702.
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74, 1464–1480.
- INFORMATION, T.D.C.F.L. (1995). Celex (Version D2.5).
- Jacoby, L. L., Yonelinas, A. P., & Jennings, J. M. (1997). The relation between conscious and unconscious (automatic) influences: A declaration of independence. In J. D. Cohen & J. W. Schooler (Eds.), *Scientific approaches to consciousness*. New Jersey: Mahwah.
- Kahneman, D. (1999). Objective happiness. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 3–25). New York, NY: Russell Sage Foundation.
- Kaplan, K. J. (1972). On the ambivalence-indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique. *Psychological Bulletin*, 77, 361–372.
- Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology*, 81, 774–788.
- Kihlstrom, J. F. (2004). Implicit methods in social psychology. In C. Sanson, C. C. Morf, & A. T. Panter (Eds.), *The sage handbook of methods in social psychology* (pp. 195–212). Thousand Oaks, CA: Sage.
- Livingston, R. W., & Brewer, M. B. (2002). What are we really priming? Cue-based versus category-based processing of facial stimuli. *Journal of Personality and Social Psychology*, 82, 5–18.
- MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109, 163–203.
- Maison, D., Greenwald, A. G., & Bruin, R. (2001). The Implicit Association Test as a measure of implicit consumer attitudes. *Polish Psychological Bulletin*, 32, 61–69.
- Maison, D., Greenwald, A. G., & Bruin, R. (in press). Predictive validity of the Implicit Association Test in studies of brands, consumer attitudes, and behavior. *Journal of Consumer Psychology*.
- Markman, A. B., & Brendl, C. M. (in press). Constraining theories of embodied cognition. *Psychological Science*.
- McFarland, S. G., & Crouch, Z. (2002). A cognitive skill confound on the Implicit Association Test. *Social Cognition*, 20, 483–510.
- Münsterberg, H. (1892). Die psychologische Grundlage der Gefühle [The psychological basis of the emotions]. International Congress of Psychology, 2nd session, pp. 132–135.
- Nosek, B. A., & Banaji, M. R. (2001). The Go/No-go Association Task. *Social Cognition*, 19, 625–666.
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics*, 6(1), 101–115.
- Nosek, B. A., Greenwald, A. G., & Banaji, M. (2004). Understanding and using the Implicit Association Test: II. Methodological issues. Unpublished manuscript.
- Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science*, 12, 413–417.
- Olson, M. A., & Fazio, R. H. (2004). Reducing the influence of extrapersonal associations on the Implicit Association Test: Personalizing the IAT. *Journal of Personality and Social Psychology*, 86, 653–667.
- Ratcliff, R., & McKoon, G. (1989). Similarity information versus relational information: Differences in the time course of retrieval. *Cognitive Psychology*, 21, 139–155.
- Rosnow, R. L., & Rosenthal, R. (1991). If you're looking at the cell means, you're not looking at only the interaction (unless all main effects are zero). *Psychological Bulletin*, 110, 574–576.
- Solarz, A. K. (1960). Latency of instrumental responses as a function of compatibility with the meaning of eliciting verbal signs. *Journal of Experimental Psychology*, 59, 239–245.
- Swanson, J. E., Rudman, L. A., & Greenwald, A. G. (2001). Using the Implicit Association Test to investigate attitude-behaviour consistency for stigmatised behaviour. *Cognition and Emotion*, 15, 207–230.
- Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology*, 72, 262–274.