Do You Remember? Rater Memory Systems and Leadership Measurement

Tiffany Keller Hansbrough

Farleigh Dickinson University, United States thansb@fdu.edu

Robert G. Lord

Durham University Business School, Durham University, UK robert.lord@durham.ac.uk

Birgit Schyns

NEOMA Business School, France birgit.schyns@neoma-bs.fr

Roseanne J. Foti

Virginia Tech, United States rfoti@vt.edu

Robert C. Liden

University of Illinois at Chicago, United States bobliden@uic.edu

Bryan Acton

Virginia Tech, United States bacton@vt.edu

In Press, The Leadership Quarterly

Corresponding author: Tiffany Keller Hansbrough, Silberman College of Business, Fairleigh Dickinson University, 285 Madison Avenue, M-MS1-03, Madison, NJ 07940, United States.

This research was supported by the grant, “Episodic and Semantic Memory Effects on Leadership Measurement and Prediction of Leadership Outcomes (W911NF-18-1-0049) U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) awarded to the first, second and fourth authors.

We would like to thank Maggie North and George Pantovic for their assistance with the Study 6 data collection.

Abstract

Despite widespread concerns about the use of retrospective accounts of leader behavior and response tendencies associated with raters who tend rely on semantic memory, little attention has been devoted to developing methods that move measurement processes beyond those based on semantic memory to those based on episodic memory. The results from a series of six studies demonstrate a) questionnaire items can be classified in terms of their emphasis on episodic or semantic memory and the language used in items is associated with different types of memory processes, b) scales based on episodic memory have a greater association with trust than do scales based on semantic memory, c) the procedure that requires raters to indicate whether their response to each item is based on semantic or episodic memory dramatically reduces the impact of liking on leadership ratings, and d) the memory source intervention that encourages raters to rely on episodic memory reduces false alarms in leadership ratings. Taken together, these results demonstrate that rater memory systems are an important component of the leadership rating process and that consideration of the type of memory elicited during that process can be used to improve leadership measurement.

Keywords: episodic and semantic memory processes, measurement, leadership ratings

Do You Remember? Rater Memory Systems and Leadership Measurement

Much measurement theory focuses on the structure of rating scales, using factor analysis or measurement models in structural equation modeling to judge the adequacy of a measure. This emphasis on items and scales (whether latent or manifest) obscures the fact that the numbers being analyzed come from people, and their cognitive processes are an equally important aspect of measurement procedures. The importance of cognitive processes has been highlighted in the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014), as well as emphasized for some time by measurement researchers (Ercikan & Pellegrino, 2014), yet raters' cognitive processes tend to receive less attention, in part, because it is not clear what aspects of cognitive processes should be examined.

Drawing from the social cognition literature, research on raters has traditionally centered on the impact of schemas and heuristics that impact ratings. For example, early work on implicit leadership theories demonstrated that the factor structures of ratings based on a hypothetical leader were equivalent to those based on actual leaders (Eden & Leivathan, 1975; Rush, Thomas & Lord, 1977). Moreover, performance expectations or information about prior performance (i.e., the performance cue effect, Larson, Lingle, & Scerbo, 1984) can introduce a cue consistent bias into ratings (Lord, 1985; Rush et al., 1977; Staw, 1975). Finally, stereotypes have been associated with lower ratings as women are rated as less competent, less influential, and less likely to have played a leadership role than their male counterparts (Heilman & Chen, 2005). Missing from this emphasis on social cognition is insight into the rater memory processes that take place during ratings. Given that raters rely on retrospective judgments when completing ratings, the emphasis on memory is particularly salient. Further, Feldman and Lynch (1988) proposed that rater use of these types of heuristics is facilitated by the accessibility of *summary*

judgments rather than behavioral information. As we will discuss later, the use of summary judgments is associated with semantic memory processes, which we differentiate from episodic memory processes.

In line with the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014), the six studies reported here focus directly on the rater and the memory system cued by specific types of items as well as the impact of a memory source intervention upon retrieval. As we will show, this approach yields new insights as to what ratings mean and how leadership measurement can be improved. We examine these rater effects in the leadership domain, because it is widely recognized that followers are an important component of the leadership process, and there is a long history of concern with rater effects on leadership ratings (e.g., Arvey & Murphy, 1998; Eden & Leviatan, 1975; Hansbrough, Lord, & Schyns, 2015; Hoyt, 2000; Roberson, Galvin, & Charles, 2007; Rush et al., 1977). However, we expect that the benefits of focusing on raters as an important aspect of measurement are much broader than concerns with accuracy and bias in leadership ratings.

Many areas of leadership research, and transformational leadership in particular, have been harshly criticized in terms of measurement quality (van Knippenberg & Sitkin, 2013). The typical leadership study in this and other areas relies on retrospective follower accounts of leader behaviors as a means to measure leadership constructs (Hunter, Bedell-Avers, & Mumford, 2007). Despite widespread reliance on follower ratings, previous research has shown that follower factors, such as personality (Bono, Hooper & Yoon, 2012; Felfe & Schyns, 2010; Hansbrough, et al., 2015), implicit leadership theories (Eden & Leviatan, 1975; Rush et al., 1977; Weiss & Adler, 1981), and liking (Brown & Keeping, 2005; Martinko, Mackey, Moss, Harvey, McAllister, & Brees, 2018; Yammarino, Cheong, Kim, & Tsai, 2020) impact ratings of

leader behavior. Although leadership measurement is often aimed at assessing *leader* behaviors and traits, early research found that ratings of leadership typically reflect the *rater's* (typically followers) cognitive and emotional processing in addition to the leader's actual behaviors (Eden & Leviatan, 1975). Therefore, these factors create ambiguity in interpreting empirical findings because leadership measures are often used to predict other variables provided by followers (e.g., trust, job satisfaction, motivation, organizational citizenship behaviors).

In addition, many measures of leader behavior rely on retrospective frequency estimates of leader behavior which require raters to aggregate multiple observations across an unspecified prior period of time comprised of multiple events (Shondrick, Dinh, & Lord, 2010). When asked to make judgments about another person, observers tend to rely on an easily accessible general impression or cognitive categorization of leaders (i.e., semantic memory) without reviewing the specific behaviors (i.e., episodic memory) on which those impressions are based (Srull & Wyer, 1989), in part, because it is unclear which prior behaviors are relevant.

Despite general awareness of such issues, there are few practical means to improve leadership measures or reduce potential confounds. One possibility suggested by Shondrick et al. (2010) is to develop leadership measures that emphasize episodic rather than semantic memory. Supporting this recommendation, Martell and Evans's (2005) work showed that training raters to rely on episodic memory reduced the impact of the performance cue effect, although it did not increase memory accuracy. Nevertheless, considering whether an item is primarily based on semantic or episodic memory may be a helpful first step to increase the validity of measures of leader behavior (Hansbrough et al., 2015).

Episodic memory is a context-specific memory for events and personal experiences which often has an emotional basis and is integrated with information about the self in the

referenced context; *semantic memory*, in contrast, provides a context-independent source of general knowledge regarding objects, facts, word meanings, etc., which tends to be the default mode of processing information due to its general relevance and ease of use in guiding retrospective judgments. However, the rich association of impressions in semantic memory also provides a conduit for many non-behavioral influences to affect ratings, clouding the interpretation of behavioral correlates. Consequently, the type of memory system used by raters may have important implications for the quality of the obtained leader ratings. Thus, there is a compelling need to move measurement processes beyond those based on semantic memory to those based on episodic memory.

Addressing this need, Studies 1 and 2 compare episodic and semantic transformational leadership items in terms of their relation to trust. Next, Studies 3, 4, and 5 replicate our results using servant leadership items and extend our findings to an organizational setting. Finally, in Study 6 we present the results of an experiment in order to directly examine the impact of episodic memory on rating accuracy and bias.

In the following sections, we first discuss memory processes, carefully distinguishing between semantic and episodic memories, examine how this distinction relates to accuracy and bias in ratings, and describe a novel application of a technique used in the memory literature to distinguish between episodic and semantic memory. Subsequently, we address how language affects social cognition using the construal level theory by Semin and Fiedler (1991), which focuses on the abstractness of verbs used to characterize people. This literature provides the basis for our argument that the language used in questionnaire items, which tends to emphasize one memory source over another, affects the relationship with outcomes.

Memory and Language Effects on Leadership Ratings

Semantic and Episodic Memory

Semantic memory. Most measures of leader behavior are developed to be general and independent of a specific work context and thus draw on semantic memory. As mentioned previously, typically the response alternatives of such measures are based on frequency estimates of previous leader behavior that require raters to generate descriptions of typical leader behavior (Hunter et al., 2007; Shondrick, et al., 2010). Furthermore, measures that require respondents to indicate the extent to which they agree with a statement about leader behavior are also summary evaluations and therefore promote reliance on heuristics and generalized impressions associated with semantic memory. Semantic memory includes general, non-contextual representations of a person in terms of trait constructs and other broadly applicable schemas, which are used as a template to process and interpret new information about that person. For example, people may rely on general schemas, such as implicit personality theories and implicit leadership theories when they have incomplete information about another person (Fiske & Taylor, 2013). Reliance on schemas often involves a pattern completion process (Smith & DeCoster, 2000), and patterns are a critical determinant of leadership perceptions (Foti & Hauenstein, 2007). More specifically, the general schemas that underlie leadership and other categories reflect patterns that typically are activated as a whole, with pattern completion processes operating at a preconscious level (Hanges, Lord, & Dickson, 2000; Lord, Brown, Harvey, & Hall, 2001; Smith & DeCoster, 2000). Thus, even when behaviors relevant to particular questionnaire items were not previously observed, raters can still rely on semantic memory to rate behavior if the item relates to a general pattern that characterizes the ratee.

One unintended consequence of measures based on summary judgments of leader behavior is that they may create false memories as raters may rely on overall global evaluations

and patterns to fill in the gaps of their memories instead of using specific leader behaviors they observed (Hansbrough et al., 2015; Shondrick et al., 2010). Such memories may also have a non-behavioral basis, such as inferences based on performance outcomes (Lord & Maher, 1991). Therefore, raters may endorse items that seem familiar (i.e., fit the pattern of behaviors typically associated with leaders) and are consistent with their general impressions even though these items describe behaviors that did not actually occur. Because raters may rely on patterns and social categories that are grounded in their cognitive schemas instead of specifically remembered leader behaviors, typical measures of leader behavior often reflect individual rater information processing and general knowledge structures rather than recall of actual leader behaviors (Hunter et al., 2007). Items that center on summary evaluations or trait based attributions should also be prone to the effects of pattern completion processes.

Episodic memory. In contrast, episodic memory is context dependent and consists of rich, vivid details that include when the event occurred and the emotions experienced during encoding (Allen, Kaut, & Lord, 2008; Tulving, 2002). Because episodic memory is based on a specific event, it enables perceivers to provide an explanation or justification for their conclusion. This is different from semantic memory, based on a generalized impression, which renders perceivers unable to provide any justification for their answer other than intuition (Smith & DeCoster, 2000). Importantly, episodic memory is the *only* kind of memory system that allows individuals to consciously re-experience the past events; therefore, it is oriented to the past in a way unlike other memory systems (Tulving, 2002). Consequently, tapping into episodic memory seems particularly relevant for obtaining accurate retrospective accounts of leader behavior.

Because episodic memory is tied to the context in which it occurred, it is more likely to activate memories of events and behaviors that *actually occurred* (Shondrick et al., 2010).

Consequently, as maintained by other researchers, ratings based on episodic memory may be more accurate than those based on semantic memory (Foti & Lord, 1987; Martell & Evans, 2005; Shondrick et al., 2010). In addition, because episodic memories locate the self in this context, items that focus on personally relevant behaviors in either transformational or servant leadership measures may be more likely to be associated with episodic memory processes. Therefore, we contend that ratings based on episodic memory may also be more strongly associated with personally relevant outcomes related to such events.

Like semantic memory, episodic memory is reconstructive (Addis & Schacter, 2012) and the reconstruction can draw from semantic memory as well as other sources of information about the event. Thus, prior episodes are not reinstated, but rather recreated. Each time a memory is retrieved, it is re-encoded by the hippocampus, therefore, the older the memory, the more traces there are of that memory and the more opportunity for that memory to be retrieved (Nadel & Moscovitch, 1997). However, recent memories are likely to be retrieved in greater detail. In situations where two or more memories can potentially be reconstructed, the memory that is retrieved is determined by the availability of the stored information, the cues that are present, and the task demands (Winocur, Moscovitch, & Bontempi, 2010). Hence, episodic memory is subject to potential distortions in memories as illustrated by the eyewitness literature (Wright & Loftus, 2008). For example, imagining based on episodic memory components can be confused with remembering the past which produces imagination-based errors (Schachter, Addis, Hassabis, Martin, Spreng, & Szpunar, 2012). Furthermore, episodic memory does not preclude the possibility that individuals may selectively attend to events consistent with stereotypes which are subsequently encoded into memory. However, it should be stressed that semantic memory also is susceptible to similar influences and may even accentuate potential biases because of the many

potential associations with general semantic information. Indeed, race (Rosette, Leonardelli, & Phillips, 2007), gender (Scott & Brown, 2006), or even performance information (Lord, 1985) likely impact ratings via the use of schemas and prototypes associated with semantic memory. In contrast, ratings based on episodic memory may be less impacted by such biases or gap filling processes.

Use of Memory Metacognitions to Distinguish Different Types of Memory

Tulving (1985) argues that raters have metacognitions about memory and are able to differentiate between ratings based on semantic memory, which are referred to as *know judgments,* and those based on episodic memory, which are referred to as *remember judgments.* Although ratings of leader behavior are likely informed by both types of memory, it should be noted that semantic and episodic memory are independent; they are processed in different regions of the brain and are subject to different age-related declines (Allen et al., 2008). Episodic memories may emphasize a vivid, initial encodings of events and context, whereas semantic memory incorporates the consolidation of a schematic version of events that incorporate the gist of events but lacks contextual details (Winocur et al., 2010). Taken together, this research suggests that determining whether an item tends to elicit a remember or know judgment may be a helpful component of measurement development. It should be noted that unlike leadership ratings, asking people to report whether their response to each item is based on a remember or know judgment is not a retrospective judgment. Instead people are reporting on the memory process they are using *in the moment* while completing leadership ratings. Whether an item produces a remember or know response may also depend on the type of language used in an item, an issue addressed in the following section.

Language and Memory

Leadership ratings provide a salient example of a task that elicits processing based on the meaning of words, and that meaning may vary depending on the abstractness of the verbs used to describe behavior. According to the linguistic category model (Semin & Fiedler, 1991), words can be represented on a continuum of concreteness to abstractness that ranges from concrete behaviors to global, general representations of traits (Jiga-Boy, Clark, & Semin, 2013). For example, the verb "helps" is more concrete than the adjective "helpful". Concrete words provide information about specific details and behaviors (e.g., observers may recall a specific instance when the target helped them). In contrast, abstract adjectives only provide a general impression of the target (e.g., the target is characterized as generally helpful independent of the context). Semin and Fiedler contend that different types of words can trigger different cognitive processes including differential memory effects. For example, it has long been known that concrete words are better remembered than abstract words (i.e., the concreteness effect, Marschark & Cornoldi, 1991; Paivio, 1986; 1991; 1995). Decades of research suggest that the concreteness effect is a robust phenomenon that applies to many different tasks, stimulus types, and memory tests (ter Doest & Semin, 2005). Here, we contend that more concrete items tend to be associated with episodic memory, whereas more abstract items tend to be associated with semantic memory.

Some measures of leader behavior are likely comprised of a mixture of concrete and abstract items. For example, van Knippenberg and Sitkin (2013) observe that measures of transformational leadership typically confound the measurement of leader behavior with attributions of its effects. The MLQ 5X (Bass & Avolio, 1996) includes items that focus on abstract attributions of charisma as well as more behaviorally based items that focus on problem solving, individual attention, and coaching of followers (van Knippenberg & Sitkin, 2013). Similarly, servant leadership scales include a variety of abstract items that focus on trait based

attributions and generalized impressions as well as items that focus on specific leader behaviors, and for this reason, a servant leadership measure was chosen to replicate results from Studies 1 and 2. Here, we argue that different types of items are associated with different types of memory. In particular, abstract items may prompt individuals to rely on generalized impressions or leadership schemas associated with semantic memory, whereas concrete items may prompt individuals to retrieve events that actually happened and therefore may be associated with episodic memory.

Hypothesis 1: The language used in leadership scale items is associated with different types of memory such that a) more abstract items are associated with semantic memory and b) more concrete items are associated with episodic memory.

The Nature of Criteria

In addition to the nature of items, the nature of the leadership criteria may also impact whether individuals rely on semantic or episodic memory. An outcome of leadership that is particularly relevant for this distinction is trust as the decision to trust another individual is likely based on an assessment of that individual's previous behaviors (Lindskold, 1978). From the observer's perspective, the decision to trust is based on "what we take to be 'good reasons' that provide compelling evidence of trustworthiness" (Lewis & Wiegart, 1985, p. 970). Moreover, trust serves as an emotional marker of relationship interdependence (Fiske, Lin, & Neuberg, 1999), and emotional reactions promote the use of episodic memory (Allen et al., 2008). Hence as outlined above, it is likely that the decision to trust is based on concrete leader behaviors that implicate the self and are encoded via episodic memory processes. Consequently, we expect that transformational leadership and servant leadership scales that emphasize episodic memory to

provide better estimates of trust than transformational or servant leadership scales that emphasize semantic memory.

Hypothesis 2: The leadership scale items based on episodic memory provide better estimates of trust than the leadership scale items based on semantic memory.

Effects of Affect and Metacognitive Processes on Memory Search Processes

Affect can pertain to either a general evaluation, such as liking, or to a specific, emotional reaction to a person in the context of a particular event. The former is used to characterize a person as a whole (e.g., generalized impression), whereas the latter pertains to how a person acted in specific events, which are nested within the person (e.g., person-parts). Lord and Dinh (2012) address this issue by noting that event level processes are nested within leaders (see also Hall & Lord, 1995, for the parts/wholes distinction in further detail). Here we consider liking as a holistic generalized impression (e.g., person wholes) that may be associated with semantic memory processes.

Srull and Wyer (1989) contend that person perception begins when observers categorize a target as likable or dislikable (in their terms, form a general evaluative concept), which then serves as an interpretive structure to make sense of the target's future behaviors. Subsequently, behaviors are recalled in a sequential search process that begins with this affective evaluation, and if this evaluation is easily retrieved, the search for specific behaviors may be curtailed unless individuals are motivated to engage in more effortful processing and have the cognitive resources available to do so (Fiske et al., 1999; Gilbert, Pelham, & Krull, 1988). Indeed, Martinko et al. (2018) contend that the rating situation is unlikely to prompt followers to engage in controlled processing. Instead, individuals often may rely on liking when they rate leaders. As noted previously, it is possible that asking subjects to indicate whether their rating for each item

was a remember or know judgment may prompt them to think more carefully about their responses which may, in turn, reduce the impact of liking on leadership ratings. Therefore, as we will examine by comparing Studies 1 and 2, the impact of liking on leadership ratings should be diminished when using this type of metacognitive instruction.

Hypothesis 3: The remember/know judgment rating procedure reduces the impact of liking on leader ratings.

Consequently, in Study 1 and Study 2 we use this metacognitive procedure as a potential experimental effect that may reduce the impact of liking on ratings. Specifically, we conducted two studies with precisely matched samples, one which included remember/know judgments as part of the rating process (Study 1) and a second study which collected leadership ratings in a typical manner. To do this, we collected the data for Studies 1 and 2 at the same time, and randomly assigned subjects to one study or the other. This procedure allowed us to gauge the potential effects of this memory metacognitive procedure on rating outcomes and test the assertion that Study 1 ratings would be less dependent on liking that Study 2 ratings. Study 2 was also used to replicate the primary findings from Study 1.

Study 1: Methods

Participants and Procedure

Using a crowdsourcing site, Amazon Mechanical Turk (MTurk), we recruited 300 participants. The criteria for participation was as follows: Participants were required to be U S. citizens, working full time, and, as a quality control mechanism, to have had over 95% of their previous assignments on MTurk accepted by requesters. We pre-tested the survey in order to provide an estimated completion time (30 minutes). The obtained average completion time of 26 minutes compared favorably to that of the pre-test. 10 surveys completed in less than 10 minutes

were excluded from the sample, resulting in a total sample size of 290 (157 males and 133 females).

The average participant was 33 years old and had been working for his/her current supervisor for 3.7 years. In terms of educational attainment, 174 participants had a BA/BS degree, 29 had a master's or terminal degree, and 87 had a high school diploma. Participants reported a wide variety of occupations including manager, sales representative, accountant, teacher, and administrative assistant. Participants rated their current leader as described below and completed the dependent variables.

As detailed below, following Tulving's (1985) methodology, participants were provided with definitions of remember and know judgments and instructions for completing leader ratings. To minimize the possibility that participants might consider one type of judgment more desirable than the other, it was stressed that both types of judgments are useful and the judgments do not differ in terms of their confidence or certainty. The instructions read as follows:

> "We have two different ways that we make judgments about other people, remembering and knowing. **Remembering is based on a vivid recollection of a specific event**. For example, we might describe someone as outgoing because we can recall specific examples of their behavior. Alternatively, **knowing is based on a general feeling or impression about a person**. It is important to note that both types of memory are useful and that one is not inherently better than the other. Moreover, remember and know judgments do not differ in terms of their confidence or certainty. For example, we can be equally confident about a judgment even though we might not associate it with a specific event. For each of the following items please rate your supervisor and then using the definitions above indicate whether your rating reflects a remember or a know judgment."

Coding of concreteness/abstractness. For the transformational leadership scale, we used the Linguistic Category Model (LCM) (Semin & Fiedler, 1991) to examine the level of abstraction of the items identified by participants as being either remember or know judgments. According to LCM, abstractness increases with the generality of verbs used in describing people or behavior, and it is most abstract when people are described by adjectives. Specifically, abstractness increases with each of the following five types of descriptions: a) Descriptive Action Verbs (DAV) that refer to observable behavior; b) Interpretive Action Verbs (IAV) that refer to a general class of behaviors that require interpretation beyond the description; c) State Action Verbs (SAV) that refer to the momentary emotional consequences of an action; d) State Verbs (SV) that refer to enduring emotional states; and e) Adjectives (ADJ) which are dispositional traits. These distinctions are then used to code the concrete/abstractness of verbal descriptions on a 1 to 5 scale where high scores describe an action or trait in more global or general terms. As demonstrated by previous research, LCM is a valid methodology to measure construal level (Freitas, Gollwitzer, & Trope, 2004; Fujita, Trope, Liberman, & Levin- Sagi, 2006).

Measures. *Transformational leadership* was assessed using the 36-item short form of the Multifactor Leadership Questionnaire (MLQ 5X, Bass & Avolio, 1996). Participants indicated on a five point scale (0-4) ranging from “not at all” to “frequently, if not always” how frequently each item fit their supervisor. The instructions were modified as described above; namely for each item participants rated their supervisor and then indicated whether each rating reflected a remember or know judgment. As described in the results section, episodic and semantic subscales were created based on whether items were identified by a majority of participants as being remember or know judgments. C*ognition-based trust* was assessed with a 5-item scale

(McAllister, 1995). A sample item is "My direct supervisor approaches his/her job with professionalism and dedication." *Affect-based trust* was assessed with a 5-item scale (McAllister, 1995). A sample item is "We have a sharing relationship. We can both freely share our ideas, feelings, and hopes." Engle and Lord's (1997) 4-item measure of *liking* was used to assess the degree to which individuals liked their immediate supervisor. Participants indicated on a five point scale ranging from strongly disagree to strongly agree the extent to which they agreed with the statements about their supervisor. All scale reliabilities are reported in Table 2.

Study 1: Results

Remember Versus Know Judgment of Items

Table 1 presents data pertaining to remember/know judgments for the MLQ items that tended to emphasize one memory source over the other. Twenty-four of the 36 items on the MLQ 5-X and 14 of the 20 transformational leadership items were considered either remember or know judgments by a clear majority (55% or greater) of the respondents. Percentages were rounded to the nearest whole percent. Moreover, the items that participants classified as know judgments, such as "instills pride in me for being associated with him/her" and "goes beyond self-interest for the good of the group" primarily centered on general impressions of leader charisma. In contrast, the items that participants classified as remember judgments, such as "seeks different perspectives when solving problems" and "spends time teaching and coaching", were behaviorally based. Therefore, items that were more behaviorally based or self-referent may tap into episodic memory.[1]

[1] We also examined whether the propensity to rely on remember or know judgments across all responses may be a function of individual differences (personality and PANAS were also measured but tangential to the focus of this study). To do so we created a proportion of the number of remember responses over all of the items divided by the total of number of scale items. This proportion was then regressed on individual differences. None of the individual

Abstractness of Remember Versus Know Items

Following the coding LCM procedures described previously (Semin & Fiedler, 1991), the first and second authors coded each of the 14 transformational leadership items identified by participants as being either a remember or know judgment in terms of abstractness and compared the ratings of the first/second author to that of a second rater who was not involved in this study (Cohen, 1960, Kappa = .58, p = .000). We anticipated that remember judgments would reflect memories of concrete events, whereas know items would be more abstract. The average abstractness for items identified as remember judgments was 1.8 whereas the average value for items identified as know judgments was 3.17 ($t = 2.62$, $p = .04$ equal variances not assumed as per the results of Levene's test for equality of variances). Thus, consistent with Hypothesis 1, the items that were classified as being know judgments were significantly more abstract than were the items that were classified as being remember judgments.

Episodic and Semantic Transformational Scales and Descriptive Statistics

We created *episodic* and *semantic* scales of transformational leadership based respectively on the items identified as being remember judgments (items # 8, 9, 13, 15, 19, 30, 31, and 32; $\alpha = .89$) and the items identified as being know judgments (items # 2, 10, 14, 18, 23, and 34; $\alpha = .87$). Table 2 reports descriptive statistics and correlations among these scales, transformational leadership items that were not classified as being remember or know judgments, and other key measures. Further, as might be expected, liking was most strongly related to affect-based trust ($r = .82$, $p = .000$), and it showed strong relations to both episodic ($r = .69$, $p = .000$) and semantic ($r = .63$, $p = .000$) transformational leadership scales.

---

differences was significantly associated with the propensity to rely on remember or know judgments (see Appendix Table A1).

The correlations between the episodic scale and semantic scale with cognition-based and affect-based trust are considered correlated correlations because they share the same criterion variable (i.e., cognition-based trust and affect-based trust). Therefore, we used Hotelling's t in order to test if these correlations were significantly different. The correlation between episodic transformational leadership and cognition-based trust was significantly different from the correlation between semantic transformational leadership and cognition-based trust ($r = .70$ vs. $r = .66$, $t = 2.033$, $p = .043$. The correlation between episodic transformational leadership and affect-based trust was also significantly different from the correlation between semantic transformational leadership and affect-based trust ($r = .74$ vs. $r = .69$, $t = 2.694$, $p = .007$).

Episodic and Semantic Transformational Leadership Scales and Trust Ratings

To test Hypothesis 2, using SPSS software, we entered the episodic and semantic transformational leadership scales into a regression to examine their unique contribution to trust. As shown in the left half of Table 3, the episodic scale was more useful than the semantic scale for estimating both cognition-based trust ($\beta = .52$, $p = .000$ vs. $\beta = .20$, $p = .02$,) and affect-based trust ($\beta = .62$, $p = .000$ vs. $\beta = .14$, $p = .09$). As suggested by prior literature (Hansbrough et al., 2015; Shondrick et al., 2010) and Hypothesis 2, the transformational leadership scale that was more strongly based on episodic memory provided better estimations of outcomes than the more

semantically-based scale. [2] [3] [4] It is also noteworthy, that the episodic scale had a somewhat stronger relation to liking and was more strongly associated with affect-based trust. This may reflect the role of affect in integrating the contextual elements of an event into a coherent episodic memory as theorized by Allen et al. (2008).

In order to test whether the difference between the episodic and semantic beta weights for cognition-based trust was statistically significant, their corresponding 95% confidence intervals were estimated via bias correcting bootstrap (1,000 re-samples). The confidence intervals overlapped by less than 50%, therefore the difference between beta weights was statistically different ($p < .05$, Cumming, 2009, see Appendix Figure A1). Specifically, half of the average of the overlapping confidence intervals was calculated (.085) and added to the episodic beta weight lower bound estimate (.349) which was .434. As the semantic upper bound estimate of .376 did

[2] We conducted Item Response Theory (IRT) analyses (Embretson & Reise, 2000; Scherbaum, Finlinson, Barden, & Tamanini, 2006) in order to gain further insights into underlying psychological factors. The results showed that both sets of items demonstrate appreciable relationships with the latent construct of transformational leadership and provide considerable information about transformational leadership across the construct. Although the differences between the episodic and semantic scales were small, the episodic scale outperformed the semantic scale in terms of item discrimination and item information. The results regarding the relative efficiency analysis indicate that the episodic scale functions as it were forty percent longer when, in fact, it is only thirty-three percent longer. Therefore, the episodic scale outperforms the semantic scale in terms of both measurement precision and relationships with relevant criteria (see Appendix Tables A4-5).

[3] We also created weighted episodic and semantic scales whereby each item was weighted by the percentage of the respondents who had indicated the item was a remember or know item. The results using weighted scales did not significantly differ from the results reported here (see Appendix Table A6).

[4] Given that there are well-known gender differences in memory processing as it relates to episodic and semantic memory (e.g., Guillem & Mograss, 2005; Herlitz, Nilsson, & Backman, 1997), we conducted additional analyses controlling for gender in Studies 1-4 where there were a proportionate number of males and females. Gender was not significant in 5 of the 6 analyses. In the analysis where gender was significant, the addition of gender did not change the beta weights reported in the text (see Appendix Tables A -3).

not exceed the value of .434, the difference between the episodic and semantic beta weights was statistically significant.

The same procedure was used to test whether the difference between the episodic and semantic beta weights for affect-based trust was statistically significant. Again, the confidence intervals overlapped by less than 50% therefore the beta weights were statistically different ($p < .05$, Cumming, see Appendix Figure A2). Specifically, half of the average of the overlapping confidence intervals was calculated (.097) and added to the episodic beta weight lower bound estimate (.423) which was .520. As the semantic upper bound estimate of .337 did not exceed the value of .520, the difference between the episodic and semantic beta weights was statistically significant.

Study 2: Methods

Given that we used a procedure in Study 1 that was different than the standard rating procedure; it is possible that the effects we found may have been due to the particular procedure used to measure metacognitions, which undoubtedly forced subjects to think more carefully about the rating process. Consequently, Study 2 was designed to address this possibility, as it was conducted at the same time as Study 1, but did not ask for remember/know judgments after rating each MLQ 5X leadership item.

Participants and Procedure

We recruited 300 participants using a crowdsourcing site, Amazon Mechanical Turk (MTurk). The criteria for participation and sampling procedure were the same as in Study 1, which was conducted at the same time with random assignment to either Study 1 or 2. Eleven surveys completed in less than 10 minutes were not included in the sample, resulting in a total sample size of 289 (145 males and 144 females).

The average participant was 32 years old and had been working for his/her current supervisor for 3.35 years. In terms of educational attainment, 169 participants had a BA/BS degree, 36 had a master's or terminal degree, and 84 had a high school diploma. Participants reported a wide variety of occupational fields and jobs, including information technology, engineering, sales, education, customer service, and reference librarian.

Measures

Transformational leadership. As in Study 1, transformational leadership was assessed using the 36-item Multifactor Leadership Questionnaire (MLQ 5X, Bass & Avolio, 1996) short form. However, this time, participants only indicated on a five point scale (0-4) ranging from "not at all" to "frequently, if not always" how frequently their supervisors showed the indicated behavior and did not indicate whether each item was a remember or know judgment. Participants also completed the same *Cognition and Affect-based trust* (McAllister, 1995), and *Liking* (Engle & Lord, 1997) measures as described in Study 1. All scale reliabilities and correlations are reported in Table 4.

Study 2: Results

Replication of Study 1

The purpose of Study 2 was to replicate the episodic and semantic transformational scale results, even though we did not ask participants to make remember/know judgments in this study. To do so, we first created episodic and semantic transformational leadership scales based on the items identified by participants from Study 1 as being remember or know judgments.

As would be expected with equivalent samples, Table 4 shows that correlations among variables were highly similar to those in Study 1 as shown in Table 2. Episodic and semantic transformational scales were highly correlated ($r = .92$, $p = .000$), and liking was strongly related

to the episodic scale (r = .78, p = .000), the semantic scale (r = .75, p = .000), and both of the trust scales (both r's = .82, p = .000).

As in Study 1, we used Hotelling's t to compare the correlated correlations. The difference in correlations between episodic transformational leadership and cognition-based trust compared to the correlation between semantic transformational leadership and cognition-based trust (r = .76 vs. r = .73) was t = 1.964, p = .051. The correlation between episodic transformational leadership and affect-based trust was significantly different form the correlation between semantic transformational leadership and affect-based trust (r = .76 vs. r = .72, t = 2.608, p = .010).

Next, we entered each scale into a regression analysis to examine their unique values in estimating cognitive and affect-based trust, respectively. As in Study 1, the episodic transformational scale provided stronger estimations than the semantic transformational scale for both cognition (β = .55, p = .000 versus β = .22, p = . 02, see the right half of Table 3) and affect-based trust (β = .62, p = .000 versus β = .16, p = .11), respectively. These results closely replicate those of Study 1. Thus, the results obtained in Study 1 appear to be driven by the memorial qualities of items rather than the between study differences involving asking for remember/know judgments for each MLQ item in Study 1 but not Study 2.

As in Study 1, we tested whether the difference between the episodic and semantic beta weights for cognition-based trust was statistically significant. The confidence intervals overlapped by less than 50%, therefore the beta weights were statistically different (p < .05, Cumming, 2009, see Appendix Figure A3). Specifically, half of the average of the overlapping confidence intervals was calculated (.101) and added to the episodic beta weight lower bound estimate (.347) which was .448. As the semantic upper bound estimate of .41 did not exceed the

value of .448, the difference between the episodic and semantic beta weights was statistically significant.

The same procedure was then used to test whether the difference between the episodic and semantic beta weights for affect-based trust was statistically significant. Again, the confidence intervals overlapped by less than 50% therefore the beta weights were statistically different ($p < .05$, Cumming, see Appendix Figure A4). Specifically, half of the average of the overlapping confidence intervals was calculated (.093) and added to the episodic beta weight lower bound estimate (.43) which was .523. As the semantic upper bound estimate of .339 did not exceed the value of .523, the difference between the episodic and semantic beta weights was statistically significant.

Effect of Metacognitive Procedure on Liking

We also considered the possibility that the procedure used in Study 1, which asks raters to indicate whether their response to each an item was based on a remember or know judgment, might encourage raters to more carefully consider their responses. In particular, as stated in Hypothesis 3, it seems plausible that taking additional time to consider one's response may increase the availability of other leader behaviors and therefore reduce the impact of liking on ratings of transformational leadership. We, therefore, compared the results of Study 1 and Study 2 to determine if the procedure used in Study 1 reduced the impact of liking on ratings of transformational leadership using all items. We found that, as compared to Study 2, where the correlation between liking and transformational leadership was $r = .80$ ($p = .000$), the Study 1 procedure significantly reduced the correlation between liking and ratings of transformational leadership to $r = .69$, ($p = .000$). The difference between the obtained correlations in Study 2 and Study 1 was significant ($z = 3.004$, $p < .01$). This result is important because it suggests that

procedures that promote more controlled processing can reduce the effects of liking on leadership ratings. However, this shift in processing does not appear to affect the general finding that episodic items are better at estimating personally relevant outcomes.

Discussion of Transformational Leadership Findings

Taken together, Studies 1 and 2 address the important issue regarding how different types of memory and language impact relationships with outcomes. Our work demonstrates that a) transformational leadership questionnaire items can be classified in terms of their emphasis on episodic versus semantic memory and b) the language used in items is associated with different types of memories; c) scales based on episodic memory provide better estimates of trust than scales based on semantic memory; and d) the procedure which requires raters to indicate whether their response to each item is a remember or know judgment significantly reduces the impact of liking on leader ratings.

One concern with Studies 1 and 2 was that they only used transformational leadership items. Thus, we conducted three additional studies using servant leadership items that replicated these results. Furthermore, we used a different measure of trust in order to illustrate that these results are not dependent upon particular trust measures. We turn now to the replication studies using servant leadership.

Following the design for our first two studies, Studies 3 and 4 were conducted with precisely matched samples, Study 3 included remember/know judgments as part of the rating process, whereas Study 4 collected leadership ratings in a typical manner. Data for Studies 3 and 4 were collected at the same time and subjects were randomly assigned to one study or the other.

Study 3: Methods

Participants and Procedure

Two hundred and seventy-one participants were recruited using MTurk. The criteria for participation were the same as described in Study 1 and 2. Fourteen participants were excluded to due to incomplete data, an identifiable response pattern, or a very rapid completion time (i.e. less than 1/3 of the mean completion time), resulting in total sample size of 257 (119 females and 137 males).

The average participant was 37 years old. One hundred and thirty-six participants had an undergraduate degree, 35 participants had a master's or terminal degree, 70 participants had either a 2 year degree or some college, and 16 participants had a high school diploma. Participants reported a wide range of occupations including sales, information technology, management, education, and financial analyst.

As described in Study 1, participants were provided with definitions of remember and know judgments and instructions for completing leader ratings. We used LCM (Semin & Fiedler, 1991) as described in Study 1 to examine the level of abstraction of the items identified by participants as being either remember or know judgments.

Measures

Servant leadership was assessed using Liden, Wayne, Zhao, and Henderson's (2008) 28-item scale. The instructions were modified as described in Study 1. Participants indicated on a seven point scale the extent to which they agreed with each statement about their manager. A sample item is "My manager sacrifices his/her own interests to meet my needs". *Trust* was assessed with a 7-item scale adapted from Robinson (1996). The wording was changed slightly to refer to "my manager" instead of "my employer". Participants indicated on seven point scales the extent to which they agreed with each statement about their manager. All scale reliabilities are reported in Table 6.

Study 3: Results

Remember Versus Know Judgment of Items

Table 5 presents data pertaining to the remember/know judgments for the servant leadership items that tended to emphasize one type of memory over the other. Eighteen of the 28 servant leadership items were considered either remember or know judgments by a clear majority (55% or greater) of the respondents. For example, participants classified "My manager wants to know about my career goals" as being a remember judgment; whereas participants classified "My manager is always honest" as a know judgment.

Abstractness of Remember Versus Know Judgments

The first and second author coded each of the 18 items identified by participants as either a remember or a know judgment as specified by LCM (Semin & Fieldler, 1991). The ratings of the first/second author were then compared to that of another individual who was not involved in the study (Kappa = .47, p = .001). As in Study 1, we anticipated that the items identified as remember judgments would be more concrete than items identified as being know judgments. The average abstractness for remember items was 2.86 whereas the average abstractness for know items was 3.82 (t = 1.64, p = .12 equal variances not assumed). Thus, the items that were classified as being know judgments were somewhat more abstract than were the items identified as being remember judgments, although this difference was not significant.

Episodic and Semantic Servant Leadership Scales

We created episodic and semantic scales of servant leadership based respectively on the items that were identified as being remember judgments (items #8, 9, 16, 17, 20, 22, and 25; α = .89) and the items identified as being know judgments (items #3, 4, 6, 10, 11, 14, 18, 19, 26, 27, and 28; α = .91). Notably, the items identified as being remember judgements centered on

solving work problems and decision making whereas the items identified as being know judgments centered on care and concern for the follower as well as general impressions of leader honesty. Thus, the content of the items identified as being remember judgments was not more focused on leader trust than was the content of the items identified as being know judgments. Correlations among all variables including servant leadership items that were not classified as remember or know judgments are depicted in Table 6. As in the previous studies, we used Hotelling's t to compare the correlations between the episodic and semantic leadership scales and trust. The correlation between episodic servant leadership and trust was significantly different from the correlation between semantic servant leadership and trust ($r = .80$ vs. $r = .71$, $t = 3.591$, $p < .001$).

Episodic and Semantic Scales of Servant Leadership and Trust Ratings

To test Hypothesis 2, we entered the episodic and semantic servant leadership scales into a regression equation in order to examine whether their unique association with trust. As shown in Table 7, as predicted, the episodic scale provided a better estimates of trust than did the semantic scale ($\beta =. 62$, $p = .000$, vs $\beta =. 24$, $p = .000$).

As in the prior studies, we tested whether the difference between the episodic and semantic beta weights was statistically significant. The confidence intervals overlapped by less than 50%, therefore the beta weights were statistically different ($p < .05$, Cumming, 2009, see Appendix Figure A5). Specifically, half of the average of the overlapping confidence intervals was calculated (.057) and added to the episodic beta weight lower bound estimate (.497) which was .554. As the semantic upper bound estimate of .343 did not exceed the value of .554 the difference between the episodic and semantic beta weights was statistically significant.

Study 4: Methods

Study 4 was conducted to allow for the possibility that the procedure used in Study 3, which encouraged subjects to think more carefully about the rating process, may have impacted the results. Therefore, Study 4 was conducted at the same time as Study 3 but subjects were not asked to make remember/know judgments in this study.

Participants and Procedure

Two hundred and sixty-nine participants were recruited using MTurk. The criteria for participation were the same as described in Study 1, 2, and 3. Twenty-two participants were excluded to due to incomplete data, an identifiable response pattern, or a very rapid completion time (i.e., less than 1/3 of the mean completion time), resulting in total sample size of 247 (138 females and 109 males).

The average participant was 37 years old. One hundred and six participants had an undergraduate degree, 44 participants had a master's or terminal degree, 77 participants had either a 2 year degree or some college, and 20 participants had a high school diploma. Participants reported a wide range of occupations including administrative assistant, sales, customer service, education, and software engineering.

Measures

Servant leadership was assessed using the 28-item servant leadership scale (Liden et al., 2008). However, in this study, participants only indicated on a seven point scale, ranging from strongly disagree to strongly agree, to what extent they agreed with each statement about their manager and did not indicate whether each item was a remember or know judgment. Participants also completed the same *Trust* measure as described in Study 3. All scale reliabilities and correlations are reported in Table 8.

Study 4: Results

Replication of Study 3

The purpose of Study 4 was to replicate the episodic and semantic servant leadership results even though we did not ask subjects to make remember/know judgments in this study. To do so, we first created episodic and semantic servant leadership scales based on the items identified by participants from Study 3 as being either remember or know judgments. As would be expected with equivalent samples, Table 8 shows that the correlations among the variables were very similar to that of Study 3 as shown in Table 6. As in the previous studies, we used Hotelling's t to compare the correlations between the episodic and semantic leadership scales and trust. The correlation between episodic servant leadership and trust was significantly different from the correlation between semantic servant leadership and trust ($r = .78$ vs. $r = .68$, $t = 4.162$, $p < .001$). Next, each scale was entered into a regression analysis to examine the unique association of each scale with trust. As in Study 3, the episodic scale was more useful in estimating trust than the semantic scale ($\beta = .66$, $p = .000$ versus $\beta = .15$, $p = .04$, see Table 7). These results closely replicate those of Study 3. Thus, the results obtained in Study 3 appear to be driven by the memorial qualities of the items rather than the remember/know procedure used in Study 3 but not in Study 4.

As in the prior studies, we tested whether the difference between the episodic and semantic beta weights was statistically significant. The confidence intervals overlapped by less than 50%, therefore the beta weights were statistically different ($p < .05$, Cumming, 2009, see Appendix Figure A6). Specifically, half of the average of the overlapping confidence intervals was calculated (.078) and added to the episodic beta weight lower bound estimate (.492) which was .570. As the semantic upper bound estimate of .291 did not exceed the value of .570, the difference between the episodic and semantic beta weights was statistically significant.

Studies 1-4 were conducted using on-line data collection procedures, MTurk. Given that participants obtained through crowdsourcing sites are paid for each task completed, they may have been motivated to quickly answer questions, which could potentially affect the advantage of episodic compared to semantic memory. Namely, when people are trying to respond quickly, they may be more likely to rely on heuristics, which should increase the use of semantic processing and possibly reduce the advantage of the episodic scale. Furthermore, it was important to extend our results to an organizational sample that did not consist of individuals who routinely participate in research studies. Accordingly, we analyzed data from a fifth study in which participants were recruited from one company and completed paper and pencil questionnaires onsite during normal paid working hours with a researcher present.

Study 5: Methods

Participants and Procedure

Two hundred and twenty-one participants were recruited from two locations of a production and distribution company located in the Midwest of the United States. It should be noted that these data were originally collected for the purposes of a different study (Panaccio, Henderson, Liden, Wayne, & Cao, 2015). Servant leadership was the only variable that was used in both our study and Panaccio et al. (2015). All organizational members were invited to participate in the research project and participation was completely voluntary. Participants completed surveys during paid working hours. Twelve participants were excluded due to incomplete data or an identifiable response pattern resulting in a total sample size of 209. Seventy-three percent of the sample was male and 27% of the sample was female. The average participant was 36 years old and had a high school diploma.

Measures

Participants completed the same *Servant Leadership* and *Trust* measures described in Study 4. All scale reliabilities and correlations are reported in Table 9. Additional measures were also collected which allowed some exploratory analyses. They included empowerment, organizational commitment, and perceived organizational support. *Empowerment* was assessed using Spreitzer's (1995) 12- item measure. Participants indicated on a seven point scale ranging from strongly agree to strongly disagree the extent to which they agreed with each statement. *Organizational commitment* was assessed using Porter, Steers, Mowday, and Boulian's (1974) measure. Following Wayne, Shore, and Liden (1997), two items were deleted and the remaining seven items were summed to form a scale. Participants indicated on a seven point scales ranging from strongly disagree to strongly agree the extent to which they agreed with each item. *Perceived organizational support* was measured using the shortened 9-item measure that was adapted from the Survey of Perceived Organizational Support (SPOS, Eisenberger, Huntington, Hutchison, & Sowa, 1986). This measure has been used in prior research (e.g., Wayne et al., 1997; Eisenberger, Fasolo, & Davis-La-Mastro, 1990).

Study 5: Results

Organizational Replication of Study 3 and 4

The purpose of Study 5 was to replicate the episodic and semantic servant leadership scale results using an organizational sample even though subjects were not asked to make remember/know judgments. As described in Study 4, we first created episodic and semantic servant leadership scales based on the items identified by participants from Study 3 as being remember or know judgments. As described previously, we used Hotelling's t to compare the correlations between the episodic and semantic scales and trust. The correlation between

episodic servant leadership and trust was not significantly different from the correlation between semantic servant leadership and trust ($r = .66$ vs. $r = .64$, $t = .652$, $p = .515$).

Next, each subscale was used in a regression analysis to examine its unique association with trust. Again, as shown in Table 7, the episodic scale better estimated trust than the semantic scale ($\beta = .42$, $p = .000$ versus $\beta = .30$, $p = .001$).

Finally, we tested whether the difference between the episodic and semantic beta weights was statistically significant. The confidence intervals did not overlap by less than 50%, therefore the beta weights were not statistically different ($p < .05$, Cumming, 2009, see Appendix Figure A7). Specifically, half of the average of the overlapping confidence intervals was calculated (.094) and added to the episodic beta weight lower bound estimate (.223) which was .317. As the semantic upper bound estimate of .48 exceeded the value of .317 the difference between the episodic and semantic beta weights was not statistically significant.

Exploratory analyses

As previously noted, the data set used in Study 5 was originally collected for a different study that included empowerment, organizational commitment, and POS, which allowed us to consider whether episodic memory was more strongly associated with additional criteria. Heretofore, our focus has been on trust as a dependent variable, and trust is specific to one's particular relationship with another, allowing one to easily access specific events. Other criteria, however, may be more abstract, making it difficult to link specific behaviors. To account for such differences, in this exploratory section we offer and test a principle that focuses on predictor-criteria congruence in terms of underlying memory processes. Specifically, we expect that *the prediction of outcomes is best when the memorial basis of the scale is congruent with the nature of the criterion construct*. One could view this Predictive Congruence Principle as an

extension of Tulving and Thomson's (1973) encoding specificity principle to predictor/criterion relations. That is abstract, general constructs which apply across contexts should be more strongly associated with semantic leadership measures, whereas more context specific, and typically event-based, criteria should be more strongly associated with episodic leadership measures. Guided by the Predictive Congruence Principle, we examined the relationship of both episodic and semantic servant leadership scales with additional criteria.

Because we are introducing our Predictive Congruence Principle here, we took an exploratory approach to analyzing these additional variables. As per the previous analyses, we entered the Servant Leadership scales based on episodic and semantic memory in a series of regression analyses to examine how much each scale was uniquely associated with each outcome. The scale based on episodic memory provided better estimates of both empowerment and organizational commitment than did the scale based on semantic memory ($\beta = .32$, $p = .003$ vs, $\beta = .08$, $p = .482$; $\beta = .32$, $p = .002$ vs. $\beta = .20$, $p = .05$)[5]. However, the Servant Leadership scale based on semantic memory was a better estimator of POS than was the scale based on episodic memory ($\beta = .55$, $p = .000$ vs. $\beta = .06$, $p = .537$). Thus, consistent with the Predictive Congruence Principle, there appear to be circumstances where semantic measures are more strongly associated with outcomes. These findings are consistent with the general literature pertaining to each variable. Considering empowerment first, it is not a global construct but rather is specific to a particular work context; the scale is constructed such that the items focus on an *individual's experiences* rather than a general description of the work environment that might

[5] We also tested whether the difference between the episodic and semantic beta weights was statistically significant. The difference between the episodic and semantic beta weights was statistically significant for empowerment ($p = .011$) and POS ($p < .001$) but was not significant for organizational commitment ($p = .19$).

result in that experience (Spritzer, 1995). Therefore, the construct of empowerment is consistent with episodic memory and as indicated by our findings is more strongly associated with the episodic servant leadership scale.

Turning to organizational commitment, it has been described as a construct based on both commitment related behaviors and attitudinal commitment such as affect (Mowday, Steers, & Porter, 1979). Therefore, organizational commitment should be consistent with both concrete behaviors associated with episodic memory and generalized impressions associated with semantic memory. Supporting this idea, although the episodic scale was significantly associated with organizational commitment, and semantic scale was not, the difference in beta weights was neither large ($\beta = .32$ versus $\beta = .20$) nor statistically significant.

Finally, perceived organizational support (POS) is a generalized perception that the organization values employees' contributions and cares about their well-being (Kurtessis, Eisenberger, Ford, Buffardi, & Adis, 2017). Employees develop perceptions of organizational support by ascribing trait-like qualities to organizations (Eisenberger, Huntington, Hutchison, & Sowa, 1986). POS is a summary judgment, or generalized impression, that the organization values employees. As our findings indicated, this conceptualization is consistent with semantic memory which is characterized by general, abstract, global impressions; hence, the semantic servant leadership scale was a much better estimator of POS than was the episodic servant leadership scale.

Study 6: Methods

As the previous studies are correlational, they are not able to directly test the effect remember judgments have on the recollection of past leadership behaviors that have occurred. As such, the purpose of Study 6 was to experimentally test the impact of a memory source

intervention (Martell & Evans, 2005) on accuracy in the recollection of previous leadership behaviors. Specifically, participants observed leadership vignettes, before rating the occurrence of leadership behaviors. We did this to test whether using only remember judgments increased accuracy in the recognition of leadership behaviors and whether it reduced biases associated with lenient responses.

Participants and Procedure

Participants in this study were undergraduates recruited from a large university in the south-eastern United States. The original sample consisted of 146 individuals. Data from participants were removed if they failed an attention check item, provided the same response to all leadership behavior items, or if they indicated that they could not understand who was speaking in the leadership vignettes (see Appendix Figure A8 for data removal process and robustness checks). The final sample consisted of 110 undergraduates. The sample consisted of 73 percent women, with an average age of 19.24 (SD = 1.16).

The stimulus materials employed in this study were four video vignettes originally created by Hanges, Lord, Day, Sipe, Gradwohl-Smith, and Brown (1997). Each vignette contained two male and two female actors portraying a work team with one designated male leader. Each vignette was approximately four to five minutes long. Each vignette was constructed to display a specific number of leadership behaviors, as defined by previous research (Lord, Foti, & DeVader, 1984). After watching the videos, participants completed a scale measuring the need for cognitive closure (Roets & Van Hiel, 2011). Participants were randomly assigned to either one of two conditions, a memory source intervention (MSI) condition (*N = 56)* or control condition (*N* = 54). Next, participants in the MSI condition received the definition of remember and know judgments, based on the Martell and Evans' (2005) procedure. This

included telling the participants that remember and know judgments do not differ in confidence or certainty. To equate for the time taken by the remember vs. know instructions, participants in the control condition were asked to list several reasons why people attend college.

At this point in the study, participants were then told that they would be answering questions about the leader and the group in the videos. They were given a set of possible behaviors that the leader performed in the videos and asked whether the behavior occurred. Participants in the MSI condition were asked *only to answer yes if their judgment was based on a remember judgment*. Those in the control condition were only asked to rate whether the leadership behavior occurred or not.

Measures

Recognition of leadership behaviors. Participants completed a behavioral recognition questionnaire consisting of a set of 17 items. Of these items, nine occurred in the videos, and eight did not occur. Behaviors were chosen based on their relative frequency.

To measure the effects of the manipulation on memory, we followed past work (Martell & Evans, 2005; Martell & Willis, 1993) and calculated the four primary metrics of memory based on signal detection theory: (1) hit rate, (2) false alarm rate, (3) memory sensitivity, and (4) response bias. *Hit rate* represents the proportion of yes responses to behaviors that occurred. *False alarm rate* represents the proportion of yes responses to behaviors that *did not* occur. *Memory sensitivity* represents the hit rate – false alarm rate. While both the hit rate and false alarm rate represent forms of memory accuracy, memory sensitivity represents a composite measure. Finally, *response bias* represents whether participants had too liberal (i.e., bias towards choosing yes) or too conservative (i.e., bias towards saying choosing no) decision criteria when rating whether a behavior occurred. It is calculated by: *false alarm rate/(1-(hit rate – false alarm*

rate)). To understand the impact of the memory manipulation on the memory of leadership behaviors, we report the results from each of the four indexes below.

Study 6: Results

Hit Rates

A one-way between-subjects ANOVA, using R software, was conducted to compare the effect of the memory source intervention on hit rates in the recognition of leadership behaviors. There was a significant effect of the MSI on hit rates across the two conditions, F (1, 108) = 32.04, $p < .001$, *partial* $\eta^2 = .23$. Those in the control condition had significantly higher hit rates ($M = .52$, $SD = .18$) than those in the MSI condition ($M = .33$, $SD = .17$).

False alarm rates

A one-way between-subjects ANOVA was conducted to compare the effect of the memory source intervention on false alarm rates in the recognition of leadership behaviors. There was a significant effect of the MSI on false alarm rates across the two conditions, F (1, 108) = 22.65, $p < .001$, *partial* $\eta^2 = .17$. Those in the control condition had significantly higher false alarm rates ($M = .72$, $SD = .15$) than those in the MSI condition ($M = .59$, $SD = .16$).

Memory sensitivity

A one-way between-subjects ANOVA was conducted to compare the effect of the memory source intervention on the memory sensitivity of leadership behaviors. There was no significant effect of the MSI on memory sensitivity across the two conditions, F (1, 108) = 1.61, $p = .207$. There was no significant difference in overall memory accuracy in the recall of past leadership behaviors across conditions (MSI condition: $M = -.26$, $SD = .22$; control condition: $M = -.21$, $SD = .19$).

Response bias

A one-way between-subjects ANOVA was conducted to compare the effect of the memory source intervention on response bias of leadership behaviors. There was a significant effect of the MSI on response bias across the two conditions, F (1, 108) = 42.59, $p < .001$, *partial* $\eta^2 = .28$. Those in the control condition had significantly more liberal response bias than those in the MSI condition (MSI condition: $M = .47$, $SD = .10$; control condition: $M = .61$, $SD = .12$).

Response bias ranges from 0-1, with 0.5 representing no bias. A number above 0.5 represents an excessively liberal decision criteria (bias towards yes), and below 0.5 represents an overly conservative decision criteria (bias towards no). As displayed in Figure 1, those in the memory source condition had a small bias towards no responses (*95% CI* = 0.44 – 0.49). However, those in the control condition had a more considerable bias towards yes responses (*95% CI* = 0.57 – 0.64).

General Discussion

The results from a series of six studies demonstrate that we can form scales that are primarily based on episodic versus semantic memory and scales based on episodic memory provide better estimations of trust. In addition, the language used in episodic items tends to be more concrete than the language used in semantic items. Finally, it is possible to develop interventions that reduce the impact of liking and false alarms on leadership ratings by reducing liberal biases. These findings are important in several ways.

First, our work underscores the importance of the words used in scales and demonstrates that language is linked to the type of memory system used by raters. Namely, we showed that the language used in more episodic items was more concrete than the language used in items tending to emphasize semantic memory. Put differently, raters reported that they tended to rely on

generalized impressions when making judgments about abstract items, but tended to rely on vivid recollection of specific events when making judgments about concrete items.

Second, scales based on episodic memory were better estimators of trust in the leader than were scales based on semantic memory in four out of five studies. Because asking participants to consider the memory basis of each item may be intrusive, we replicated the results using traditional response procedures which enabled us to rule out the possibility that the results were an artifact due to the methodology. The superiority of items based on episodic memory suggests that measures based on concrete behaviors may be more useful than those based on generalized impressions, at least for personally relevant outcomes such as trust. However, this statement must be qualified by recognizing that in our applied sample of working adults, differences between scales were not significant when predicting trust, although they were in the predicted direction. It should be noted the significant effects were not due to greater variance in the episodic scales. Instead, the standard deviations for the semantic scales are higher than those of episodic scales in Studies 1, 3, 4, and 5. This suggests that there may be more valid variance in the episodic scales, which is consistent with the results of our IRT analyses (see Appendix Tables A 4-5). Furthermore, the results were replicated with three different measures of trust that included both cognitive and affective trust which suggests that the dependent variable did not favor episodic memory. Therefore, determining whether an item emphasizes remember or know judgment may be a useful way to screen items when developing scales. It is noteworthy that the absence of significant results in Study 5 could be due to several factors: the applied nature of the sample, the much higher percentage of males subjects than any of the other studies, completing measures at the same time and in a work setting, or a combination of these factors. Future research should examine such factors.

Third, asking raters to consider whether their ratings reflect remember or know judgments significantly reduces the impact of liking on ratings. One explanation, consistent with Srull and Wyer's (1989) general evaluative judgment heuristic, is that asking raters to consider the memory basis of each item encourages more effortful processing and promotes engagement in a more extensive memory search. Therefore, as shown in Studies 1 and 2, the use of this procedure reduced the impact of general evaluations (e.g., liking). This effect is distinct from reducing the role of affect, which may have an important role in encoding and retrieving information from episodic memory (Allen et al., 2008; Naidoo, Kohari, Lord, & DuBois, 2010). These results may also be explained by the type information that is typically available to individuals when completing ratings of leader behavior. For example, Hastie and Park (1986) contend that summary judgments are routinely used in decision making due to their high availability. In this case, asking raters to consider whether each item reflected a remember or a know judgment may have increased the availability of other leader behaviors. Consistent with this interpretation, Baltes and Parker (2000) found that having subjects recall relevant behaviors before making ratings reduced the impact of performance expectations (i.e., performance cue effect) on ratings, an effect we equate with the use of semantic memory.

Finally, the memory source intervention used in Study 6 that encouraged raters to rely on only episodic memory when completing ratings reduced false alarms in the recognition of leader behaviors, although it did not improve overall memory sensitivity. Furthermore, individuals completing ratings in the memory source condition tended to use more conservative decision criteria. In contrast, individuals in the control condition had higher hit rates and higher false alarms. This suggests that individuals in the control condition were inclined to endorse items *in general* as evidenced by a significantly more liberal response bias. In short, this study

demonstrated a causal relation between remember instructions and how participants reported memories for leader behavior. Although not increasing memory sensitivity, the memory source condition did reduce leniency bias, a finding that directly parallels the results of Martell and Evans' (2005) training procedure. Therefore, the memory source manipulation leads to more conservative responses. In addition, a reduction in bias is consistent with reduced impact of liking shown in Study 1 versus Study 2, which can also be interpreted as causal because there are differences between conditions to which participants were randomly assigned. However, the memory source manipulation did not increase the hit rate, therefore, the overall memory sensitivity (hit rate-false alarms) was not increased. The memory source manipulation focuses on recall rather encoding, therefore, the manipulation may not impact which leader behaviors are encoded into memory. A well replicated effect in the memory literature is that recall is best when encoding conditions match those of retrieval (Shrondrick, et al., 2010), a principle called *transfer-appropriate processes*. It is possible to design manipulations that focus on encoding, such as frame of reference training, however, as noted by Sulsky and Day (1992), frame of reference training increased bias. The transfer-appropriate processing principle suggests memory sensitivity would improve only when both encoding and retrieval conditions emphasized remember judgments.

Taken together, these results may help move the field away from the criticism that most measures of leadership primarily reflect gap-filling processes associated with implicit leadership theories and liking (Hunter et al., 2007; Martinko et al., 2018; Rush et al., 1977). This may be true when measurement processes tend to elicit semantic memory and reliance on implicit leadership theories (See Lord, Epitropaki, Foti, & Hansbrough, 2020). However, it may be possible to reduce bias and increase memory sensitivity if raters are trained to use remember

judgments in *both* encoding and retrieval processes. It is also possible to shift subjects away from using person schemas and toward greater reliance on scripts as a way to increase memory sensitivity (Foti & Lord, 1987).

Validity of the Remember/Know Procedure

The remember/know procedure was developed to distinguish between episodic and semantic memory (Tulving, 1985). However, there is a robust debate in the literature whether remember/know judgments represent different types of memory or differing degrees of confidence. Consistent with a dual process interpretation, neurophysiological studies support the contention that remember/know judgments reflect different types of memory (Eldridge, Sarfatti, & Knowlton, 2002). For example, only remember judgments are impacted by levels of processing (Gardiner, 1988) and hippocampal activity is elevated at retrieval for remember judgments but for not know judgments (Diana & Wang, 2018; Eldridge, Knowlton, Furmanski, Bookheimer, & Engel, 2000). Conversely, consistent with a signal detection interpretation, there is also empirical evidence that remember/know judgments represent differing degrees of memory strength (e.g., Donalson, 1996; Wixted, 2009). However, a signal detection interpretation does not allow for a description of the types of memories retrieved nor does it allow for the possibility that people can experience high confidence know judgments (Wixted & Mickes, 2010). Wixted (2009) calls for researchers to equate remember and know judgments for strength in order to disentangle memory strength (i.e., level of confidence) from recollection and familiarity. Notably, our instructions were carefully worded to indicate that remember and know judgments do not differ in terms of memory, their confidence or certainty. Therefore, the remember/know judgments reported here are more consistent with a dual processing interpretation.

Future Research and Implications

Our results provide the initial foundation to develop a theory of measurement that incorporates the rater into the measurement process by considering rater metacognitive processes. Such memorial insights by raters can provide important information regarding the memory processes used at the item level as well as how the stimulus provided by each item in a measure interacts with raters' memory systems. As detailed below, future research may wish to extend this method to other leadership or dependent variable scales, examine the accessibility and retrieval of different types of memories during the rating process, and use this method in scale development. Extensions to rater training are also feasible as suggested by Martell and Evans' (2005) experimental finding that metacognitive training can reduce bias in ratings.

Studies 1-5 used both transformational leadership (MLQ-5X) and servant leadership (SL-28) to demonstrate that scales based on episodic memory generally were better estimators of trust than scales based on semantic memory. Nevertheless, future research should examine other leadership measures to establish boundary conditions for the finding that more concrete items based on episodic memory are more strongly associated with outcomes. Such research should not ignore the dependent variable, as with more abstract and general criteria (rather than personally relevant items, such as trust), semantic items may be more strongly associated with outcomes, albeit, not because of accuracy in behavioral ratings. Indeed, the exploratory analyses presented in Study 5 show that scales based on semantic memory were better estimators of POS which is based on generalized impressions. Further extensions would be of particular interest as it relates to LMX because liking is a key component of LMX. The dimension of affect is defined as "the mutual affection members of a dyad have for each other based primarily on interpersonal attraction" (Dienesch & Liden, 1986, p. 625) and the items (Liden & Maslyn, 1998) center on generalized impressions (e.g., "I like my supervisor very much as person"). Moreover, the

affective outcomes typically associated with LMX, such as satisfaction and organizational commitment, also may be based on generalized impressions rather than specific, concrete events. Therefore, based on the Predictive Congruence Principle, it is possible that scales based semantic memory may better estimate outcomes associated with LMX. Likewise, scales based on semantic memory might be particularly relevant for research focused on generalized impressions or schematic processing, such as implicit leadership theories.

Our findings have implications for other leadership scales as well. In deciding which measures to use to replicate Studies 1 and 2, the first and second authors first coded several popular leadership scales using the LCM (Semin & Fieldler, 1991). Based on this coding, we chose not to use measures of abusive behavior (Tepper, 2000) or the Leader Behavior Descriptive Questionnaire (Stogdill, 1963), because items tended to be at the concrete end of these scales, which suggests that these measures might tend to emphasize episodic memory. However, it should be noted that, at least with the LBDQ, the finding that ratings are affected by performance information are well replicated (see Lord, 1985 for a review of studies showing this "performance cue effect"). Such results imply a strong semantic component to this scale.

Together these results imply that there might be several ways to move raters toward episodic memory. For example, the personal relevance of many transformational and servant leadership items in addition to concrete language may encourage people to rely on episodic memory in responding to items. Future research may also wish to examine whether negative, particularly abusive, leadership behaviors are associated with episodic memory. For example, the emotions literature suggests that people are more likely to rely on schematic processing when in a positive mood (Isen & Daubman, 1984; Isen, 1993), whereas negative emotion improves memory for an event (Diana & Wang, 2018). Moreover, prospect theory (Kahnman & Tversky,

1979) posits that negative outcomes are more salient. Taken together, this suggests that negative leadership behaviors are more likely to be encoded into episodic memory.

Our work could also be extended by using a reaction time paradigm to examine the possibility that semantic memory is more accessible and therefore is more likely to be used when completing leadership ratings (Diana, Reder, Arndt, & Park, 2006). Here, we anticipate faster reaction times for items that emphasize semantic memory but weaker relationships with criteria. Future research could also investigate other strategies to enhance the accessibility of episodic memory. For example, researches could design a study that manipulates response latencies coupled with a memory source intervention that encourages raters to rely on only on episodic memory (e.g., Martell & Evans, 2005). Doing so would enable us to pinpoint why raters prioritize semantic memory – is it a function of search time or the instructions or both. As already mentioned, priming script schemas rather than person schemas might increase accuracy (Foti & Lord, 1987).

Future research may wish to further examine the impact of the language used in scale items. For example, it is possible that scale items could be rewritten to increase the accuracy of ratings. In particular, ter Doest and Semin (2005) report that individuals remember concrete words better than abstract words. Taken together with our results, we might expect that scales based on more concrete items would be less likely to promote the use of gap filling processes than would scales based on more abstract items. This notion could be tested by using signal detection theory to compare the hit rates and false alarms of different versions of scales. Scales that emphasize episodic memory may also foster higher levels of agreement among raters. For example, Morgeson (2005) found substantial group level agreement in leader ratings when raters were focused on a specific problem or event. Thus, it may possible to increase both accuracy and

inter-rater agreement by creating items that focus on specific events rather than general impressions. It should be noted that this approach differs from behaviorally anchored rating scales (BARS; Smith & Kendall, 1963). Our approach to item development focuses squarely on the *rater and memory processes* raters are using during the rating process instead of a focus on items independent of the memory processes they elicit in raters.

Developing items that trigger *only* episodic memory is an interesting conundrum. Episodic memory is based on memory of a vivid event. As such, if the specific leader behavior did not occur or if the respondent did not have the opportunity to witness the behavior, episodic memory would not be available during recall. This is particularly relevant for low baseline leader behaviors, such as abusive supervision. Asking respondents to only complete items if they had had the opportunity to observe the leader behavior in question would be a useful first step to increase the likelihood of tapping into episodic memory.

Limitations

Items were classified in terms of their memory basis if 55% or more of the respondents considered them either remember or know judgments. Yet, a third or more of respondents classified the item in question differently. This suggests that a portion of the distinction between the items based on episodic and semantic memory is a function of the rater. It is difficult to pull apart the rater effects on remember and know judgments in Studies 1 and 3 because the rater was not held constant. However, we did test for some individual differences in the propensity to rely on episodic memory and also controlled for gender differences. None of these factors impacted our results (see Appendix Tables A 1-3). It is possible that other factors such as binding capacity or working memory capacity may impact behavioral encoding and the availability of episodic memory.

In order to provide an adequate comparison of episodic and semantic memory, it is necessary to select items that differ substantially in terms of R/K percentage, which we arbitrarily defined as at least a 10% difference, and there also needs to be enough items to make a reliable scale. Those two criteria supported picking 55% or more as the criteria for allocating items to semantic and episodic scales (see Appendix Table A7) which was consistent with the results. The difference in regression weights went down with more lenient criteria (53%) for cognitive trust as the dependent variable, but they went up slightly for affective trust, likely reflecting the strong affective basis of both episodic memory and affective trust. Indeed, encoding events into episodic memory strongly depends on their affective basis. Also, as scale reliability decreased due to fewer items if we used more stringent criteria such as 57% or greater, the regression weights for semantic and episodic measures were no longer significantly different (see Appendix Table A8). Because the results depended on having enough items with clear differences in episodic proportions, it was particularly important to replicate findings with another measure, which we did with the servant leadership measure.

Although the episodic and semantic scale were comprised of different items, they were highly correlated and therefore not independent. If the scales were uncorrelated, we would expect that the coefficients in the regressions would be similar to the raw correlations. Therefore, the observation that they differ is an indicator of collinearity. Yet, the portion of the variance that was unique to the episodic scales was consistently larger than that associated with semantic scales. Further, the difference in beta weights between the episodic and semantic scales associated with trust was statistically different across 4 studies and two different measures of trust. Although the difference between the beta weights in Study 5 was not statistically different, we found the same pattern of results with a sample that was markedly smaller than the other

studies, and was predominantly male. Nevertheless, measurement error can also affect the conclusions drawn from analyses, particularly when there is a high degree of collinearity.

It is noteworthy that memories are consolidated over time (McClelland, McNaughton, & O'Reilly, 1994; Winocur et al., 2010), and through this process the gists of episodic memories are combined to create a more schematic, semantic memory that is distributed across neocortical networks (the episodic memory depends on the hippocampal system). Winocur et al. (2010) argue that even after consolidation, both episodic and semantic memory remain available, which is consistent with our finding that for all items some subjects reported using semantic memory and other subjects reported that they used episodic memory. Which type of memory is used during the leadership rating task may depend on the retrieval circumstance, which in our studies varied across items and a memory source intervention. Given this dual memory trace logic, it is quite understandable that episodic scales and semantic scales were highly correlated for both transformational and servant leadership, as they reflect overlapping memory structures that had a common origin.

The relation of semantic memory to consolidation processes has two important implications. First, ratings of a supervisor may tend to become more semantic as tenure with that supervisor increases and many episodes with that supervisor are consolidated over time[6]. In addition, focusing leadership ratings on short time periods (i.e., the previous day or week) may emphasize episodic memory because the memory has not yet been consolidated, whereas a focus on a longer time period or an unspecified temporal focus for ratings may favor semantic memory. As such, it is possible that the impact of the memory source intervention on accuracy

[6] The authors wish to thank Ron Riggio for this suggestion.

and bias may become more pronounced over time as it may facilitate the retrieval of older memories.

Our results may also be impacted by endogeneity including omitted variables, common source bias, and measurement error. First in terms of omitted variables, while we did not find individual differences such as personality or gender impacted the results, there is always the possibility that a different omitted variable impacted the results. Given that people have different memories based on different experiences with different leaders, we would expect individual differences to matter. It is possible that memory might act as a mediator between individual differences and outcomes. Future studies could explore this possibility.

It should be noted that we only examined single source data collected at one time. Given that we were interested in personally relevant variables (i.e., trust), the use of subordinate reports was appropriate. As noted by Podsakoff et al. (2012), the only known fix would have been to include an instrumental variable. For example, future research may wish to include the time of day as an instrumental variable because of diurnal variation in memory. Moreover, by collecting data at the same point in time, we were able to create a more difficult test that did not advantage episodic memory. Specifically, the results of Arnulf, Larsen, Martinsen, and Bong (2014)'s study demonstrated the main source of quantitative variation in the surveys measuring leadership and organization outcomes was the degree of semantic overlap among the items. Furthermore, using single source data generalizes our findings to the "typical leadership study" (Hunter et al., 2007). However, future research should examine how scales based on episodic memory might predict future outcomes that come from multiple sources. Indeed, as detailed previously, we anticipate that scales based on episodic memory might increase inter-rater agreement.

The interrater agreement for the LCM coding, as represented by the Kappa statistic, was .58 for Study 1 and .47 in Study 3. Although the Kappa was significant in both studies, it represents moderate agreement (Landis & Koch, 1997). We did not use LCM (Semin & Fiedler, 1991) to code outcomes. LCM is used as it relates to person perception therefore it is less applicable for items that focus on an individual's internal experience. Nevertheless, based on the Predictive Congruence Principle, we might expect semantic leadership scales to be more strongly related to scales comprised of abstract items. While the level of abstraction as assessed by LCM tends to correlate with different types of memories, they are not synonymous. For example, the empowerment items focus on the psychological experience of empowerment and include word that are highly valenced (e.g., confident, important, and meaningful). Emotion is associated with episodic memory because it helps bind the memory to the source. Therefore it is possible that items that are highly valenced may also tap into episodic memory.

Four of our six samples were comprised of heterogeneous individuals drawn from an online sample. While there has been some criticism of the use of online samples (Harms & Desimone, 2015), several studies suggest that the behavior of MTurk participants is comparable to that of laboratory subjects. For example, Paolacci, Chandler and Ipeirotis (2010) were able to replicate the results of decision making experiments on MTurk samples. Moreover, Horton, Rand, and Zeckhauser (2011) in a replication of prior laboratory studies, found that both laboratory subjects and MTurk workers irrationally cooperative during a one-shot Prisoner's Dilemma game. As concluded by Mason and Suri (2011), evidence that MTurk is a valid means of collecting data is consistent and continues to accumulate. Nevertheless, concerns have been raised about the data quality and representativeness of MTurk samples. In particular, Feitosa, Joseph, and Newman (2015) found that data collected from non-English speakers contributed to

poor data quality. Moreover, Paolacci et al. (2009) caution that MTurk samples that may be disproportionately young and female as well as include "professional" Turkers. However, it should be noted that our samples were neither disproportionately young nor female. The lab setting for Study 6 also enabled us to control possible distractions and answer any questions about the remember vs. know instructions.

Although Study 6 provided the opportunity for stronger inferences due to its experimental design, given the nature of memory source manipulation, it is possible that participants in the memory source condition were aware that the study focused on memory processes. Participants in the memory source intervention (MSI) condition were asked to indicate that the behavior in question occurred only if it was based on a remember judgment, possibly creating a *demand effect* (Lonati, Quiroga, Zehneder, & Antonakis, 2018) which might promote more conservative ratings. The instructions used in the experiment were identical to the memory source manipulation used by Martell and Evans (2005). However, we took steps to minimize the demand of characteristics while maintaining the integrity of the memory source manipulation. The experiment was conducted by individuals who were unaware of the experimental hypotheses or the purpose of the study. To reduce the social desirability of remember judgments, the instructions stated that both types of judgments are useful and that remember and know judgments do not differ in terms of their confidence or certainty. As such, there should not have been any indication of the hypotheses or any preferred response from participants. Finally, while individuals in the MSI condition had a minor conservative decision bias (i.e., tendency to rate "No" on recall of behaviors), they did not have a strong conservative decision bias (i.e., response bias close to .5). Nevertheless, future research should give additional consideration to the role of demand characteristics in designing memory source interventions.

Finally, it is important to note that episodic memory may be unusual. First, in terms of person perception, individuals give priority to categorization-based knowledge structures and only move on from schematic processing if they have the motivation and cognitive resources available to so do (Brewer, 1988; Fiske & Neuberg, 1990). Upon retrieval, general evaluations are quicker to access and individuals may end their memory search once a response has been found. Furthermore, leadership measures center on questions about the *leader.* If raters did not encode the information in terms of leadership, responding to questions that refer to this categorization may pose challenges. Finally, the content of some of the items used in leadership scales does not ask for episodic memory and instead focuses on generalized impressions or whether people like their boss (e.g., Yammarino et al., 2020). Nevertheless, we believe that it is possible to move people toward episodic memory by a variety of strategies detailed here. Under some circumstances, this might increase validity.

Conclusions

Rater memory systems are a crucial component of leadership ratings due to the retrospective nature of the rating task. The type of memory expressed upon retrieval depends, in part, on the cues that are present in the items and setting (Wincocur et al., 2010). Our results suggest that scales based on episodic memory and a memory source intervention may act as memory probes that move raters toward episodic processing. Taken together, this represents a fundamental step toward a better understanding of the role of rater memory processes in leadership ratings and how this information can be used to improve leadership measurement. We expect that there are many ways future research can build on these findings to further the understanding of memory processes on rating leader behavior or other types of behavior (e.g., performance appraisals).

References

Addis, D. R., & Schacter, D. L. (2012). The hippocampus and imagining the future: Where do we stand? *Frontiers in Human Neuroscience, 5*, 2-15.

Allen, P., Kaut, K., & Lord, R. G. (2008). Emotion and episodic memory. In E. Dere, A. Easton, L. Nadel, & J. P. Huston (Eds.), *Handbook of behavioral neuroscience: Episodic Memory Research, Vol.18* (pp. 115-132). Elsevier Science.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Bong, C. H. (2014). Predicting survey responses: How and why semantics shape survey statistics on organizational behavior. *PloS One, 9*.

Arvey, R. D., & Murphy, K. R. (1998). Performance evaluation in work settings. *Annual Review of Psychology, 49*, 141-168.

Baltes, B. B., & Parker, C. P. (2000). Reducing the effects of performance expectations on behavioral ratings. *Organizational Behavior and Human Decision Processes, 82*, 237-267.

Bass, B. M., & Avolio, B. J. (1996). *Manual for the Multifactor Leadership Questionnaire*. Palo Alto, CA: Mindgarden .

Bono, J. E., Hooper, A. C., & Yoon, D. J. (2012). Impact of rater personality on transformational and transactional leadership ratings. *Leadership Quarterly, 23*, 132-145.

Brewer, M. B. (1988). A dual-process model of impression formation. In R.S. Wyer & T. K Srull (Eds.), *Advances in Social Cognition, Vol. 1*, (pp-1-36). Hillsdale, NJ: Lawrence Erlbaum Associates.

Brown, D. J., & Keeping. L. M. (2005). Elaborating the construct of transformational leadership: The role of affect. *Leadership Quarterly, 16*, 245-273.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed,). New Jersey: Lawrence Erlbaum.

Cumming, G. (2009). Inference by eye: Reading the overlap of independent confidence intervals. *Statistics in Medicine, 28,* 205-220.

Diana, R.A., Reder, L.M., & Park, H. (2006). Models of recognition: A review of arguments in favor of a dual-process account. *Psychometric Bulletin and Review, 13,* 1-21.

Diana, R. A., & Wang, F. (2018). Episodic Memory. In J. Wixted (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, Learning and Memory, 1*, 67 – 100 (pp.1-33). New York: Wiley.

Dienesch, R.M., & Liden, R.C. (1986). Leader-member exchange model of leadership: A critique and further development. *Academy of Management Review, 11*, 618-634.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24,* 523-533.

Eden, D., & Leviatan, U. (1975). Implicit leadership theory as a determinant of the factor structure underlying supervisory behavior scales. *Journal of Applied Psychology, 60,* 736-741.

Eisenberger, R., Fasolo, P., & Davis-LaMastro, V. (1990). Perceived organizational support and employee diligence, commitment, and innovation. *Journal of Applied Psychology, 75,* 51-59.

Eisenberger, R., Huntington, R., Hutchison, S., & Sowa, D. (1986). Perceived organizational support. *Journal of Applied Psychology, 71,* 500-507.

Eldridge, L .L, Knowlton, B. J., Furmanski, C., Bookheimer, S., & Engel, S. A. (2000). Remembering episodes: A selective role for the hippocampus during retrieval. *Natural Neuroscience, 3,* 1149-1152.

Eldridge, L. L., Sarfatti, S., & Knowlton, B. J. (2002). The effect of testing procedure on remember-know judgments. *Psychonomic Bulletin & Review, 9,* 139-145.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Mahwah, New Jersey: Lawrence Erlbaum.

Engle, E. M., & Lord, R. G. (1997). Implicit leadership theories, self-schemas and leader-member exchange. *Academy of Management Journal, 40,* 988-1010.

Ercikan, K., & Pellegrino, J. (2017). *Validation of Score Meaning for the Next Generation of Assessments*. New York: Routledge.

Feitosa, J., Joseph, D. L., & Newman, D. A. (2015). Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English. *Personality and Individual Differences, 75,* 47-52.

Feldman, J. M., & Lynch, J. G. (1998). Self-generated validity and other effects of measurement on belief, attitude, and intention. *Journal of Applied Psychology, 73,* 421-435.

Felfe, J., & Schyns, B. (2010). Followers' personality and the perception of transformational leadership: Further evidence for the similarity hypothesis. *British Journal of Management, 21,* 393-410.

Fiske, S. T., Lin, M., & Neuberg, S. L. (1999). The continuum model: Ten years later. In S. Chaiken & Y. Trope (Eds.), *Dual-Process Theories in Social Psychology* (pp. 231-254). New York: The Gilford Press.

Fiske, S. T., & Taylor, S. E. (2013, 2nd ed.). *Social Cognition: From Brains to Culture*. New York: McGraw-Hill.

Foti, R. J., & Hauenstein, N. M. (2007). Pattern and variable approaches in leadership emergence and effectiveness. *Journal of Applied Psychology, 92,* 347-355.

Foti, R. J., & Lord, R. G. (1987). Prototypes and scripts: The effects of alternative methods of processing information. *Organizational Behavior and Human Decision Processes, 39,* 318-341.

Freitas, A. L., Gollwitzer, P., & Trope, Y. (2004). The influence of abstract and concrete mindsets on anticipating and guiding others' self-regulatory efforts. *Journal of Experimental Social Psychology, 40,* 739-752.

Fujita, K., Trope, Y., Liberman, N., & Levin-Sagi, M. (2006). Construal levels and self-control. *Journal of Personality and Social Psychology, 90,* 351-367.

Gardiner, J. M. (1988). Functional aspects of recollective experiences. *Memory & Cognition, 16,* 309-313.

Gilbert, D. T., Pellham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology, 54,* 733-739.

Guillem, M., & Mograss, M. (2005). Gender differences in memory processing: Evidence from event-related potentials. *Brian and Cognition, 57,* 84-92.

Hall, R, J., & Lord, R.G. (1995). Multi-level information processing explanations of followers' leadership perceptions. *Leadership Quarterly 6,* 265-287.

Hanges, P. J., Lord, R. G., Day, D. V., Sipe, W. P., Gradwohl - Smith, W. G., & Brown, D. J. (1997, April). Leadership and gender bias: Dynamic measures and nonlinear modeling. *Paper presented at 12th Annual Conference of the Society for Industrial and Organizational Psychology*, St. Louis, MO.

Hanges, P. J., Lord, R. G., & Dickson, M. W. (2000). An information processing perspective on leadership and culture: A case for connectionist architecture. *Applied Psychology: An International Review, 49,* 133-161.

Hansbrough, T. K., Lord, R. G., & Schyns, B. (2015). Reconsidering the accuracy of leadership ratings. *Leadership Quarterly, 26,* 220-237.

Harms, P. D., & Desimore, J. (2015). Caution! MTURK workers ahead-Fines doubled. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8,* 183-190.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review, 93,* 258-268.

Heilman, M.E., & Chen, J. J. (2005). Same behavior, different consequences: Reactions to men's and women's altruistic citizenship. *Journal of Applied Psychology, 90,* 905-916.

Herlitz, A., Nilsson, L. G., & Backman, L. (1997). Gender differences in episodic memory. *Memory and Cognition, 25,* 801-811.

Horton, J. J., Rand, D. G., & Zeckahauer, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics, 14,* 399-425.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5,* 64-86.

Hunter, S. T., Bedell-Avers, K. E., & Mumford, M. D. (2007). The typical leadership study: Assumptions, implications, and potential remedies. *Leadership Quarterly, 18,* 435-446.

Isen, A. M. (1993). Positive affect and decision making. In M. Lewis & J. Haviland (Eds.), *Handbook of Emotions* (pp.261-277). New York: Gilford Press.

Isen, A. M., & Daubman, K. A. (1984). The influence of affect on categorization. *Journal of Personality and Social Psychology, 47,* 1206-1217.

Jiga-Boy, G. M., Clark, A. E., & Semin, G. R. (2013). Situating construal level: The function of abstractness and concreteness in social contexts. *Social Cognition, 31,* 201-221.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263-292.

Kurtessis, J. N., Eisenberger, R., Ford, M. T., Buffardi, L. C., Stewart, K. A., & Adis, C. A. (2017). Perceived organizational support: A meta-analytic evaluation of organizational support theory. *Journal of Management, 43,* 1854-1884.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Larson, J. R., Jr., Lingle, J. H., & Scerbo, M. M. (1984). The impact of performance cues on leader-behavior ratings: The role of selective information availability and probabilistic response bias. *Organizational Behavior and Human Performance, 33,* 323-349.

Lewis, J. D., & Weigart, A. (1985). Trust as a social reality. *Social Forces, 63,* 967-985.

Liden, R.C., & Maslyn, J.M. (1998). Multidimensionality of leader-member exchange: An empirical assessment through scale development. *Journal of Management, 24,* 43-72.

Liden, R. C., Wayne, S. J., Zhao, H., & Henderson, D. (2008). Servant leadership: Development of a multidimensional measure and multi-level assessment. *Leadership Quarterly, 19,* 161-177.

Lindskold, S. (1978). Trust development, the GRIT proposal and the effects of conciliatory acts on conflict and cooperation. *Psychological Bulletin, 85,* 772-793.

Lonati, S, Quiroga, B. F., Zehnder, C., & Antonakis, J. (2018). On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management, 64,* 19-40.

Lord, R. G. (1985). Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology, 70,* 66-71.

Lord, R. G., Brown, D. J., Harvey, J. L., & Hall., R. J. (2001). Contextual constraints on prototype genration and their multilevel consequences for leadeship percpetions. *Leadership Quarterly, 12,* 133-152.

Lord, R.G., & Dinh, J.E. (2012). Aggregation processes and levels of analysis as organizing structures for leadership theory. In D. V. Day & J. Antonaki (Eds.), *The Nature of Leadership* (pp.29-65). Los Angeles, CA: Sage.

Lord, R. G., Epitropaki, O., Foti, R. J., & Hansbrough, T. K. (2020). Implicit leadership and followership theories and dynamic processing of leadership information. *Annual Review of Organizational Behavior and Organziational Psychology, 7,* 49-74.

Lord, R. G., Foti, R. J., & De Vader, C. L. (1984). A test of leadership categorization theory: Internal structure, information processing, and leadership perceptions. *Organizational Behavior and Human Performance, 34,* 343-378.

Lord, R. G., & Maher, K. J. (1991). *Leadership and Information Processing: Linking Perceptions and Performance*. Boston: Routledge.

Marschark, M., & Cornoldi, C. Imagery and verbal memory in C. Cornoldi & A. McDaniel (Eds), *Imagery and Cognition* (pp. 133-182). New York: Springer.

Martell, R. F., & Evans, D. P. (2005). Source-monitoring training: Toward reducing rater expectancy effects in behavioral management. *Journal of Applied Psychology, 90,* 956-963.

Martell, R. F., & Willis, C. E. (1993). Effects of Observers′ Performance Expectations on Behavior Ratings of Work Groups: Memory or Response Bias?. *Organizational Behavior and Human Decision Processes, 56,* 91-109.

Martinko, M. J., Mackey, J. D., Moss, S. E., Harvey, P., McAllister, C. P., & Brees, J. R. (2018). An exploration of the role of subordinate affect in leader evaluations. *Journal of Applied Psychology, 103,* 738-752.

Mason, W., & Suri, S. (2011). Conducing behavioral research on Amazon's Mechanical Turk. *Behavioral Research Methods, 44,* 1-23.

McAllister, D. J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal, 38,* 24-59.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1994). Why there are complimentary learning systems in the hippocampus and neocortex: Insights from the successes and

failures of connectionist models of learning and memory. *Psychological Review, 102*, 419-457.

Morgeson, F. P. (2005). The external leadership of self-managing teams: Intervening in the context of novel and disrupting events. *Journal of Applied Psychology, 90,* 497-508.

Mowday, R. T., Steers, R. M., & Porter, L. W. (1979). The measurement of organizational commitment. *Journal of Vocational Behavior, 14,* 224-247.

Nadel, L., & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology, 7,* 217-227.

Naidoo, L. J., Kohari, N. E., Lord, R. G., & DuBois, D. A. (2010). "Seeing" is retrieving: Recovering emotional content in leadership ratings through visualization. *Leadership Quarterly, 21, 886-900.*

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*, 411-419.

Panaccio, A., Henderson, D. J., Liden, R. C., Wayne. S. J., & Cao, X. (2015). Toward an understanding of when and why servant leadership accounts for employee extra-role behaviors. *Journal of Business and Psychology, 30,* 657-675.

Pavino, A. (1986). *Mental representation: A dual-coding approach*. New York: Oxford University Press.

Pavino, A., (1991). *Images in mind: The evolution of a theory*. New York: Harvester Wheatsheaf.

Pavino, A. (1995). Imagery and memory. In M.S. Gazzaniga (ed.), *The Cognitive Neurosciences* (pp. 977-986). Cambridge, MA: MIT Press.

Porter, L. W., Steers, R. M., Mowday, R. T., & Boulian, P. V. (1974). Organizational comment, job satisfaction, and turnover among psychiatric technicians. *Journal of Applied Psychology, 59,* 603-609.

Roberson, L., Galvin, B .M., & Charles, A. C. (2007). When group identities matter: Bias in performance appraisal. *Academy of Management Annals, 1,* 617-650.

Robinson, S. L. (1996). Trust and breach of the psychological contract. *Administrative Science Quarterly, 41,* 574-599.

Roets, A., & Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the need for closure scale. *Personality and Individual Differences, 50, 90-94.*

Rosette, A.S., Leonardelli, G.J., & Phillips, K.W. (2008). The white standard: Racial bias in leader categorization. *Journal of Applied Psychology, 93,* 758-777.

Rush, M. C., Thomas, J. C., & Lord, R. G. (1977). Implicit leadership theory: A potential threat to the internal validity of leader behavior questionnaires. *Organizational Behavior and Human Performance, 20,* 93-110.

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V.C., Spreng, R. N., & Szpunar, K. K. (2012). The future of memory: Remembering, imagining and the brain. *Neuron Review, 76,* 677-694.

Scherbaum, C. A., Finlinson, S., Barden, K., & Tamanini, K. (2006). Applications of item response theory to measurement issues in leadership research *Leadership Quarterly, 17, 366-386.*

Scott, K. A., & Brown, D. J. (2006). Female first, leader second? Gender bias in the encoding of leadership behavior. *Organizational Behavior and Human Decision Processes, 101,* 230-242.

Semin, G. R., & Fiedler, K. (1991). The linguistic category model, its bases, applications, and range. *European Review of Social Psychology, 2,* 1-30.

Shondrick, S. J., Dinh. J. E., & Lord, R. G. (2010). Developments in implicit leadership theory and cognitive science: Applications to improving measurement and understanding alternatives to hierarchical leadership. *Leadership Quarterly*, *21,* 959-978.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*, 108-131.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*, 149-155.

Spreitzer, G. M. (1995). Psychological empowerment in the workplace: Dimensions, measurement, and validation. *Academy of Management Journal, 38,* 1442-1465.

Srull, T. K., & Wyer, R. S. (1989). Person memory and person judgment. *Psychological Review, 96*, 58-83.

Staw, B. M. (1975). Attribution of the "causes" of performance. *Organizational Behavior and Human Performance, 13,* 414-432.

Stogdill, R. M. (1963). *Manual for the Leader Behavior Description Questionnaire: Form XII*. Columbus, OH: Ohio State University Bureau of Business Research, College of Commerce and Administration.

Sulsky, L. M., & Day, D. V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology, 77,* 501-510.

Tepper, B. J. (2000). Consequences of abusive supervision. *Academy of Management Journal, 43*, 178-190.

ter Doest, L., & Semin, G. R. (2005). Retrieval contexts and the concreteness effect: Dissociations in memory for concrete and abstract words. *European Journal of Cognitive Psychology, 17,* 859-881.

Tulving, E. (1985). How many memory systems are there? *American Psychologist, 40,* 385-398.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual review of Psychology, 53,* 1-25.

Tulving, E., & Thomson, D. (1973). Encoding specificity and retrieval in episodic memory processes. *Psychological Review, 80,* 352-372.

van Knippenberg, D., & Sitkin, S. B. (2013). A critical assessment of charismatic-transformational leadership research: Back to the drawing board? *Academy of Management Annals, 7,* 1-60.

Wayne, S. J., Shore, L. M., & Liden, R. C. (1997). Perceived organizational support and leader-member exchange: A social exchange perspective. *Academy of Management Journal 40,* 82-111.

Weiss, H. M., & Adler, S. (1981). Cognitive complexity and the structure of implicit leadership theories. *Journal of Applied Psychology, 66,* 69-78.

Winocur, G., Moscovitch, M., & Bontempi, B. (2010). Memory formation and long-term retention in humans and animals: Convergence towards a transformational account of hippocampal-neocortical interactions. *Neuropsychologia, 48,* 2339-235.

Wixed, J. T. (2009). Remember/know judgments in cognitive neuroscience: An illustration of the underrepresented point of view. *Learning & Memory, 16,* 406-412.

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117,* 1025-1054.

Wright, D. B., & Loftus, E. F. (2008). Eyewitness memory: In G. Cohen & M. A. Conway (eds.), *Memory in the Real World* (3rd ed., pp. 91-106). Howe, UK: Psychological Press.

Yammarino, F.J., Cheong, M., Kim, J., & Tsai, C.-Y. (2020). Is Leadership More Than "I Like My Boss"? In: M. R. Buckley, A .R. Wheeler, J. E. Baur & J. R. B. Halbesleben (eds.) *Research in Personnel and Human Resources Management*, Vol. 38 (pp. 1-55), Bingley, UK: Emerald.

Table 1: MLQ 5X Items Considered either Remember or Know Judgments (Study 1)

(Transformational items in bold)

Remember items (N=8) (% remember responses in parentheses)
1 (57.6), 3 (54.5), 7 (60.0), **8 (55.5), 9 (56.2),** 11 (61.7), **13 (64.8), 15 (59.3),** 16 (55.9), **19 (55.2),** 27(57.9), **30 (57.6), 31 (55.2), 32 (61.0)**, 33 (59.0), 35 (66.2)
Know items (N=6) (% know responses in parentheses)
2 (58.6), 10 (57.9), 14 (54.8), 17 (58.6), **18 (58.3),** 20 (55.5), **23 (63.1),** 24 (56.2), **34 (59.7)**
Unclassified items*
4 (54.1), 5 (52.8), **6 (53.1)**, 12 (54.1) , **21 (48.3)** , 22 (53.1), **25 (47.2)**, **26 (51.0)**, 28 (46.7), **29 (53.1), 36 (51.4)**

*Episodic percentage, semantic percentage =1-epsisodic percentage

Table 2: Correlations among Transformational Leadership, Episodic and Semantic Transformational Leadership Scales, Liking, and Trust (Study 1)

	M	**SD**	**1**	**2**	**3**	**4**	**5**	**6**
1. Unclassified TL Items	3.22	.81	(.78)					
2. Episodic scale	3.22	.88	.85**	(.89)				
3.Semantic scale	3.10	.92	.83**	.89**	(.87)			
4. Affect-based trust	3.42	1.10	.71**	.74**	.69**	(.93)		
5. Cognition-based trust	3.79	.94	.63**	.70**	.66**	.71**	.(89)	
6. Liking	3.78	1.00	.63**	.69**	.63**	.82**	.77**	(.94)

Note: *p <.05; ** p<.01, Reliabilities in brackets in the diagonal

N= 290

Table 3: Estimated Coefficients and Effect Sizes for Cognition-Based and Affect-Based Trust with Episodic and Semantic Scales of Transformational Leadership

	STUDY 1			STUDY 2		
	B	95% CI	β	B	95% CI	β
Cognition-based trust						
Episodic Scale	.55**	.36 to .74	.52**	.52*	.34 to .70	.56*
Semantic Scale	.21*	.03 to .39	.20*	.23*	.04 to .43	.22*
Adjusted R^2			.49			.58
Cohen's f^2 Episodic Scale[7]			.12			.15
Cohen's f^2 Semantic Scale			.02			.02
Affect-based Trust						
Episodic Scale	.77**	.56 to .98	.62**	.62**	.43 to .81	.62**
Semantic Scale	.17	-.03 to .37	.14	.17	-.04 to .38	.14
Adjusted R^2			.55			.58
Cohen's f^2 Episodic Scale			.18			.17
Cohen's f^2 Semantic Scale			.02			.02

Note: ** p<.01, * p<.05, B= unstandardized coefficients, β= standardized coefficients. Study 1 *N* = 290, Study 2 *N* = 289.

[7] Cohen's f^2 (Cohen, 1988) is commonly used for calculating global effect size. However, a variation of Cohen's f^2 for local effect size is more relevant here in order to account for the proportion of variance uniquely accounted for by each scale $f^2 = \frac{R^2_{AB} - R^2_{A}}{1 - R^2_{AB}}$

Table 4: Correlations among Transformational Leadership, Episodic and Semantic Scales of Transformational Leadership, Liking, and Trust (Study 2)

	M	**SD**	**1**	**2**	**3**	**4**	**5**	**6**
1.Unclassified TL Items	3.29	.91	(.84)					
2. Episodic Scale	3.73	1.08	.90**	(.89)				
3. Semantic Scale	3.21	.98	.88**	.92**	(.88)			
4. Affect-based trust	3.31	1.08	.74**	.76**	.72**	(.92)		
5. Cognition-based trust	3.71	1.00	.73**	.76**	.73**	.73**	(.90)	
6. Liking	3.73	1.07	.76**	.78**	.75**	.82**	.82**	(.95)

Note: *p <.05; ** p<.01, Reliabilities in brackets in the diagonal

N= 289

Table 5: Servant Leadership Items Considered either Remember or Know Judgments (Study 3)

Remember items (N=7) (% remember responses in parentheses)
8 (56.4), 9 (60.3), 16 (59.1), 17 (58.4), 20 (65.4), 22 (58.8), 25 (56.4),
Know items (N=11) (% know responses in parentheses)
3 (58.8), 4 (71.2), 6 (64.6), 10 (64.6), 11 (58.8), 14 (56.0), 18 (62.4), 19 (59.5), 26 (63.8), 27 (55.6), 28 (61.9)
Unclassified items (N=10)
1, 2, 5, 7, 12, 13, 15, 21, 23, 24

Table 6: Correlations among Servant Leadership, Episodic and Semantic Servant Leadership Scales, and Trust (Study 3)

	M	**SD**	**1**	**2**	**3**	**4**
1. Unclassified SL items	5.01	1.07	(.87)			
2. Episodic scale	5.25	1.15	.87**	(.89)		
3. Semantic scale	4.22	1.26	.86**	.77**	(.91)	
4. Trust	5.39	1.44	.80**	.80**	.71**	(.94)

Note: *p <.05; ** p<.01, Reliabilities in brackets in the diagonal

N= 257

Table 7: Estimated Coefficients and Effect Sizes for Trust with Episodic and Semantic Scales of Servant Leadership

	STUDY 3			STUDY 4			STUDY 5		
Trust	B	95% CI	β	B	95% CI	β	B	95% CI	β
Episodic Scale	.78**	.63 to 92	.62**	.79**	.63 to .96	.66**	.51**	.30 to .72	.42**
Semantic Scale	.27**	.14 to .40	.24**	.14*	.01 to .28	.15*	.36**	.16 to .56	.30**
Adjusted R^2			.66			.60			.47
Cohen'$_{f2}$ Episodic Scale			.47			.38			.11
Cohen's f^2 Semantic Scale			.01			0			.06

Note: ** p<.01, * p<.05, B= unstandardized coefficients, β= standardized coefficients. Study 3 *N* =257, Study 4 *N* = 247, Study 5 *N*=209

Table 8: Correlations among Servant Leadership, Episodic and Semantic Servant Leadership Scales, and Trust (Study 4)

	M	**SD**	**1**	**2**	**3**	**4**
1. Unclassified SL Items	5.0	1.09	(.89)			
2. Episodic scale	5.24	1.10	.91**	(.87)		
3.Semantic scale	4.26	1.34	.86**	.82*	(.94)	
4. Trust	5.47	1.33	.78**	.78**	.68**	(.93)

Note: *p <.05; ** p<.01, Reliabilities in brackets in the diagonal

N= 247

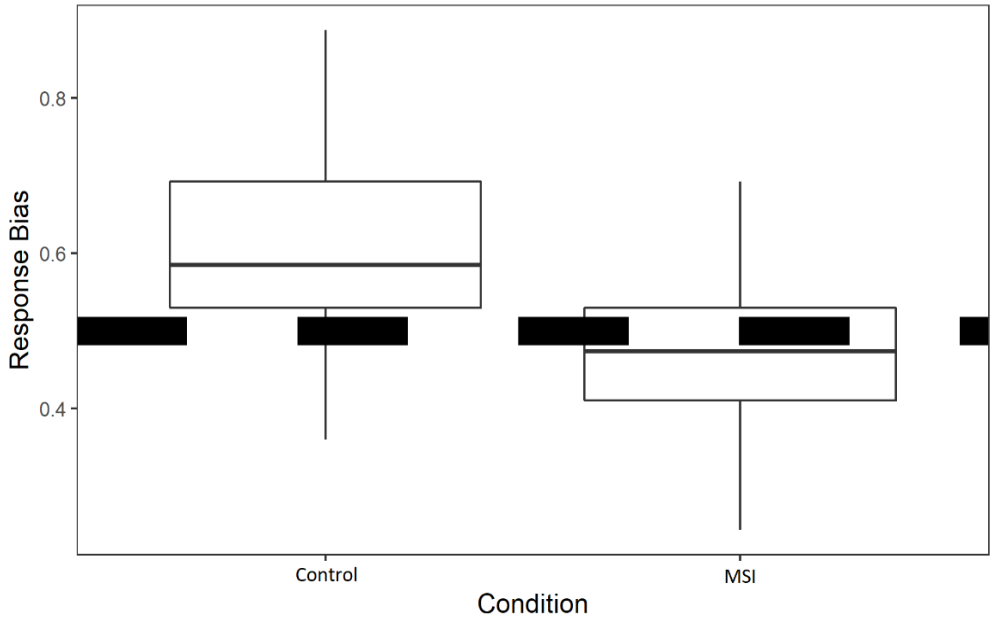
Table 9: Correlations among Servant Leadership, Episodic and Semantic Servant Leadership Scales, and Trust (Study 5)

	M	**SD**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1. Unclassified SL Items	4.78	.94	(.85)						
2. Episodic scale	5.01	1.02	.86**	(.86)					
3. Semantic scale	4.14	1.04	.85**	.81**	(.90)				
4. Trust	4.99	1.24	.71**	.66**	.64**	(.89)			
5. Empowerment	5.78	.74	.38**	.38**	.34**	.18**	(.85)		
6. Commitment	5.85	.97	.47**	.48**	.46**	.26**	.53**	(.88)	
7. POS	4.67	1.10	.53**	.50**	.60**	.41**	.44**	.59**	(.90)

Note: *p <.05; ** p<.01, Reliabilities in brackets in the diagonal

N= 209

Figure 1 Response Bias Across Condition (Study 6)



Note: This figure represents the distribution of response bias across subject in the control and MSI condition. Response bias ranges from 0-1. The dashed line represents 0.5, indicating no response bias. Below the line indicates a conservative rating bias (tendency to rate no, a behavior did not occur), and above the line indicates a liberal rating bias (tendency to rate yes, a behavior did occur). *MSI* = Memory Source Intervention.

Appendix

Table A1: Estimated Coefficients for R Sum with Individual Differences (Study 1)

	B	95% CI	β
Openness	-.16	-1.51 to 1.12	-.01
Conscientiousness	-.23	-2.0 to 1.52	-.01
Agreeableness	1.11	-.33 to 2.56	.07
Extraversion	.01	-1.65 to 1.66	.00
Neuroticism	.35	-1.03 to 1.72	.02
Honesty	-.44	-1.86 to .97	-.03
Positive affect	1.25	-.21 to 2.71	.09
Negative affect	-.23	-1.74 to 1.27	-.02
Adjusted R^2			.001

Note: R sum is the individual propensity to rely on remember judgments across all responses. This variable was created by computing a proportion which was the number of a subject's remember responses summed over the all of the items divided by their total number of scale items (i.e., 36). B= unstandardized coefficients, β= standardized coefficients. Study 1 N = 290.

Table A2: Estimated Coefficients for Cognition-Based and Affect-Based Trust with Episodic and Semantic Scales Transformational Leadership Controlling for Gender

	STUDY 1			STUDY 2		
	B	95% CI	*β*	B	95% CI	β
Cognition-based trust						
Gender	-.04	-.20 to .11	-.02	-.17*	-.32 to -.02	-.08*
Episodic Scale	.55**	.36 to .74	.52**	.53**	.35 to .70	.56**
Semantic Scale	.21*	.03 to .39	.20*	.23*	.03 to .42	.22*
Adjusted R^2			.49			.59
Affect-based Trust						
Gender	-.10	-.27 to .08	-.04	-.13	-.30 to .03	-.06
Episodic Scale	.77**	.56 to .98	.62**	.63**	.44 to .82	.63**
Semantic Scale	.17	-.03 to .37	.14	.17	-.04 to .38	.15
Adjusted R^2			.55			.59

Note: ** p<.01, * p<.05. B= unstandardized coefficients, β= standardized coefficients. Study 1 *N* = 290, Study 2 *N* = 289.

Table A3: Estimated Coefficients for Trust with Episodic and Semantic Scales of Servant Leadership Controlling for Gender

	STUDY 3			STUDY 4		
Trust	B	95% CI	β	B	95% CI	β
Gender	.00	-.11 to .11	.00	.01	-.09 to .12	.01
Episodic Scale	.78**	.63 to 92	.62**	.79**	.63 to .96	.66**
Semantic Scale	.27**	.14 to .40	.24**	.15*	.01 to .28	.15*
Adjusted R^2			.66			.60

Note: ** p<.01, * p<.05. B= unstandardized coefficients, β= standardized coefficients Study 3 *N* =257, Study 4 *N* = 247

Table A4: Parameter estimates from Samejima's graded response IRT model (Study 1)

	a	b_1	b_2	b_3	b_4	Max IIF
Know Items						
#2	1.57	-2.41	-0.86	0.55	1.85	0.72
#10	2.45	-1.08	-0.33	0.46	1.28	1.79
#14	2.29	-1.33	-0.48	0.31	1.49	1.57
#18	2.96	-1.27	-0.61	0.26	1.30	2.56
#23	2.09	-1.89	-0.79	0.15	1.36	1.28
#34	1.87	-1.77	-0.80	0.30	1.68	1.04
Remember Items						
#8	1.53	-2.28	-0.73	0.34	2.00	0.71
#9	1.58	-2.32	-1.06	0.03	1.56	0.75
#13	2.02	-2.11	-1.12	-0.18	1.24	1.21
#15	1.63	-1.55	-0.40	0.41	1.53	0.82
#19	1.51	-2.31	-1.32	-0.42	0.98	0.71
#30	2.59	-1.25	-0.47	0.38	1.36	1.95
#31	4.09	-1.19	-0.48	0.15	1.08	4.62

#32	3.65	-1.30	-0.42	0.41	1.37	3.60

Note. $N = 290$. α = discrimination parameter; b = difficulty parameter. Max IIF = maximum value of the item information function, IRT results for each scale separately as per equating procedure Scherbaum et al., 2006).

Table A5 Test Level Information for the Know and Remember Items (Study 1)

Theta	-2.8	-2.4	-2.0	-1.6	-1.2	-0.8	-0.4	0.0	0.4	0.8	1.2	1.6	2.0	2.4	2.8
Item type															
Know	2.81	4.03	5.78	7.93	9.46	9.87	9.79	9.70	9.56	9.08	9.20	8.16	5.76	3.67	2.39
Remember	3.70	4.95	6.96	11.09	14.79	14.46	15.33	14.73	14.17	13.22	13.86	10.35	6.08	3.69	2.48
Remember items compared to know items															
	1.31	1.23	1.20	1.40	1.56	1.47	1.57	1.52	1.48	1.46	1.51	1.27	1.06	1.01	1.04

Table A6 Estimated Coefficients for Cognition-Based and Affect-Based Trust with Weighted Episodic and Semantic Scales of Transformational Leadership (Study1)

	B	95% CI	*β*
Cognition-based trust			
Weighted Episodic Scale	.96**	.64 to 1.29	.52**
Weighted Semantic Scale	.35*	.04 to .66	.20*
Adjusted R^2			.49
Affect-based Trust			
Weighted Episodic Scale	1.30**	.94 to 1.66	.60**
Weighted Semantic Scale	.32	-.03 to .66	.15
Adjusted R^2			.55

Note: ** p<.01, * p<. 05. B= unstandardized coefficients, β= standardized coefficients. Study 1 N=290.

Table A7 Decision Criteria for Episodic and Semantic Transformational Leadership Scales, Item Inclusion, and Reliability (Study 1)

Decision Criteria	**Number of Remember Items (TL)**	**Alpha**	**Number of Know Items (TL)**	**Alpha**	**Difference in R/K Proportion**
60%	2 (#13, 32)	.73	2 (23 , 34)	.61	.20
59%	3 (#13, 15 , 32)	.76	3 (#2 , 23 , 34)	.70	.18
58%	4 (#13, 15, ,30, 32)	.82	5 (#2 , 10 ,18, 23, 34)	.84	.16
57%	4 (#13 , 15, 30, 32)	.82	5 (#2 , 10 ,18 , 23 , 34)	.84	.14
56%	6 (#8 , 9 , 13, 15, 30, 32)	.85	5 (#2, 10 ,18 , 23, 34)	.84	.12
55%	8 (#8 , 9, 13, 15, 19, 30, 31, 32)	.89	6 (#2 ,10, 14, 18, 23 , 34)	.87	.10
54%	8 (#8 , 9 , 13 , 15 , 19 , 30 , 31, 32)	.89	6 (#2 ,10, 14, 18, 23, 34)	.87	.08
53%	10 (#6, 8, 9, 13, 15, 19, 29, 30, 31 , 32)	.89	7 (#2 ,10, 14, 18, 23, 25, 34)	.86	.06
52%	10 (#6, 8, 9, 13 , 15 , 19 , 29 , 30 , 31 , 32)	.89	8 (#2 , 10 , 14, 18 , 21, 23, 25 , 34)	.88	.04
51%	12 (#6,, 8 , 9 , 13 , 15 , 19 , 26, 29 , 30 , 31 , 32, 36)	.91	8 (#2 , 10, 14, 18, 21, 23, 25, 34)	.88	.02

Table A8: Estimated Coefficients for Cognition-Based and Affect-Based Trust with Episodic and Semantic Scales of Transformational Leadership with Different Decision Criteria

	57% Criteria			55% Criteria			53% Criteria		
Cognition-based trust	B	95% CI	β	B	95% CI	β	B	95% CI	β
Episodic Scale	.39**	.25 to .54	.40**	.55**	.36 to .74	.52**	.41**	.20 to .61	.36**
Semantic Scale	.35**		.34**	.21*	.03 to .39	.20*	.38**	.18 to .58	.35**
Adjusted R^2			.50			.49			.47
Affect-based trust									
Episodic Scale	.33**	.16 to .50	.29**	.77**	.56 to .98	.62**	.86**	.64 to 1.08	.65**
Semantic Scale	.56**	.39 to .74	.47**	.17	-.03 to .37	.14	.14	-.08 to .35	.11
Adjusted R^2			.52			.55			.56

Note: ** p<.01, * p<.05. B= unstandardized coefficients, β= standardized coefficients. N=290

Figure A1 Confidence Intervals for Beta Coefficients for Episodic and Semantic Scales of Transformational Leadership (Study 1)

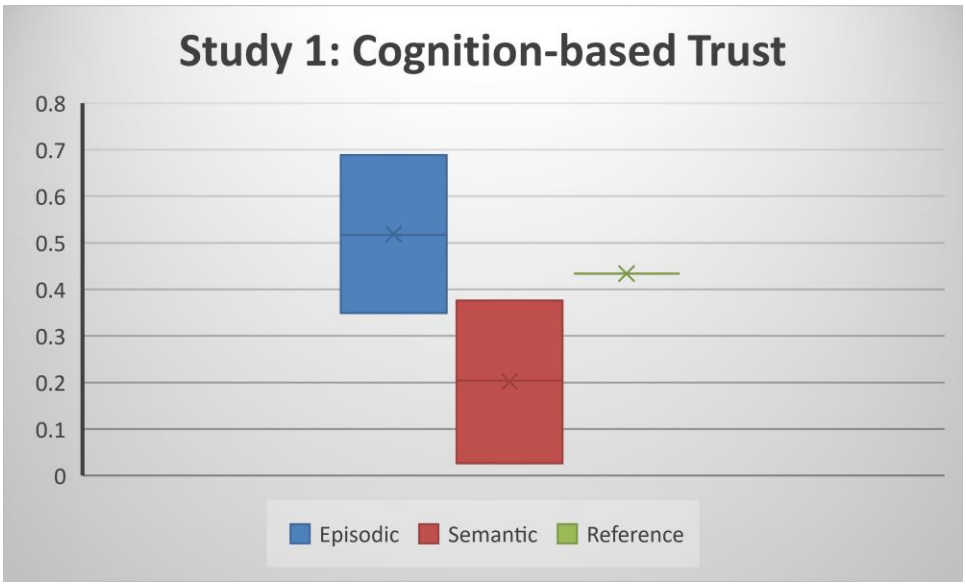


Figure A2 Confidence Intervals for Beta Coefficients for Episodic and Semantic Scales of Transformational Leadership (Study 1)

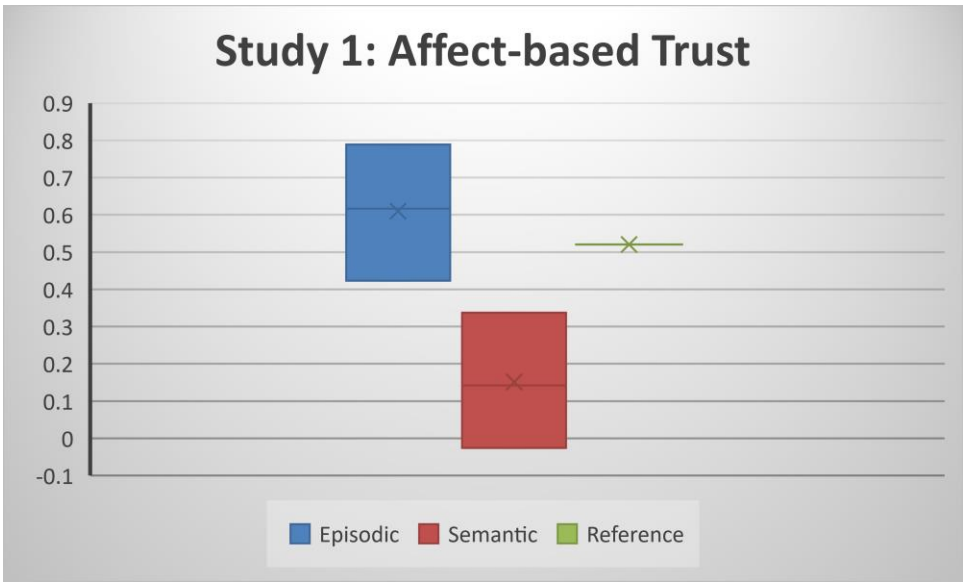


Figure A3 Confidence Intervals for Beta Coefficients for Episodic and Semantic Scales of Transformational Leadership (Study 2)

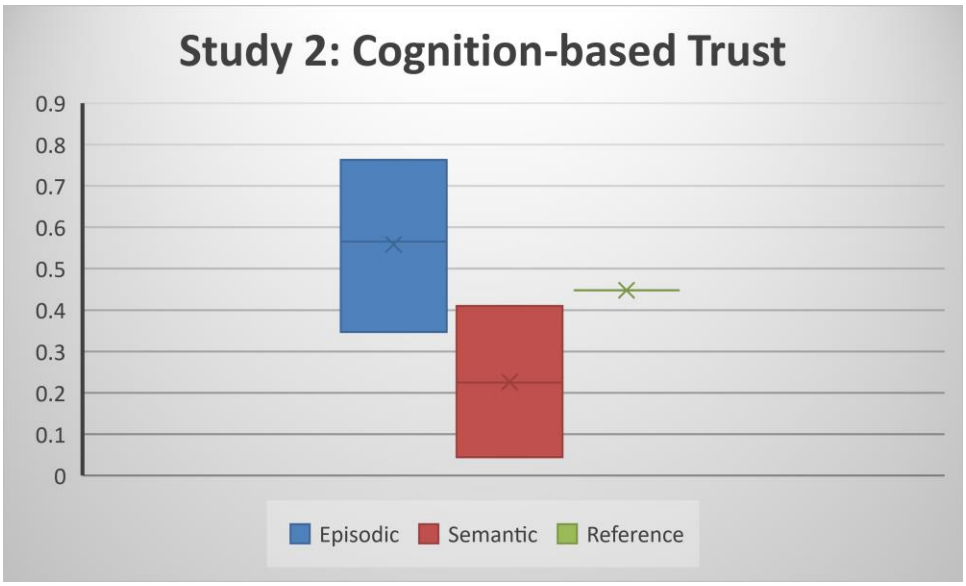


Figure A4 Confidence Intervals for Beta Coefficients for Episodic and Semantic Scales of Transformational Leadership (Study 2)

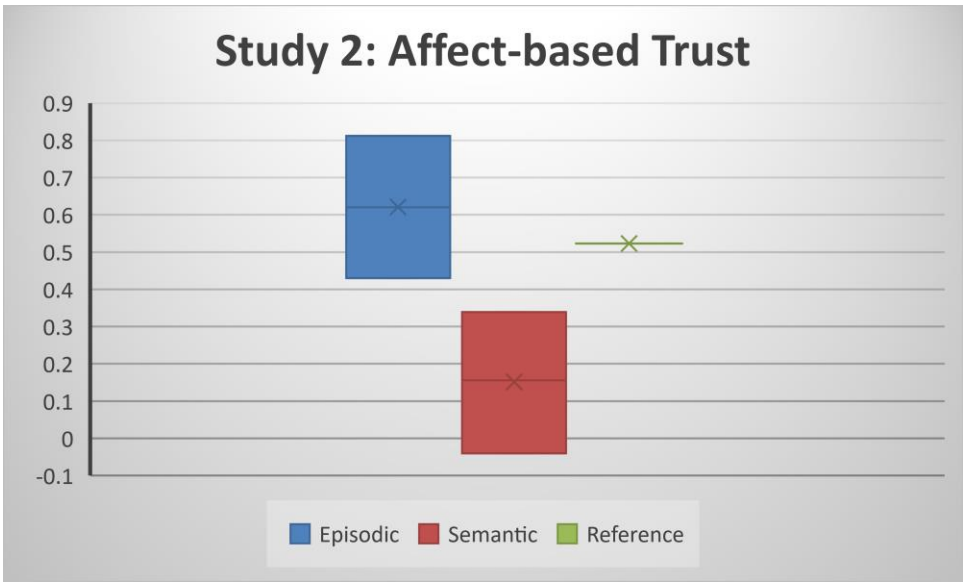


Figure A5 Confidence Intervals for Beta Coefficients for Episodic and Semantic Scales of Servant Leadership (Study 3)

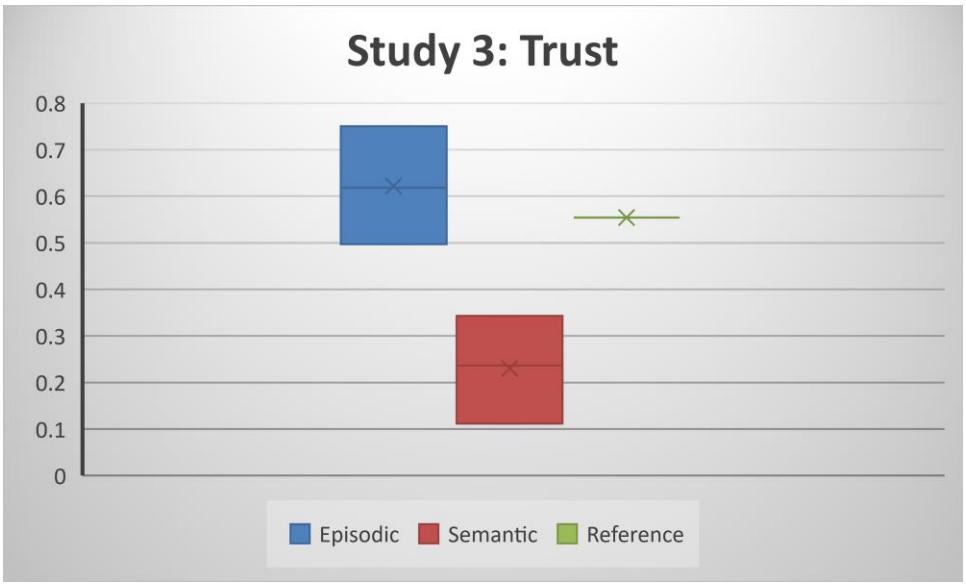


Figure A6 Confidence Intervals for Beta Coefficients for Episodic and Semantic Scales of Servant Leadership (Study 4)

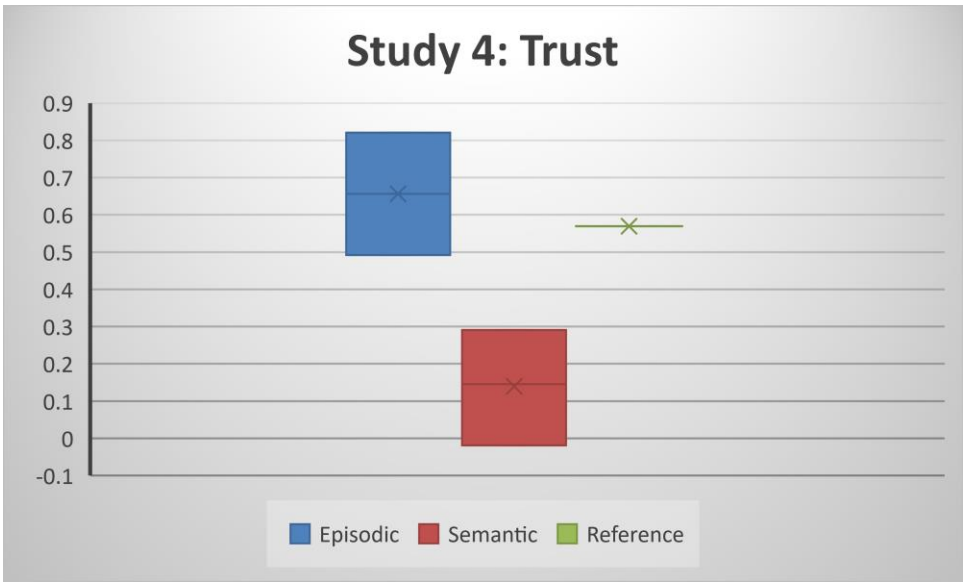


Figure A7 Confidence Intervals for Beta Coefficients for Episodic and Semantic Scales of Servant Leadership (Study 5)

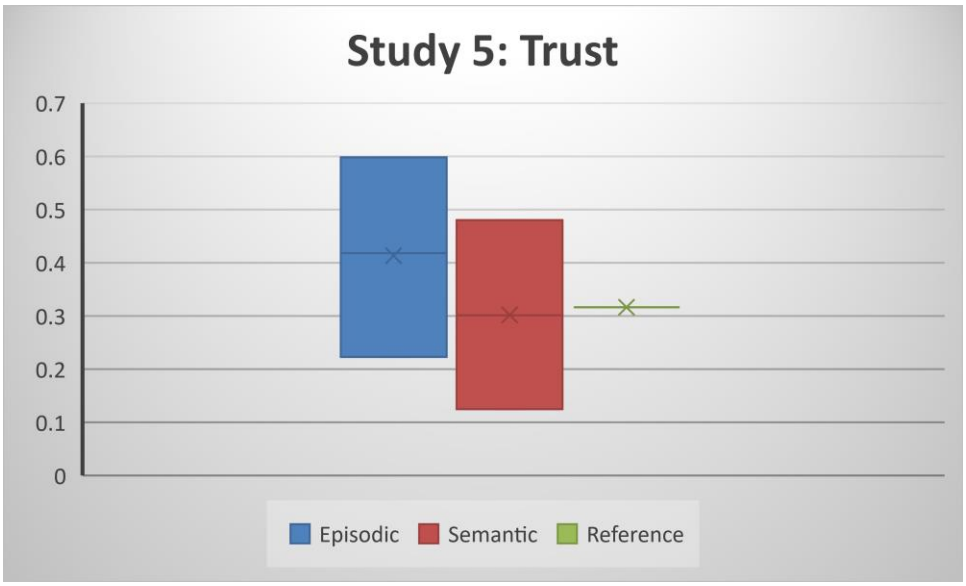


Figure A8 Subject Removal Process for Study 6