The Need to Evaluate

W. Blair G. Jarvis and Richard E. Petty
Ohio State University

Five studies tested the hypothesis that stable individual differences exist in the chronic tendency to engage in evaluative responding. In 2 studies, the 16-item Need to Evaluate Scale (NES) was developed and demonstrated to possess high internal consistency, a single factor structure, high test-retest reliability, and convergent and discriminant validity. Three additional studies supported the predictive validity of the NES. In one, high-NES participants were more likely to report having attitudes toward a variety of important social and political issues than low-NES participants. In another study, high-NES participants wrote more evaluative thoughts in a free thought listing about unfamiliar paintings than low-NES participants. In a final study, high-NES participants wrote more evaluative thoughts in a free thought listing about a typical day in their lives than low-NES participants. Implications for research in social and personality psychology are discussed.

Strictly speaking, nearly all cognition and perception is evaluative. . . . It is not possible to view a social object or a social act without the same time making an assessment on dimensions closely corresponding to good/bad, pleasant/unpleasant, etc. (Markus & Zajone, 1985, p. 210)

A pervasive evaluative factor in human judgment regularly appears first . . . the attitudinal variable in human thinking, based as it is on the bedrock of rewards and punishments both achieved and anticipated, appears to be primary. (Osgood, Suci, & Tannenbaum, 1957, p. 72)

As the citations above illustrate, evaluation, defined as the assessment of the positive and/or negative qualities of an object, is assumed to be among the most pervasive and dominant human responses. The vast amount of research directed at understanding how people evaluate social issues, themselves, outgroups, future behaviors, products, and other objects that they encounter on a daily basis exemplifies how pervasive evaluative responding is considered to be. As we explain shortly, when moderators of evaluative responding have been proposed, these moderators typically have been conceptualized as situational variables or variables associated with particular attitude objects or classes of objects. Importantly, the possibility that people reliably vary in both the likelihood and extent of evaluative responding across situations and objects seems to have escaped consideration in the literature. It has neither been proposed nor demonstrated that the extent or dominance of evaluative responding can be a function of the individual, independent of the particular situational factors or specific attitude objects encountered. Thus, the primary goal of this article is to examine the hypothesis that individuals differ in the extent to which they chronically engage in evaluative responding. We refer to this individual difference as the need to evaluate.1

To examine this hypothesis, we first evaluate the empirical evidence and theoretical assertions often used to support the view that evaluation is a pervasive and dominant response for most people in most situations. We argue that despite the research and theory used to support the pervasiveness of evaluative responding, the possibility of individual differences remains quite plausible. We then describe the development of the Need to Evaluate Scale (NES)—a measure designed to assess individual differences in the propensity to engage in evaluation. After examining the psychometric properties of the NES, we discuss this conceptualization in more detail later and examine whether such a conceptualization is adequate to explain our obtained data.

1 The term "need" is used here as it was intended to be understood in the development of the Need for Cognition Scale (Cacioppo & Petty, 1982), the Personal Need for Structure Scale (Neuberg & Newsome, 1993), and the Need for Closure Scale (Webster & Kruglanski, 1994). That is, we do not assume any biological basis for this individual difference, though one is possible. We presume initially that the underlying nature of the need to evaluate will be similar to what McClelland, Koestner, and Weinberger (1989) referred to as a self-attributed motive (i.e., a motive based in the self-concept and typically assessed through self-report). We discuss this conceptualization in more detail later and examine whether such a conceptualization is adequate to explain our obtained data.
onstrate its convergent and discriminant validity through comparison with other established individual difference measures. Then, supporting the predictive validity of the NES, we demonstrate that those high in the need to evaluate are more likely to have formed attitudes toward a variety of social and political issues. Two additional studies further demonstrate the predictive validity of the NES while offering an opportunity to explore in more detail the nature of the need to evaluate. In these studies, individuals high in the need to evaluate are shown to engage in greater spontaneous evaluative responding than those who are low. Finally, we discuss the possible nature of the need to evaluate in light of our data, examine the limitations of the current research, and discuss the implications of the NES for theory and research.

The Presumed Dominance of Evaluation in Human Judgment

There are at least four sources of evidence for the assertion that evaluation is a pervasive and dominant response for most people across the many situations and objects they encounter. This evidence comes from: (a) the factor analysis of adjective ratings, (b) research on the automatic activation of attitudes, (c) the common observation that most people can easily report their attitudes toward a wide variety of objects when asked, and (d) a compelling though primarily theoretical literature on the functionality of attitudes.

Factor Analysis of Adjective Ratings

In their classic work on the dimensions of meaning in human judgment, Osgood et al. (1957) consistently found that evaluation emerged as the central dimension of meaning in judgments of a wide variety of objects. In these studies, participants were typically presented with various objects (e.g., my mother; symphony) and were to rate each object on up to 76 bipolar adjective scales. Factor analyses of these data sets consistently revealed a dominant first factor which accounted for between half and three-quarters of the total variance among ratings. On this dominant factor, which Osgood et al. labeled the “evaluative factor,” loaded scales such as good/bad, nice/awful, pleasant/unpleasant, and other clearly evaluative dimensions. Collectively, the second and third factors to emerge, labeled “potency” and “activity” respectively, consistently accounted for only half of the variance that could be accounted for by the evaluative factor.

These factor analytic results in no way attenuate the likelihood of individual differences in evaluative responding. This is because a dominant evaluative factor could have occurred in one of two ways. First, as commonly assumed, the evaluative factor could have accounted for a similar proportion of variance in most participants’ judgments (i.e., with each participant having equally strong evaluative responses across the set of objects rated). Alternatively, the same apparent dominance could have emerged if the evaluative factor was accounting for an even greater proportion of variance in the judgments of some participants, while accounting for a proportionately smaller amount of variance in other participants’ judgments.

Even if it turned out that the evaluative factor accounted for an equivalently large proportion of variance for all participants, this would not mean that all participants were equally likely to engage in spontaneous evaluative responding. In the Osgood et al. (1957) research, participants were never given the opportunity to choose the semantic differentials that they felt characterized their judgments about the objects. In all of the factor analyses reported, participants were required to make judgments about the object on each of the scales provided. Thus, what the Osgood et al. findings reflect is the shared meaning between the adjective scales and not necessarily the meaning elicited spontaneously by any given object or set of objects. Even if many of the participants in these studies would not have otherwise evaluated the objects presented, the fact that they were able to evaluate the objects, if asked, is a sufficient condition to produce the dominance of the evaluative factor.

Automatic Evaluative Responding

A second line of research appearing to support the pervasiveness and dominance of evaluative responding comes from relatively recent studies on the automatic activation of attitudes (Bargh, 1994; Bargh, Chaiken, Govender, & Pratto, 1992; Fazio, 1995; Fazio, Powell, & Herr, 1983; Fazio, Sanbonmatsu, Powell, & Kardes, 1986). This work has provided compelling evidence that attitudes (evaluations) are often activated from memory automatically on mere exposure to an attitude object and that the magnitude of this activation is proportional to the strength of the attitude. Although this evidence offers compelling support for the pervasiveness of attitudes toward common objects (Bargh et al., 1992), as yet there has been no evidence to suggest either (a) that such effects and their magnitudes are similar across most individuals, or (b) that there is consistency between individuals in the relative dominance of evaluative, relative to other, forms of responding.

First, as with the Osgood et al. (1957) results, one cannot tell from the reported analyses (which collapse over participants) whether the effect is being driven by most participants, or by some subset for whom the effect is quite strong. Furthermore, all of the objects examined in the automatic activation paradigms have been objects with which most people are quite familiar (e.g., cockroaches, ice cream). As Bargh (1994) argued, an object might need to elicit only minimal evaluative responding in order to achieve an automatic evaluative association in memory. Given the high frequency of exposure to these objects used in these studies, it would not be surprising that even those who infrequently engage in spontaneous evaluation had evaluated these objects at least a sufficient number of times to cause an evaluative association in memory. Presumably, some people might initially evaluate an object and then seldom if ever engage in evaluation of that object again. In contrast, other people could retrieve and update their evaluation of that object on a regular basis (Petty, Jarvis, & Evans, 1996). Thus, even though most people might evaluate objects to which they are frequently exposed to an extent that is sufficient to produce automatic activation of that evaluation in memory, the extent of such responding could still reliably vary between individuals.

In addition, we should also recognize that the automatic activation of evaluative associations does not imply the domi-
nance of evaluative thought over other types of thought across all people. According to most network models of memory (e.g., Anderson, 1983; Collins & Loftus, 1975), most object representations in memory have multiple associations (e.g., color, size, etc.) to which activation will spread on initial activation of the object. The automatic attitude activation effect primarily demonstrates that for common objects, one such association is evaluative. However, these findings do not suggest the strength or precedence of this association relative to the other object associations in memory. It seems reasonable that although most people would have an evaluative association with many common objects in memory, the dominance of this association relative to others could generally be greater for some people than for others.

**Ease of Evaluative Responding**

Further evidence for the prevalence of evaluative responding comes from the common observation of the invariable ease with which people are able to report attitudes toward an enormous array of stimuli. In fact, this common observation has previously been offered as support for the pervasiveness of attitudes (e.g., Greenwald, 1989). Certainly, if the ability of participants to report attitudes were any indication of the prevalence of attitudes, one would have to conclude that nearly all people do in fact have attitudes about nearly everything. As an illustrative case, consider Bargh et al.'s (1992) study using the automatic attitude activation paradigm. In this study, participants were asked to report their attitudes toward 92 objects. The mean response latency for indicating one's attitude toward each of these objects tended to fall in the range of 600 to 900 ms, with less than one percent of all trials taking more than 2.5 s. The ability of participants to report an evaluation is certainly not limited to common objects. People appear to have no trouble reporting attitudes toward unfamiliar attitude objects about which they have engaged in relatively little thought, such as comprehensive exams for college seniors (e.g., Petty, Cacioppo, & Goldman, 1981), the XT-100 answering machine (Chaiken & Mateswaran, 1994), and even objects of negligible meaning such as nonsense syllables, ideographs, and faces of strangers (Zajonc, 1968).

Thus, the relative ease with which people can report attitudes toward a variety of objects is clear. However, whether this suggests that evaluative responding is equally pervasive among all people is not. Furthermore, these results do not indicate whether respondents would have engaged in evaluation if this was not required by the task. As suggested earlier in regard to the Osgood et al. (1957) analyses, the fact that people can evaluate with ease does not necessarily imply that they had an a priori evaluation to give, or that they would have evaluated the same object spontaneously had they not been asked to do so. The question of whether reported attitudes reflect a priori evaluations is not a new issue. Converse (1964, 1970) argued that people would often report attitudes when they have none to give—even when the option of not responding is made quite clear. Converse showed that the inconsistencies in a substantial portion of participants' attitude reports over time could best be accounted for by a model positing that these responses were essentially random, and offered the idea that these people actually possessed "non-attitudes" (see also Fazio, 1995). Converse (1970) suggested that the prevalence of nonattitudes in some individuals was likely due to ability factors, such as lack of knowledge about or attention given to particular objects. However, this difference is plausibly a function, at least in part, of individual differences in the motivation to evaluate.

**Attitudes Are Functional**

Another factor contributing to the assumption that evaluation is a pervasive response is the notion that general evaluations (i.e., attitudes) serve various important and even essential functions for people (e.g., Katz, 1960; Herek, 1987; Shavitt, 1989; Smith, Bruner, & White, 1956). For example, theorists have argued that attitudes are adaptive in that they (a) prepare individuals to respond to and cope with life events (e.g., Allport, 1935; Fazio, 1995), (b) increase control over positive and negative outcomes (e.g., Festinger, 1954; Katz, 1960; Petty & Cacioppo, 1986a; Pittman, 1993), (c) facilitate social interaction (e.g., Chaiken, Liberman, & Eagly, 1989; Smith et al., 1956; Snyder & Debono, 1985), (d) help one's environment to be more organized, structured, and understandable (e.g., Katz 1960), (e) protect and enhance one's self-image and esteem (e.g., Katz, 1960), (e) allow expression of personal values (e.g., Katz, 1960; Snyder & Debono, 1985), and so forth. With such benefits to be had, it certainly seems reasonable that people should be chronically motivated to evaluate. However, even if all individuals were highly motivated by the functions described above, evaluation is not the only way to satisfy these motives. For example, the motive to achieve knowledge and understanding could be satisfied by evaluating the object, but the individual could presumably also satisfy the knowledge motive through other nonevaluative means such as by attempting to analyze the object's structural properties or functional relations to other objects. In fact, some researchers have argued that the preferred means of satisfying the motive to gain control over one's environment is through an effortful analysis of causal relations rather than through evaluation (Edwards & Weary, 1993; Gleicher & Weary, 1991; Pittman & D'Agostino, 1985; but see Pittman, 1993).

**Theoretical Basis for Expecting Individual Differences**

In addition to the fact that current research and theory do not preclude individual differences in the propensity to evaluate, there are also reasons to expect such differences. For example, Kelly (1955) argued that individual differences in the chronic use of social category information would develop out of the variance in people's personal histories. Kelly argued that variations in peoples' frequency of experience with social constructs (e.g., kindness, selfishness) would lead to corresponding variations in the likelihood that they would apply such constructs in the screening of input stimuli. For example, those frequently exposed to "kind" behaviors should become more

---

2 We are not suggesting a "demand" interpretation of this data. We are simply suggesting that many of the evaluations people report, however real, do not require an a priori existence and could be generated or constructed in response to an attitudinal inquiry (Petty, Priester, & Wegener, 1994; Wilson & Hodges, 1992).
likely to apply the construct “kind” to future behaviors than those who have infrequently experienced kindness. Because it seems quite reasonable that individuals would vary in the extent to which they experience and are exposed to positive and negative objects and events in their lives, it seems to follow that this variation could lead in a similar way as described by Kelly to chronic differences in the likelihood that these evaluative constructs would be applied to input stimuli in the future.

Additionally, it has been argued that variations in verbal exposure to social constructs can lead to an increased likelihood of their chronic application. Higgins and King (1981) cited the example of the child whose parents frequently comment on other people’s attractiveness. In addition to the traditionally expected social learning and modelling effects (which should also produce individual differences), they argue that the construct “attractive” would likely become chronically accessible for this child. In the same way, children raised in environments where evaluative constructs are frequently and openly applied by family and peers should be more susceptible to chronic activation of these constructs than others. Although chronic activation of a construct does not necessarily imply its chronic usage, these two events are likely to be highly correlated (Higgins, 1989). In fact, some evidence supports the view that general procedures for applying evaluative constructs are also subject to varying degrees of automaticity (see Bargh, 1990; Smith 1984, 1989).

In addition to these rather cognitive bases for individual differences in evaluative responding, there are reasons to expect motivational bases as well. First, it could be predicted on the basis of self-perception theory (Bem, 1972) that those who engage in frequent evaluation (e.g., for reasons discussed above) would infer that this is something they like to do—thus providing a motive for further evaluative responding. Second, recall that functional theorists have argued that attitudes provide various incentives for people to engage in evaluative responding. Rewards such as enhanced control over one’s environment, positive social interaction, enhanced self-image and self-esteem, enhanced opportunity for self-expression, and an enhanced understanding of one’s environment would seem to provide a motivational impetus for evaluative responding. Given that research has supported that people are differentially motivated by such rewards (e.g., desire for control, self-monitoring, and need for structure) it is reasonable that people would be differentially motivated to engage in evaluation as a means to attain them.

Importance of Individual Differences in Evaluative Responding

The goal of the present research is to examine the hypothesis that some people are consistently more prone to engage in evaluation than others. There are a number of reasons why examining this hypothesis is important. First, it is a basic research issue as to whether or not individual differences in evaluative responding exist. If such differences can be demonstrated, then this would offer the first empirical support for the contention that the assumed pervasiveness and dominance of evaluative responding depends at least in part on the individual. If such individual differences in evaluative responding can be demonstrated, then this raises important questions, such as why such differences occur, how they develop, and so forth.

The need-to-evaluate construct also could help to identify those individuals who, in day to day life, will spontaneously engage in the sorts of evaluative behaviors that are demanded by many of the situations found in social science investigations. In attitude and survey research, for example, people are instructed to report their evaluations of social issues, political candidates, and so forth, usually without regard to the possibility that people might not have spontaneously evaluated these objects. If the elicited evaluative responses are better indicators of spontaneous attitudes for some people than others, the need-to-evaluate construct could be very useful in identifying those individuals for whom psychological theories will generalize beyond the experimental situation.

In addition, if people vary in their tendency to engage in evaluative responding, the need-to-evaluate construct could be expected to moderate well-established evaluative phenomena, especially in those cases where evaluation is not explicitly demanded by the situation. As just one example, Tesser and his colleagues have shown that merely thinking about an object can polarize evaluations (see Tesser, 1978; Tesser, Martin, & Mendolia, 1995, for reviews). Because the effects of mere thought on attitude polarization presumably depend on the spontaneous generation of evaluative, relative to other, types of thought, an individual difference measure that captured such a tendency would predict a greater effect of mere thought for those who chronically engaged in evaluative responding.

There might also be meaningful practical implications of a need to evaluate. For example, based on research that has shown evaluative responding to have positive health implications (see Fazio, 1995), identifying one’s chronic tendency to either engage in or avoid evaluative responding might help to account for variance in an individual’s ability to cope with stress and negative life events. In addition to the potential health implications of the need to evaluate, there also are implications for advertisers and government agencies who spend billions of dollars each year attempting to influence public attitudes. Specifically, some of this communication might not have the intended effect on its recipients—at least on those for whom evaluation is not a dominant response.

In sum, we argued that the existence of stable individual differences in evaluative responding is plausible and remains unexamined. We further argued that such an individual difference, if measurable, would serve as a useful contribution to the field, both theoretically and practically. Five studies are reported and discussed that were designed to measure, validate, and examine the nature and implications of individual differences in the need to evaluate.

Construction of the Need to Evaluate Scale

The first stage of scale construction involved generating items for the Need to Evaluate Scale (NES). For this purpose, data from 357 male and female participants were obtained in four pilot studies. All participants were enrolled in introductory psychology classes at Ohio State University, and all participated for partial course credit. Initial scale construction was an iterative process whereby items were generated, tested, and analyzed, and then either retained for further testing, revised, or deleted.
Through this process we reduced a pool of 46 items to a highly reliable 16-item scale (Cronbach's $\alpha = .87$). The retained items are listed in Table 1. In each pilot study, participants were asked to rate the extent to which each of the scale items were characteristic of them on a 5-point scale labeled extremely uncharacteristic (1), somewhat uncharacteristic (2), uncertain (3), somewhat characteristic (4), and extremely characteristic (5). The usefulness of this particular response scale has been demonstrated previously in the development and validation of the Need for Cognition Scale (Cacioppo & Petty, 1982). Items were assessed according to conventional a priori standards (e.g., see Allen & Yen, 1979; Briggs & Cheek, 1986; Comrey, 1988). Necessary statistical criteria for item retention were: (a) an item-total correlation of greater than .30, (b) an average interitem correlation of greater than .20, (c) a mean greater than 2 and less than 4 on the 5-point scale, and (d) a standard deviation of at least 1. Items also were required to meet the following qualitative criteria: (a) face validity, (b) clarity of meaning, and (c) a presumed unitary relation with the proposed construct. The next goal was to thoroughly examine the psychometric properties of the 16-item NES.

Study 1: Psychometric Properties of the NES

Our primary goals in the first study were: (a) to assess the factor structure of the NES, (b) to determine the internal consistency of the scale, and (c) to assess the stability of participants' scores on the NES over a 10-week period.

### Method

#### Participants and Procedure

As in the pilot studies, all participants were enrolled in introductory psychology classes at Ohio State University and participated for partial course credit. Data for the first sample came from the responses to the 16 retained items provided by the 131 female participants in the last pilot study. Because this sample included only female participants, the same items were administered to a second sample of 77 male and 83 female participants, and a third sample of 138 male and 128 female participants. This allowed the opportunity both to replicate the obtained factor structure, and to ensure that the results were not gender specific. Participants were tested in large groups ranging in size from 35 to 55. As in the pilot studies, participants were asked to rate the extent to which each of the 16 scale items were characteristic of them on a 5-point scale (described above). In each of the three samples, the covariances between each of the 16 items of the NES were submitted to an iterative principal factors analysis. Confirmatory factor analyses (described in detail below) were subsequently performed on these data to test and compare the plausibility of four alternative factor structures. The internal consistency of the 16-item scale in each sample also was assessed. Finally, the scale was readministered to a subsample of participants ($n = 70$) 10 weeks later to determine the test-retest reliability of the NES.

---

### Table 1

Need to Evaluate Scale, Factor Loadings, and Cronbach’s Alpha From Three Samples

<table>
<thead>
<tr>
<th>Item</th>
<th>Item wording</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I form opinions about everything.</td>
<td>.61</td>
<td>.58</td>
<td>.51</td>
</tr>
<tr>
<td>2</td>
<td>I prefer to avoid taking extreme positions. (R)</td>
<td>-.47</td>
<td>-.60</td>
<td>-.38</td>
</tr>
<tr>
<td>3</td>
<td>It is very important to me to hold strong opinions.</td>
<td>.75</td>
<td>.72</td>
<td>.71</td>
</tr>
<tr>
<td>4</td>
<td>I want to know exactly what is good and bad about everything.</td>
<td>.48</td>
<td>.45</td>
<td>.42</td>
</tr>
<tr>
<td>5</td>
<td>I often prefer to remain neutral about complex issues. (R)</td>
<td>-.55</td>
<td>-.52</td>
<td>-.59</td>
</tr>
<tr>
<td>6</td>
<td>If something does not affect me, I do not usually determine if it is good or bad. (R)</td>
<td>-.39</td>
<td>-.40</td>
<td>-.43</td>
</tr>
<tr>
<td>7</td>
<td>I enjoy strongly liking and disliking new things.</td>
<td>.59</td>
<td>.52</td>
<td>.47</td>
</tr>
<tr>
<td>8</td>
<td>There are many things for which I do not have a preference. (R)</td>
<td>-.42</td>
<td>-.53</td>
<td>-.35</td>
</tr>
<tr>
<td>9</td>
<td>It bothers me to remain neutral.</td>
<td>.52</td>
<td>.56</td>
<td>.56</td>
</tr>
<tr>
<td>10</td>
<td>I like to have strong opinions even when I am not personally involved.</td>
<td>.71</td>
<td>.64</td>
<td>.76</td>
</tr>
<tr>
<td>11</td>
<td>I have many more opinions than the average person.</td>
<td>.61</td>
<td>.66</td>
<td>.59</td>
</tr>
<tr>
<td>12</td>
<td>I would rather have a strong opinion than no opinion at all.</td>
<td>.69</td>
<td>.53</td>
<td>.47</td>
</tr>
<tr>
<td>13</td>
<td>I pay a lot of attention to whether things are good or bad.</td>
<td>.49</td>
<td>.37</td>
<td>.33</td>
</tr>
<tr>
<td>14</td>
<td>I only form strong opinions when I have to. (R)</td>
<td>-.54</td>
<td>-.66</td>
<td>-.53</td>
</tr>
<tr>
<td>15</td>
<td>I like to decide that new things are really good or really bad.</td>
<td>.42</td>
<td>.45</td>
<td>.38</td>
</tr>
<tr>
<td>16</td>
<td>I am pretty much indifferent to many important issues. (R)</td>
<td>-.47</td>
<td>-.35</td>
<td>-.35</td>
</tr>
<tr>
<td>Cronbach's Alpha</td>
<td>.87</td>
<td>.86</td>
<td>.83</td>
<td></td>
</tr>
</tbody>
</table>

Note. Factor loadings are standardized regression coefficients from single-factor solutions obtained from iterative principle factor analyses. The analyzed matrices contained covariances between raw scores—no items were reverse scored prior to analysis. This affects only the direction of loadings and not their magnitudes. Reverse scoring (R) would normally be used for this item.
Results and Discussion

Exploratory Factor Analyses

Sample 1. The first sample consisted of 131 female introductory psychology students. Exploratory factor analysis of the data provided by this group revealed a dominant first factor. The eigenvalues obtained from this analysis are presented in Figure 1. The first factor accounted for 41% of the total scale variance, and 77% of the total common variance (i.e., excluding unique variance associated with each item). This can be contrasted with the second strongest factor which accounted for only 9% of the total variance and 17% of the common variance. Given a single factor solution, the factor loadings also were highly consistent with those predicted (see Table 1, Sample 1). All loadings were in the expected directions, and all were at least moderately strong (absolute values ranging from .39 to .75). Thus, this analysis provides at least good preliminary evidence that the variance in the scale is being driven primarily by one major factor.

Sample 2. The data for the second sample were obtained from 83 female and 77 male introductory psychology students. Men ($M = 53.21$) and women ($M = 51.05$) did not differ in their average scores on the NES, $F(1, 158) = 1.67, p = .20$. An examination of the eigenvalues for the preliminary factor analysis (see Figure 1), reveals that the dominance of the first factor in the first sample was not dependent on its all-female composition. In the second sample, the first factor accounted for 41% of the total scale variance, and 75% of the total common variance—nearly identical to those proportions obtained with the all-female sample. Again the second strongest factor was far less pronounced, accounting for only 9% of the total scale variance, and for 17% of the common variance.

Again, all factor loadings were consistent with those predicted (see Table 1, Sample 2). As with the first sample, all loadings were in the expected direction and were of at least moderate strength (absolute values ranging from .35 to .72). If the first factor extracted from the data in both samples was truly the same construct, then we also would expect a high degree of correspondence in the factor loadings between samples. As expected, the absolute values of the factor loadings between the first two samples were highly correlated ($r = .69, p < .01$) as were the raw factor loadings ($r = .99, p < .001$).

Sample 3. The data for the third sample were obtained from 128 female and 136 male introductory psychology students. Men ($M = 53.20$) and women ($M = 51.05$) marginally differed in their average scores, $F(1, 264) = 3.49, p = .06$. Consistent with Samples 1 and 2, the data revealed a dominant first factor which accounted for 33% of the total scale variance and 68% of the total common variance (see Figure 1). Although accounting for slightly more variance than in the previous samples, the second factor again was far less pronounced, accounting for 12% of the total variance and 24% of the total common variance.

All factor loadings again were consistent with those predicted (see Table 2, Sample 3). All were in the expected direction and again were all of at least moderate strength (absolute values ranging from .33 to .71). If the first factor extracted represents the same construct as that extracted in the first two samples, there should be a high degree of correspondence between the loadings on the first factor across the three samples. Strong support for factor structure replication was obtained. The absolute values of the factor loadings were highly correlated with those obtained in both Sample 1 ($r = .80, p < .001$) and Sample 2 ($r = .76, p < .001$). The raw factor loadings were also highly correlated with those obtained in both Sample 1 ($r = .99, p < .001$) and Sample 2 ($r = .99, p < .001$).

Two-factor solution. Although a single dominant factor seems to account for the majority of variance in the scale items, a second factor nevertheless has emerged in each of the three samples and has accounted for approximately 16–24% of the shared variance. To examine the nature of this second factor, we collapsed all of the data from the three samples. Again using an iterative principle factors analysis, we extracted two factors, and used a direct quartimin rotation (Jennrich & Sampson, 1969; see also Browne, 1988) to determine the loadings of the scale items on each factor (see Table 2). After examining the loadings, we generated two hypotheses to account for the source of the additional common variance.

Hypothesis 1: Correlated measurement error. The first difference noted between items loading on the two factors was that with one exception, all of the items loading on the second factor were the reverse-scored items. This suggested that the second factor might be a methodological one. A number of studies have shown that when participants are asked multiple questions on the same rating scale, the correlation between those items will frequently be artificially inflated in a positive direction due to correlated measurement error between scale items (Alwin, 1969; see also Browne, 1988). The absolute values of the additional common variance were also highly correlated with those obtained in both Sample 1 ($r = .99, p < .001$) and Sample 2 ($r = .99, p < .001$).

3 All analyses reported were also performed separately on each sample. Results were highly consistent across samples and so the data were merged into a single data file ($n = 557$).

4 Promax and Harris-Kaiser rotation methods were also examined to ensure the obtained solution was not method dependent. Results across the three methods of rotation were highly consistent.
loadings for Two-Factor Solution for 16 Retained Items (n = 557)

<table>
<thead>
<tr>
<th>Item</th>
<th>Item wording</th>
<th>Factor 1 loading</th>
<th>Factor 2 loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I form opinions about everything</td>
<td>.43</td>
<td>-.24</td>
</tr>
<tr>
<td>3</td>
<td>It is very important to me to hold strong opinions.</td>
<td>.55</td>
<td>-.30</td>
</tr>
<tr>
<td>4</td>
<td>I want to know exactly what is good and bad about everything.</td>
<td>.55</td>
<td>.01</td>
</tr>
<tr>
<td>7</td>
<td>I enjoy strongly liking and disliking new things.</td>
<td>.67</td>
<td>.05</td>
</tr>
<tr>
<td>10</td>
<td>I like to have strong opinions even when I am not personally involved.</td>
<td>.50</td>
<td>-.35</td>
</tr>
<tr>
<td>11</td>
<td>I have many more opinions than the average person.</td>
<td>.43</td>
<td>-.30</td>
</tr>
<tr>
<td>12</td>
<td>I would rather have a strong opinion than no opinion at all.</td>
<td>.43</td>
<td>-.23</td>
</tr>
<tr>
<td>13</td>
<td>I pay a lot of attention to whether things are good or bad.</td>
<td>.56</td>
<td>.10</td>
</tr>
<tr>
<td>15</td>
<td>I like to decide that new things are really good or really bad.</td>
<td>.61</td>
<td>.12</td>
</tr>
<tr>
<td>2</td>
<td>I prefer to avoid taking extreme positions. (R)</td>
<td>-.08</td>
<td>.48</td>
</tr>
<tr>
<td>5</td>
<td>I often prefer to remain neutral about complex issues. (R)</td>
<td>.08</td>
<td>.71</td>
</tr>
<tr>
<td>6</td>
<td>If something does not affect me, I do not usually determine if it is good or bad. (R)</td>
<td>.05</td>
<td>.54</td>
</tr>
<tr>
<td>8</td>
<td>There are many things for which I do not have a preference. (R)</td>
<td>-.06</td>
<td>.44</td>
</tr>
<tr>
<td>9</td>
<td>It bothers me to remain neutral.</td>
<td>.23</td>
<td>-.43</td>
</tr>
<tr>
<td>14</td>
<td>I only form strong opinions when I have to. (R)</td>
<td>-.04</td>
<td>.66</td>
</tr>
<tr>
<td>16</td>
<td>I am pretty much indifferent to many important issues. (R)</td>
<td>.09</td>
<td>.56</td>
</tr>
</tbody>
</table>

Correlation between Factors 1 and 2: -.40

Note. Factor loadings are standardized regression coefficients from a two-factor solution obtained from an iterative principle factors analysis, using a direct quartimin rotation. The analyzed matrix contained covariances between raw scores—no items were reversed prior to analysis. Reverse scoring (R) would normally be used on this item.

Correlated measurement error might produce a second factor for the reverse-scored items (RSIs) for the following reason. First, the likelihood of any set of items loading on a single factor is maximized when all items are maximally correlated with each other (i.e., regardless of valence). When one uses RSIs in a scale, the likelihood of their loading on the same factor as the non-reverse-scored items (NRSIs) is maximized when the two sets of items are maximally negatively correlated. Importantly, a methodological source of positive correlation between all scale items would attenuate the true negative relation of the RSIs with the NRSIs. However, this positive inflation would simultaneously strengthen the relation among the RSIs themselves, and among the NRSIs themselves. Consequently, items that should correlate positively with each other will appear more related than they are (i.e., RSIs with themselves, NRSIs with themselves), and items that should correlate negatively with each other will appear less related than they are (i.e., RSIs and NRSIs with each other). Thus, any correlated measurement error would increase the likelihood that RSIs and NRSIs will load on separate factors.

To test this hypothesis, we could extract any such correlated measurement error using causal modelling techniques. If the cause of the second factor is largely due to correlated measurement error, then a two-factor solution should not fit the data better than a single-factor solution when this method variance is isolated and extracted.

Hypothesis 2: Preference for neutrality. A second common-
In the following analyses, each of these hypotheses is tested using confirmatory factor analysis techniques.

**Confirmatory Factor Analyses**

Although the exploratory factor analyses support the position that a single dominant factor is responsible for the majority of variance in the scale, confirmatory factor analysis techniques that use structural equation modelling allow a better test of this hypothesis (for a discussion, see Judd, Jessor, & Donovan, 1986). In this set of analyses, we examined and compared four alternative models using RAMONA V3.8 (Browne & Mels, 1993). Figure 2 presents these four models. Model A tested the hypothesis that a single factor drives all of the variance in the scale items. Model B made the additional assumption that some constant influence of correlated measurement error exists across all the scale items. Model C tested the alternative hypothesis that rather than measurement error, it is a second conceptual factor that is the source of the additional variance. Finally, Model D examined whether adding a second conceptual factor could improve model fit with method variance extracted. Each of these models was compared to a null model (a model positing no common factors) and, where possible, with each other to assess relative and incremental fits to the data.

**Null model**. Testing the fit of the null model tests the hypothesis that all of the scale items are independent and that no factor exists that accounts for covariation between them. As expected, this model did not fit the data well ($\chi^2 = 2508.76$, $df = 120$, $p < .0001$; Maximum Wishart Likelihood solution). It is important to note that the chi-square statistic obtained for poorly fitting models (e.g., null models) is extremely sensitive to the type of solution used (e.g., Maximum Wishart Likelihood [MWL] Generalized Least Squares [GLS]). Thus, the solution one chooses can have a major impact on the fit indices of alternative models since most involve a comparison with the null model's chi-square statistic. Because in this case the MWL solution indicated a better fit of the null model than the GLS solution, we focus our discussion on the MWL solutions, as these will produce the most conservative fit estimates for the alternative models. However, fit indices calculated as a function of both MWL and GLS solutions are presented in Table 3.

**Model A**. The first alternative to the null model we examined was that of a single factor accounting for all of the covariance between the scale items (see Figure 2 and Table 2, Model A). As expected, this model fit the data substantially better than the null model ($\chi^2_{null} - \chi^2_{Model A} = 1817.70$, $df = 16$, $p < .0001$). However, other supplemental fit indices including Root Mean Square Error of Approximation (RMSEA; Browne & Cudek, 1992; Steiger & Lind, 1980), the Tucker-Lewis Index (Tucker & Lewis, 1973), the Bentler-Bonett Normed Fit Index (Bentler & Bonett, 1980), and the Comparative Fit Index (Bentler, 1990), all indicated this model to be less than optimal. That is, although this model represents a great improvement over the null model, substantial covariance between the scale items remains for which the model does not account. Again, this is not surprising given that the exploratory factor analyses indicated that the first factor accounted for only 75% of the shared variance between scale items. Although this does not put into question the dominance of the first factor, the question remains as to the source of the remaining common variance. With the next three models, the two hypotheses outlined above are tested to determine the source of this remaining variance.

**Model B**. The second alternative model examined was equivalent to Model A but with the additional assumption that there exists in the data some amount of correlated method error, such that some degree of positive correlation across all items is due to measuring all items on the same rating scale. To test this hypothesis, a latent method factor with causal paths leading to all of the scale items was added to Model A (see Figure 2, Model B). Because the hypothesized method factor should theoretically influence each scale item to the same degree, all of these paths were constrained to equality.

The results of this analysis revealed that a small degree of positive correlation did exist across both RSIs and NRSIs. Although the influence of this method factor was not large ($\beta = .30$), when it was taken into account the fit of the single factor solution improved substantially as indicated by the chi-square difference ($\chi^2_{Model A} - \chi^2_{Model B} = 276.80$, $df = 1$, $p < .0001$) and by all supplementary fit indices. Examining the new factor loadings reveals that when this method factor was taken into account, the loadings of the NRSIs became slightly less positive, and the loadings of the RSIs became slightly more negative. This would be expected if correlated measurement error had been attenuating the true negative relation between the RSIs and NRSIs.

**Model C**. The next model examined tested the hypothesis that the two-factor solution accounts for the data better than the single-factor solution (see Figure 2, Model C). To test this hypothesis, two conceptual latent variables were modelled: (1) **NEVAL(+)**, to represent the Need to Evaluate, and (2) **NEVAL(-)**, to represent the Preference for Neutrality. Determination of causal paths followed from the two-factor exploratory analysis reported earlier. Items that loaded more strongly on the first factor in this previous analysis (i.e., items 1, 3, 4, 7, 10, 11, 12, 13, 15) were modeled as indicators of **NEVAL(+)**, and those that loaded more strongly on the second (i.e., items 2, 5, 6, 8, 9, 14, 16) were modeled as indicators of **NEVAL(-)**.

As expected, the two-factor solution fit the data significantly better than did the single-factor solution (Model A) as indicated by the chi-square difference ($\chi^2_{Model A} - \chi^2_{Model C} = 231.54$, $df = 1$, $p < .0001$) and by all supplementary fit indices. However, comparing Model C with Model B reveals that the two-factor solution did not fit the data better than the single-factor solution with method error extracted. In fact, Model B fit the data slightly better than Model C on all measures of fit.\(^7\)

\(^7\) A difference test between chi-square statistics is possible only when one model is nested within the other. This occurs only when all of one model's parameters represents a subset of another model's parameters. Because neither Model B nor Model C is nested within the other, a test of the chi-square difference between these models is not possible.
Model D. Finally, it was possible that both method variance and variance due to a second conceptual factor could be influencing participants’ scores (see Figure 2, Model D). If a second factor (Preference for Neutrality) was truly affecting participants’ scores, then modeling this second factor should enhance model fit even with the method variance extracted. Comparing the results of Model D with Model B reveals that the assumption of a second conceptual factor added negligible improvement when method variance was extracted. First, the chi-square difference between Model D and Model B, although significant, is proportionally small ($\chi^2_{\text{Model D} - \chi^2_{\text{Model B}} = 11.63, df = 1, p < .001}$). Second, all other fit indices indicate that the fit of these two models is essentially equivalent. Finally, and most importantly, we can see that when the method variance is extracted, the correlation between the two conceptual factors is very high ($r = -.86, p < .001$). Thus, even if one accepted that the second factor has conceptual meaning, these results suggest that the second factor, Preference for Neutrality, is essentially the inverse of the Need to Evaluate. The enhanced relationship between the factors is consistent with the rationale that correlated measurement error would have attenuated the true negative relation between the reverse and non-reverse scored items. With the extraction of this method variance, the two sets of items are highly negatively related, resulting in the two-factor solution offering little improvement in either statistical or conceptual terms.

Internal Consistency

Table 4 presents the means, standard deviations, and item-total correlations for each of the 16 retained items on the NES. The item-total correlations indicate that each of the 16 retained items is reliably correlated with the entire scale (range = .36 to .65). Cronbach’s alpha coefficients for the 16-item scale were calculated separately for each of the samples and overall. These analyses indicated the NES to possess a high degree of internal consistency across samples (alphas = .87, .82, .86 for samples...
Table 3

Goodness of Fit Indexes for Four Alternative Confirmatory Factor Analysis Models for Both Maximum Wishart Likelihood (MWL) and Generalized Least Squares (GLS) Solutions

<table>
<thead>
<tr>
<th>Model</th>
<th>( \chi^2 )</th>
<th>RMSEA</th>
<th>TLI</th>
<th>BBNFL</th>
<th>CFI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null model (df = 120)</td>
<td>2508.76</td>
<td>.19</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>one factor</td>
<td>691.06</td>
<td>.10</td>
<td>.72</td>
<td>.73</td>
<td>.75</td>
</tr>
<tr>
<td>Model B (df = 103)</td>
<td>414.26</td>
<td>.07</td>
<td>.85</td>
<td>.84</td>
<td>.87</td>
</tr>
<tr>
<td>one factor + method</td>
<td>459.52</td>
<td>.08</td>
<td>.83</td>
<td>.82</td>
<td>.85</td>
</tr>
<tr>
<td>Model D (df = 102)</td>
<td>402.63</td>
<td>.07</td>
<td>.85</td>
<td>.84</td>
<td>.87</td>
</tr>
<tr>
<td>two factor + method</td>
<td>420.90</td>
<td>.08</td>
<td>.93</td>
<td>.93</td>
<td>.94</td>
</tr>
</tbody>
</table>

Goodness of Fit Index (GLS)

<table>
<thead>
<tr>
<th>Model</th>
<th>( \chi^2 )</th>
<th>RMSEA</th>
<th>TLI</th>
<th>BBNFL</th>
<th>CFI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null model (df = 120)</td>
<td>5386.39</td>
<td>.28</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>one factor</td>
<td>885.70</td>
<td>.12</td>
<td>.83</td>
<td>.84</td>
<td>.85</td>
</tr>
<tr>
<td>Model B (df = 103)</td>
<td>429.99</td>
<td>.08</td>
<td>.93</td>
<td>.92</td>
<td>.94</td>
</tr>
<tr>
<td>one factor + method</td>
<td>481.61</td>
<td>.08</td>
<td>.92</td>
<td>.91</td>
<td>.93</td>
</tr>
<tr>
<td>Model D (df = 102)</td>
<td>420.90</td>
<td>.08</td>
<td>.93</td>
<td>.93</td>
<td>.94</td>
</tr>
</tbody>
</table>

Note. At the most basic level, interpretation of these fit indexes is as follows. All indicate, in some form, the amount of discrepancy between the actual sample covariance matrix and the covariance matrix predicted by the model's estimated parameters. The Tucker-Lewis Index (TLI), Bentler-Bonett Normed Fit Index (BBNFI), and Comparative Fit Index (CFI) each involve a direct comparison of the target model with the null model, and differ only in terms of adjustments made for the number of parameters involved. All range from 0 to 1, with 1 indicating the target model has accounted for all of the discrepancy indicated in the null model. Conventionally, as each approaches .9, the model can be said to fit the data very well. Readers interested in the subtle differences between these measures are referred to Bentler (1990) for a discussion (see also Tucker & Lewis, 1973). In contrast, the Root Mean Square Error of Approximation (RMSEA) is not dependent on a comparison with the null model and is thus an indicator of absolute rather than relative fit. RMSEA is corrected for the number of parameters and has a lower bound of 0, with .05 indicating perfect fit. As RMSEA descends and approaches .05, the model can be said to fit the data very well (see Browne & Cudek, 1992).

Results

The correlation between first and second administrations of the scale was .84.

Study 2: Relation of the NES to Other Scales

The results of Study 1 offered evidence that the Need to Evaluate Scale is a reliable measure capturing a stable individual difference. The purpose of Study 2 was to examine the relation of the NES to some relevant existing individual difference scales in order to begin assessing its discriminant and convergent validity. In light of the number of correlations to be examined, we have structured this section such that the obtained results and relevant discussion will immediately follow our rationale for measuring each individual difference.

In short, we felt that it was necessary to demonstrate that NES scores were not redundant with measures of participants' motivation to think in general, to gain control over their environment, to obtain structure or closure, to provide socially desirable responses, or by their affective intensity. However, we did expect, based on functional theories of attitudes, that those individuals who valued such functions would be likely to score somewhat higher on the NES than others. For this reason, we expected at least small positive correlations between the NES and participants' motivation to gain control over their environment and to gain a structured understanding of their world. Because evaluation represents one form of thought, we also expected at least a small positive correlation with an assessment of motivation to think in general. Finally, because evaluation is often affective in nature, we expected at least a small positive correlation with affective intensity. On the other hand, we did not expect the NES to correlate with measures of self-monitoring or social desirability.

Method

Participants and Procedure

Participants were 600 (307 male, 293 female) undergraduates enrolled in introductory psychology classes at Ohio State University who participated for partial course credit. Participants were tested in large groups ranging in size from 35 to 55. All participants were administered the NES and a subset of two to four of the following scales: the 18-item Need for Cognition Scale (Cacioppo, Petty, & Kao, 1984), the Desire for Control Scale (Burger & Cooper, 1979), the Internal Versus External Control Scale (Rotter, 1966), the Personal Need for Structure Scale (Neuberg & Newsome, 1993), the Need for Closure Scale (Webster & Kruglanski, 1994), the Self-Monitoring Scale (Snyder, 1974), the Affect Intensity Scale (Larsen, Diener, & Emmons, 1986), the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1964).

Overall, men (M = 52.80) obtained a slightly larger score on the NES than did women (M = 50.88), \( F(1, 598) = 4.52, p < .05 \). Correlations of the NES with the eight scales noted above are presented in Table 5 and are described below.

Results

As noted earlier, 70 participants (36 male, 34 female) from Sample 3 completed the NES at two points in time, 10 weeks apart. The results of the 10-week test–retest analysis indicated participants' scores on the NES to be highly stable over time.
Table 4
Item Means, Standard Deviations, and Item-Total Correlations (n = 557)

<table>
<thead>
<tr>
<th>Item</th>
<th>Item wording</th>
<th>Mean</th>
<th>SD</th>
<th>ITC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I form opinions about everything.</td>
<td>3.69</td>
<td>1.07</td>
<td>.52</td>
</tr>
<tr>
<td>2</td>
<td>I prefer to avoid taking extreme positions. (R)</td>
<td>3.03</td>
<td>1.23</td>
<td>.43</td>
</tr>
<tr>
<td>3</td>
<td>It is very important to me to hold strong opinions.</td>
<td>3.27</td>
<td>1.16</td>
<td>.65</td>
</tr>
<tr>
<td>4</td>
<td>I want to know exactly what is good and bad about everything.</td>
<td>3.38</td>
<td>1.15</td>
<td>.41</td>
</tr>
<tr>
<td>5</td>
<td>I often prefer to remain neutral about complex issues. (R)</td>
<td>3.44</td>
<td>1.15</td>
<td>.49</td>
</tr>
<tr>
<td>6</td>
<td>If something does not affect me, I do not usually determine if it is good or bad. (R)</td>
<td>3.20</td>
<td>1.14</td>
<td>.38</td>
</tr>
<tr>
<td>7</td>
<td>I enjoy strongly liking and disliking new things.</td>
<td>2.91</td>
<td>1.07</td>
<td>.46</td>
</tr>
<tr>
<td>8</td>
<td>There are many things for which I do not have a preference. (R)</td>
<td>2.81</td>
<td>1.10</td>
<td>.40</td>
</tr>
<tr>
<td>9</td>
<td>It bothers me to remain neutral.</td>
<td>2.95</td>
<td>1.20</td>
<td>.50</td>
</tr>
<tr>
<td>10</td>
<td>I like to have strong opinions even when I am not personally involved.</td>
<td>3.01</td>
<td>1.16</td>
<td>.64</td>
</tr>
<tr>
<td>11</td>
<td>I have many more opinions than the average person.</td>
<td>3.16</td>
<td>1.04</td>
<td>.56</td>
</tr>
<tr>
<td>12</td>
<td>I would rather have a strong opinion than no opinion at all.</td>
<td>3.57</td>
<td>1.21</td>
<td>.48</td>
</tr>
<tr>
<td>13</td>
<td>I pay a lot of attention to whether things are good or bad.</td>
<td>3.55</td>
<td>1.03</td>
<td>.36</td>
</tr>
<tr>
<td>14</td>
<td>I only form strong opinions when I have to. (R)</td>
<td>3.25</td>
<td>1.21</td>
<td>.53</td>
</tr>
<tr>
<td>15</td>
<td>I like to decide that new things are really good or really bad.</td>
<td>3.42</td>
<td>1.02</td>
<td>.36</td>
</tr>
<tr>
<td>16</td>
<td>I am pretty much indifferent to many important issues. (R)</td>
<td>3.63</td>
<td>1.07</td>
<td>.36</td>
</tr>
<tr>
<td>Total Scale</td>
<td></td>
<td>3.27</td>
<td>.63</td>
<td>.85</td>
</tr>
</tbody>
</table>

Note. (R) Indicates item was reverse scored. ITC = item-total correlation.

Table 5
Correlations Between NES and Other Scales

<table>
<thead>
<tr>
<th>Scale</th>
<th>Correlation with NES</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affective Intensity</td>
<td>.17*</td>
<td>109</td>
</tr>
<tr>
<td>Desire for Control</td>
<td>.22*</td>
<td>103</td>
</tr>
<tr>
<td>External Locus of Control</td>
<td>-.05</td>
<td>136</td>
</tr>
<tr>
<td>Need for Cognition</td>
<td>.35**</td>
<td>88</td>
</tr>
<tr>
<td>Need for Closure</td>
<td>-.06</td>
<td>87</td>
</tr>
<tr>
<td>Personal Need for Structure</td>
<td>.03</td>
<td>95</td>
</tr>
<tr>
<td>Desire for Structure</td>
<td>.06</td>
<td>95</td>
</tr>
<tr>
<td>Response to Lack of Structure</td>
<td>-.02</td>
<td>95</td>
</tr>
<tr>
<td>Self-Monitoring</td>
<td>-.04</td>
<td>103</td>
</tr>
<tr>
<td>Social Desirability (Marlowe-Crowne)</td>
<td>-.05</td>
<td>83</td>
</tr>
</tbody>
</table>

Note. NES = Need to Evaluate Scale.  
*p < .05.  **p < .001.

thought (Cacioppo & Petty, 1982; Cacioppo, Petty, Feinstein, & Jarvis, in press). If individual differences in the need to evaluate are determined primarily by differences in people's motivation to think, then a high correlation should be obtained between need for cognition and the NES. Alternatively, if the motivation to think represents only one of many factors that could influence one's motivation to evaluate, or the motivation to evaluate represents only one of many reasons for why people engage in thinking, then a smaller correlation would be expected. The relation between the NES and the Need for Cognition Scale was moderate and positive (r = .35, p < .001). The modest correlation between need for cognition and the NES fits well with current perspectives in attitude theory which hold that evaluation by no means requires effortful thought (e.g., see Chaiken et al., 1989; Petty & Cacioppo, 1981, 1986b).

Motivation to Control

As discussed earlier, many theorists have argued that one function that global evaluations (i.e., attitudes) can serve is that they help the individual to gain control over his or her environment. Presumably, knowing what is good and bad allows people to maximize positive outcomes and minimize negative outcomes because those objects associated with reward will be approached and those associated with punishment avoided (e.g., Festinger, 1954; Katz, 1960; Petty & Cacioppo, 1986a; Smith et al., 1956). Burger and Cooper (1979) developed an individual difference measure of chronic levels of control motivation. The relation between the NES and control motivation was significant, but relatively small (r = .22, p < .05). This suggests that chronic control motivation may in fact be only one of many motives associated with a stable motive to evaluate. Interestingly, the NES was independent of Rotter’s (1966) Internal Versus External Control Scale (r = -.05, p < .05), suggesting that although the desire for control might relate to the need to evaluate, the perception that one is actually in control of his or her environment appears unrelated.

Motivation to Attain Structure and Knowledge

Katz (1960; Katz & Stotland, 1959) proposed that one important function that attitudes serve is that of helping the person to structure, understand, and generally make sense of the world. Neuberg and Newsome (1993) have recently provided validation for the Personal Need for Structure (PNS) Scale developed
by Thompson, Naccarato, and Parker (1989). This scale measures individual differences in the chronic motive to simplify and structure the experiential world. Although we expected at least a small positive correlation with the NES, these measures were independent (r = .03, ns).8

The need for closure has been defined by Kruglanski (1989, p. 14) as a “desire for a definite answer on some topic, any answer as opposed to confusion and ambiguity.” The Need for Closure Scale (Webster & Kruglanski, 1994) was developed in order to measure chronic individual differences in such a motive. Because an evaluation of any particular object offers at least one answer to the individual who is motivated by closure, a moderate correlation with the NES might be expected. However, one could also argue that some people high in the need for closure could be averse to evaluative responding given the subjective nature of evaluation and the individuals’ need for definite answers. In any case, the need to evaluate could not be predicted by participants’ scores on the Need for Closure Scale (r = -.06, ns).

Self-Monitoring

The self-monitoring construct (Snyder, 1974) reflects an individual’s tendency to be chronically concerned with self-image. High self-monitors are motivated to make and maintain a positive image in social settings, whereas low self-monitors are more concerned with expressing their true inner selves. If attitudes facilitate the development and maintenance of social interactions, then it might be expected that high self-monitors could be averse to evaluative responding given the subjective nature of evaluation and the individuals’ need for definite answers. High self-monitors are motivated to make and maintain a positive image in social settings, whereas low self-monitors are more concerned with expressing their true inner selves. Thus, one might also expect that those low in self-monitoring could be highly motivated to engage in evaluative responding. It could also be that both high and low self-monitors are equally likely to engage in evaluative responding—but for different reasons. If so, then no overall correlation with NES would be obtained. In fact, the need to evaluate could not be predicted from an individual’s score on the self-monitoring scale (r = -.04, ns).

Affect Intensity

The Affect Intensity Scale (Larsen et al., 1986) measures individual differences in the intensity with which people experience their emotions across a wide number of situations. People who score high on this scale experience emotion more strongly than those who are low (Larsen & Diener, 1987). Given that attitudes and opinions are often affective in nature, one might expect a positive correlation between the NES and affective intensity. However, one could also argue that this correlation would be limited by the fact that attitudes can also be based on cognition rather than affect (Crites, Fabrigar, & Petty, 1994; Zanna & Rempel, 1988). Consistent with the notion that attitudes have multiple bases, the correlation between the NES and the Affective Intensity Scale was positive but small (r = .17, p < .05).

Social Desirability

The Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1964) measures an individual’s motive to avoid social disapproval (Crowne, 1979). If participants who score high (or low) on the NES are doing so simply to avoid making a negative impression, then we would expect to see a positive (negative) correlation with the Marlowe-Crowne. The correlation between the Marlowe-Crowne and the NES was nonsignificant (r = -.05, ns), suggesting this is not the case.9

Summary

No measure that we compared with the NES could account for more than 16% of the variance in participants’ scores on the NES, suggesting that the NES is tapping a unique individual difference. The fact that no correlation exceeded .35 suggests good discriminant validity, especially when one considers that a number of these established scales have been shown to be more highly correlated with each other than that (e.g., desire for control with need for cognition, r = .48, Thompson, Chaiken, & Hazelwood, 1993). It was, however, proposed that an individual’s need to evaluate could be multiply determined, and that although no other single individual difference could fully predict the need to evaluate, a number of factors could be expected to contribute. This view was supported by low positive correlations between the NES and need for cognition, desire for control, and affect intensity.

Study 3: Existence of Attitudes Toward Social and Political Issues

So far, we have demonstrated that the NES is a stable unidimensional construct that is not redundant with a number of other relevant individual differences. In Study 3, we attempted to demonstrate the predictive validity of the NES. We reasoned that one implication of individual differences in evaluative responding is that those who score high on the NES should be more likely than those who score low to form global evaluations (i.e., attitudes) toward the objects they encounter. To test this implication, we adapted items from the 1990 National Election Survey and asked college students for their attitudes toward social and political issues. As is often standard procedure in the National Election Survey, we provided participants with an explicit option to indicate that they had no opinion on each issue. If those high in the need to evaluate engage in evaluative

---

8 Because the sample of participants who filled out the PNS also happened to fill out the Need for Cognition Scale, this allowed an exploratory analysis of the relation between the PNS and NES separately for individuals who were differentially motivated to engage in effortful thought. Interestingly, the relation between the need for structure and the need to evaluate was reliable—but only for those who were low in need for cognition (r = -.44, p < .001; high need for cognition: r = -.17, ns). An additional 2 (Need for Cognition: low, high) X 2 (Need for Structure: low, high) analysis of variance based on median splits with the need to evaluate as the dependent variable revealed this interaction to be reliable, F(1, 91) = 4.27, p < .05. These analyses suggest that the need for structure may be a partial determinant of the need to evaluate, but primarily for those who do not enjoy thinking for its own sake.

9 In a more recent study, we administered the NES and Paulhus’ (1988) Balanced Inventory of Desirable Responding to sample of 116 participants (50 female, 66 male). Further support for a null relation between the NES and socially desirable responding was obtained (r = .12, ns).
responding more often than those who are low, then this tendency should result in an increased likelihood of attitude formation. Thus, we predicted that those scoring high in the need to evaluate would report having more attitudes than those scoring low across the divergent set of social and political issues.

For control purposes, participants also were measured on their need for cognition and need for closure. We reasoned that the tendency to engage in effortful thought and to obtain closure could also be related to the tendency to form and maintain attitudes. Thus, the utility of the NES in predicting the likelihood of attitudinal responding was compared with these variables using stepwise and hierarchical regression analyses, as well as through analyses of variance splitting participants into high, moderate, and low levels of the need to evaluate.

Method

Participants

Participants were 52 (27 male, 25 female) students enrolled in introductory marketing classes at Ohio State University. All participants received course credit. The experimental sessions were conducted in groups ranging in size from one to five, but each participant completed the study in a private cubicle.

Procedure

When the students arrived, they were instructed that they would be participating in four different experiments within the hour-long session. The first experiment was simply the administration of the NES, Need for Cognition, and Need for Closure Scales. Participants were asked to complete these scales as part of a series of scale validation experiments being conducted by the psychology department. Following the administration of the individual difference measures, two short studies unrelated to the experiment reported here were conducted. This was done to minimize the salience of participants' responses to the individual difference scales when completing the questionnaire described below. All participants were asked to complete a final questionnaire entitled "Political Survey—National Issues." The cover page of the questionnaire informed participants that the questionnaire was a cooperative effort between the psychology and political science departments and that its purpose was to measure attitudes and opinions toward a number of current social and political issues. The need for accuracy in responding was emphasized. After completing this questionnaire, all participants were fully debriefed and thanked.

Political Questionnaire

We constructed a 29-item questionnaire that asked participants for their attitudes toward various national issues such as environmental protection, mandatory national service, legalized abortion, capital punishment, preferential hiring of minority groups, and so forth. Items were adapted from the 1990 American National Election Survey conducted by the Center for Political Studies at the University of Michigan. Only items that were still current issues and that required minimal specific information in order to give an opinion (e.g., items not making reference to specific people or events) were included. Response option formats for each item also were consistent with the original survey where possible. One of the response options for each question was "no opinion."

Results and Discussion

Participants' scores on the NES ranged from 34 to 76, with a mean of 53.0 and a median of 51. Men \( M = 53.76 \) and women \( M = 51.8 \) did not differ in their average scores, \( F(1, 50) < 1 \). According to a tertiary split, those scoring in the upper tertile of the NES (scores ≥ 57 ) were classified as high, the middle tertile as moderate, and the lower tertile (scores ≤ 46 ) as low in the need to evaluate. In each of our Studies 3, 4, and 5, our planned comparisons were directed at comparing participants classified as high or low, however, all omnibus tests of interactions and main effects reported are based on data from all three levels of the need to evaluate. Furthermore, all planned comparisons reported were calculated using error terms from all participants because this typically provides the most reliable estimate (Kirk, 1982).

Analysis of Variance

To examine the relative prevalence of social and political attitudes as a function of the need to evaluate, we performed a one-way analysis of variance on the number of "no opinion" responses given by participants in response to the 29 attitude questions as a function of the three levels of the need to evaluate. Based on the expectation that participants high in the need to evaluate would be most likely to have attitudes toward these issues, we simply predicted a main effect such that participants high in the need to evaluate would less frequently endorse the no opinion response than their low need to evaluate counterparts. The predicted negative relation between the need to evaluate and the number of no-opinion responses offered by participants was obtained, \( F(2, 49) = 3.52, p < .04 \). Participants who were high in the need to evaluate endorsed the no-opinion response option less frequently \( (M = 4.24) \) than those who were low \((M = 4.42)\), \( t(33) = 2.34, p < .02 \), or moderate \( (M = 4.11) \), \( t(33) = 2.307, p < .02 \). Participants low and moderate in the need to evaluate did not differ reliably \( (t < 1) \).

Regression Analyses

We also performed stepwise and hierarchical multiple regression analyses analyzing the number of no-opinion responses as a function of participants' gender and scores on the NES, and Need for Cognition and Need for Closure Scales. The stepwise regression analysis revealed participants' scores on the NES to be the strongest predictor of no-opinion responses, and in the expected negative direction \( (\beta = - .34) \), \( t(50) = -.248, p < .02 \). Participants low in the need to evaluate entered the regression equation first, with no other variables achieving significance in the equation.

A hierarchical regression analysis also was performed in which participants' gender and scores on the Need for Cognition and Need for Closure Scales were initially forced into the regression equation. Consistent with the stepwise analysis, subsequent entry of need to evaluate resulted in a significant increment in variance accounted for \( (\beta = -.32) \), \( t(48) = -2.13, p < .04 \). No other variables were reliable in the equation. Finally, the interaction product terms were entered hierarchically according to the recommended procedure for testing interactions by Cohen and Cohen (1983). None of the control variables reliably interacted with NES or with each other to predict evaluative responding.

Summary

The results of Study 3 supported the predictive validity of the NES. We proposed that the unitary construct being measured.
by the NES was a chronic tendency to engage in evaluative responding and supported this by demonstrating that high NES scores were associated with an increased frequency of global attitude formation.

**Study 4: Spontaneous Evaluative Thought**

In Study 4 we attempted to further demonstrate the predictive validity of the NES as well as explore an issue raised by McClelland and his colleagues about whether a postulated motivational trait reflects a self-attributed versus an implicit motive (McClelland et al., 1989; Weinberger & McClelland, 1990). Self-attributed motives are viewed as components of the self-concept. They are the motives that an individual believes he or she possesses. According to McClelland, self-attributed motives impact on behavior when situational variables make the relevant component of the self-salient, but they otherwise lay dormant and are uninfluential in determining behavior. Because the NES requires respondents to be aware of their propensity to evaluate, if it accounts for anything, it should predict evaluative responding in situations that cue participants as to the evaluative nature of the task. Study 3 used such an explicitly evaluative task—responding to an opinion survey—and found that the NES was predictive of responses. That is, those who scored high on the NES were more likely to express opinions than those who scored low on the NES. This study suggests that the NES can be valuable in predicting behavior when explicitly evaluative tasks are used. McClelland argued, however, that self-attributed traits rarely predict behavior when the task contains no motivationally relevant cues. For example, Weinberger and McClelland (1990) noted that in one study (Patten & White, 1977) a self-report measure of achievement motivation predicted performance on a laboratory task only when there was some external incentive for achievement. When no such incentive was present, self-attributed need for achievement did not relate to performance.

In contrast to self-attributed motives which McClelland argued are unique to humans and are acquired after language development, implicit motives are argued to be either biologically based or learned in the prelinguistic period. Furthermore, in contrast to self-attributed motives that predict behavior in structured situations where the motive is salient, implicit motives are thought to predict behavior in unstructured situations where the motive is not explicitly cued (Weinberger & McClelland, 1990). Although it is clear that the NES predicts evaluative responding in a structured situation in which evaluation is explicitly cued, can it also predict behavior in the unstructured and spontaneous situations where implicit motives are thought to operate? There is reason to believe that the need to evaluate can be an implicit as well as a self-attributed motive. That is, all organisms have built-in biological mechanisms for approaching and avoiding stimuli in the environment. For lower organisms, these mechanisms are relatively fixed, but in humans these mechanisms include the development of stored evaluations (attitudes) that guide approach and avoidance tendencies (see Cacioppo, Petty, & Berntson, 1991). Thus, the need to evaluate, like other implicit motives identified by McClelland (e.g., need for achievement), plausibly has a biological as well as an environmental basis (though a biological basis may not be necessary to achieve implicit status). If the need to evaluate is an implicit as well as a self-attributed motive, then it should be manifest in spontaneous as well as structured situations. In Study 4 we attempted to assess participants’ tendencies to engage in evaluative responding when no situational cues directed participants’ attention toward doing so. Any such tendencies would be expected to reflect participants’ implicit motives to evaluate—that is, their tendency to engage in a behavior when the relevant aspect of the self-concept has not been directly cued by the task (McClelland et al., 1989). We reasoned that if the NES can reliably predict individual differences in participants’ implicit motivation to engage in evaluative responding, then the NES would be tapping more than an aspect of participants’ self-concepts and could prove useful in accounting for evaluative behavior in both structured and unstructured situations.

**Overview**

To examine the ability of the NES to predict participants’ implicit motivation to evaluate, participants in Study 4 were presented with 24 paintings from various styles and periods and were instructed simply to list their thoughts regarding each painting. Behavior in such non-demanding, free-response situations have been argued to be ideal for measuring implicit motivation (e.g., Koestner & McClelland, 1990). Thoughts were subsequently coded as evaluative or nonevaluative. If the NES were strictly measuring a self-attributed motive, then participants’ NES scores should be unrelated to the prevalence of evaluative thought in this situation that did not explicitly prime evaluation.

For control purposes, participants were again measured on two potentially related constructs: need for cognition, and need for closure. We also obtained measures of participants’ general interest in, knowledge of, and direct experience with art, as well as their interest in the specific paintings used for this experiment. The predictive utility of the NES was compared with these variables using stepwise and hierarchical regression analyses, as well as through multivariate analyses of variance splitting participants into high, moderate, and low scores on the NES.

**Method**

**Participants**

Participants were 35 female students enrolled in introductory psychology classes at Ohio State University. All participants received course credit for their participation. All participants completed the study individually.

That is, it also seems possible that implicit motives could be learned but are not present in consciousness. If so, the implicit/self-attributed distinction for motives becomes similar to a distinction in the attitudes and social cognition literatures between implicit and explicit attitudes (see Greenwald & Banaji, 1995). For example, an individual might view him or herself as unprejudiced (self-attributed or explicit attitude), but yet act in a prejudiced manner (implicit attitude). The self-attributed unprejudiced attitude would operate mostly in situations that were seen as explicitly relevant to the attitude, whereas the implicit prejudiced attitude would operate mostly in more spontaneous situations.
Procedure

Instructions. Upon arrival, participants were introduced to the experimenter who told them that they would be helping in selecting stimulus materials for a future study. Specifically, their task was to write down their initial responses to various paintings that they would see. Participants were provided with a 5-page booklet that clearly marked the space to be used for indicating each thought. Space was provided for up to three thoughts for each of 24 paintings. The experimenter noted that there was room to write down three thoughts for each painting and that although they could use short sentences to express their thoughts, they should avoid writing just one or two words.

Stimuli. Slides of 33 paintings were obtained with the intention of presenting a widely varied selection of stimuli. These paintings, chosen from numerous styles and periods, were approximately equal in terms of the number of positive, negative, and neutral images, as well as the number of simple and complex images. Consultation with the curator of the Ohio State University art library aided in this selection process. Whether paintings were classified as positive, negative, or neutral was determined on the basis of subject matter (e.g., people playing, people fighting, portraits), imagery (e.g., suggesting serenity, death, thoughtfulness), use of color (e.g., soft blues, harsh reds, neutral browns), and general aesthetic appeal. For example, one of the positive paintings was Monet's Rouen Cathedral at Sunset; one of the neutral paintings was Copley's Paul Revere, and one of the negative paintings was Bacon's Head Surrounded by Slides of Beef. Slides were projected onto a wall at a distance of approximately 7 feet from the participant, with image dimensions of approximately $3' \times 2'$.

Free thought listing task. After ensuring that the participants understood all instructions, the thought listing task was begun. Each of the 24 selected slides was presented for a duration of 35 s. Twenty-five seconds into each trial, the experimenter gave warning of the next trial (e.g., by announcing "Painting #2"). This gave the participant 10 s to finish writing her thoughts about the current painting.

Coding evaluative thoughts. Responses were scored as evaluative if they either implied an evaluation (e.g., "I would not hang this in my home."); "Pretty colors"); or were explicitly evaluative (e.g., "I really like the colors," "Very gross and disturbing"). Any response that was seemingly void of implicit or explicit evaluation was scored as nonevaluative (e.g., "Two men working," "Lots of color"). Because in this study participants often did not use complete sentences, a number of thoughts were ambiguous with respect to evaluative content. For ambiguous thoughts, raters were instructed that if it seemed at all likely that the thought could have occurred without its causing the person to feel favorable or unfavorable toward the painting or the image (e.g., in cases such as "lots of violence going on"); then the thought was to be scored as nonevaluative.

For each of the 24 paintings, participants received an evaluative content score for their responses that could range from 0 to 3, because up to three thoughts could be listed for each painting. Similarly, for each painting, the participant received a nonevaluative thought score that also ranged from 0 to 3. The interrater reliability of two independent raters of evaluative content was .92, with discrepancies being resolved through discussion. The scores for all 24 paintings were subsequently averaged to obtain an overall evaluative and an overall nonevaluative thought score for each participant, with each ranging from 0 to 3.

Post-test attitude measurement. After completion of the thought-listing task, participants were told that we also were interested in "how much people will tend to like or dislike these paintings." The experimenter told participants that she would measure how much they liked each painting on a scale from 1 to 10. Measuring attitudes toward the various paintings served as a check for the valence manipulation discussed earlier. Each trial in this task simply required the participant to look at the slide and then announce how much she liked it on a scale from 1 to 10 (with 1 = do not like it at all; 10 = like it very much). The experimenter recorded all responses.

Intervening task and individual difference measures. In order to reduce the salience of the thought-listing task prior to participants responding to the individual difference questionnaires, participants completed a lengthy and demanding memory experiment that was unrelated to the current study. Upon completion of the unrelated memory task, which lasted approximately 25 min, participants were given a booklet containing the Need for Cognition, the Need for Closure, and the Need to Evaluate Scales. At the end of the booklet were four questions regarding the participant's general interest in, knowledge of, and direct experience with art, as well as interest in the paintings used in the experiment. Finally, participants were fully debriefed and thanked for their participation.

Results and Discussion

Participants' scores on the NES ranged from 36 to 74, with a mean of 53.6 and a median of 53. Participants were again classified as high, moderate, or low in need to evaluate according to a tertile split of their scores. Those scoring in the upper tertile (scores ≥ 59) were classified as high, the middle tertile as moderate, and the lower tertile (scores ≤ 48) as low in need to evaluate.

Analyses

Analysis of variance. To examine the relative prevalence of evaluative versus nonevaluative thought as a function of NES score, we performed a 2 (Type of thought: evaluative, nonevaluative) × 3 (Need to evaluate: low, moderate, high) analysis of variance on the average number of thoughts listed per painting.11 We predicted a two-way interaction such that participants high in the need to evaluate would generate more evaluative thoughts toward the paintings, but equal or fewer nonevaluative thoughts than participants low in the need to evaluate.

Figure 3 presents the mean evaluative and nonevaluative thought scores as a function of scores on the NES. The NES did not reliably influence the overall number of thoughts per stimulus listed by participants, $F(2, 32) = 1.25$, $ns$. (Ms = 2.29, 2.49, 2.58, for low, moderate, and high NES participants respectively). Second, the predicted two-way interaction between NES and Thought Type was obtained, $F(2, 32) = 5.62, p < .01$. Planned comparisons between the obtained means revealed that high-NES participants listed more evaluative thoughts per stimulus ($M = 1.36$) than low-NES participants ($M = .61$), $F(1, 32) = 30.23, p < .001$. High-NES participants also reported fewer nonevaluative thoughts per stimulus ($M = 1.23$) than low-NES participants ($M = 1.68$), $F(1, 32) = 11.26, p < .01$. Thus, the positive relation between the NES and evaluative thought was not a consequence of a global increase in all types of thought.

Regression analyses. We also performed stepwise and hierarchical multiple regression analyses analyzing the mean number of evaluative thoughts per painting as a function of need to evaluate, need for cognition, need for closure, and the four separate art variables (i.e., knowledge of art, appreciation and

11 In studies 4 and 5, we also analyzed evaluative thoughts as a proportion of total thoughts. Results were similar to those presented in all cases.
enjoyment of art, direct experience with art, and interest in the paintings chosen for this experiment). The results of this analysis showed participants’ scores on the NES to be the strongest predictor of evaluative thought, which was in the expected positive direction ($\beta = .52$), $t(33) = 3.40, p < .002$. Need to evaluate was entered into the regression equation first, with no other variables achieving significance in the equation. The same analysis using nonevaluative thoughts as the dependent variable revealed no reliable predictors.

A second regression analysis also was performed, in which all variables except need to evaluate were initially forced into the regression equation. Consistent with the stepwise analysis, subsequent entry of need to evaluate resulted in a significant increment in variance accounted for ($\beta = .46$, $t(27) = 2.48, p < .02$). Interestingly, participants’ self-reported knowledge about art was also a marginally reliable predictor ($\beta = .39$), $t(27) = 2.05, p < .06$. Finally, the interaction product terms were entered between NES and each of the control variables. No interactions were reliable.

Stimulus valence. In the following analysis, whether the obtained effects were being driven by a particular class of stimuli was examined. Specifically, whether the increase in evaluative responding for participants high in the need to evaluate was limited to positive, neutral, or negative stimuli was tested.

Recall that the valence of imagery in the 24 paintings was manipulated such that positive, neutral, and negative images were equally represented. To check the success of this manipulation, we averaged each participant’s post-test attitudinal responses to the eight paintings in each valence set. A 3 (Stimulus Valence: positive, neutral, negative) × 3 (Need to Evaluate: low, moderate, high) analysis of variance was performed on these data. This analysis revealed that the valence manipulation was successful, $F(2, 64) = 34.58, p < .001$. Comparing post-test attitudes toward each group of paintings revealed that the positive paintings ($M = 6.77$) were rated more favorably than the neutral paintings ($M = 5.37$), $F(1, 64) = 19.39, p < .001$, which were rated more favorably than the negative paintings ($M = 4.06$), $F(1, 64) = 16.71, p < .001$. Furthermore, overall favorability toward the paintings did not vary as a function of the NES, $F(2, 32) = .41, n.s.$, nor did the NES interact with stimulus valence to affect attitude judgments, $F(4, 64) = .81, n.s.$ Thus, participants at the three levels of the need to evaluate were in agreement with respect to the favorability of the paintings used in this experiment. That is, even though they engaged in less evaluative thinking, when explicitly asked to express an attitude, those low and moderate in the need to evaluate reported the same attitudes as people high in the need to evaluate.

To test the possibility that the thought listing results were dependent on the valence of the stimuli, a 3 (Stimulus Valence: positive, neutral, negative) × 3 (Need to Evaluate: low, moderate, high) analysis of variance was performed on the dependent variable, evaluative thought. Again, the main effect for the need to evaluate was reliable, $F(2, 32) = 8.07, p < .01$. This analysis also revealed a main effect for stimulus valence such that paintings with positive imagery elicited greater evaluative thought ($M = 1.14$) than either neutral ($M = .93$) or negative ($M = .96$) paintings, $F(2, 64) = 7.57, p < .001$. Finally, the influence of the need to evaluate on evaluative thought did not reliably interact with stimulus valence, $F(4, 64) = 1.44, n.s.$ That is, evaluative thought increased as a function of the need to evaluate, and this relationship was stable across positive, neutral, and negative stimuli.

Summary

Study 4 represents a conceptual replication and extension of Study 3. Supporting the predictive validity of the NES, it was again found that participants high in the need to evaluate were more likely to make evaluative responses. However, this time it also was demonstrated that the ability of the NES to predict evaluative responding did not require a situation in which participants were cued to the evaluative nature of the task. This finding suggests that the nature of the need to evaluate is not strictly limited to an aspect of the self-concept and is not dependent on participants’ explicit motivation to engage in evaluative responding. It was also shown that the positive relation between the need to evaluate and evaluative responding did not depend on the valence of the stimuli used. This suggests that the need to evaluate does not arise strictly from a heightened preference for positivity or a reduced aversion to negativity. Finally, it was again demonstrated that the prediction of evaluative responding was unique to the NES among other stable individual differences.

Study 5: Autobiographical Narratives

Although Study 4 demonstrated that the NES could predict evaluative responding in a context devoid of explicit cues suggesting the evaluative nature of the task, one could argue that the nature of the stimuli (i.e., paintings) might have subtly cued participants that the task was an evaluative one. In fact, recall that two thirds of the paintings had been intentionally chosen to represent positive or negative stimuli. Thus, it is possible that the evaluative nature of many of the paintings could have cued participants’ self-attributed motives to evaluate much in the
same way as any explicit cue would have. For this reason, in Study 5, participants were again asked to freely list their thoughts, but this time in a task that was not at all explicitly evaluative. In this study, participants were asked simply to provide short autobiographical narratives describing the events of their previous day. It was hypothesized that participants' NES scores would be positively related to the prevalence of evaluative relative to nonevaluative thoughts in their descriptions of the previous day.

Results and Discussion

In Study 5, participants' scores on the NES ranged from 34 to 80 out of a possible 80, with a mean of 52 and median also of 52. Men (M = 52.96) and women (M = 51.04) did not differ in their average scores, F < 1. Participants again were classified as high, moderate or low in the need to evaluate according to a tertary split of participants' scores on the NES. Those scoring in the upper tertile (scores ≥ 56) were classified as high, the middle tertile as moderate, and the lower tertile (scores ≤ 47) as low.

Analysis of Variance

To examine the relative prevalence of evaluative versus nonevaluative thought segments in participants' narratives as a function of the need to evaluate, a 2 (Type of thought Segment: evaluative, nonevaluative) x 3 (Need to evaluate: low, moderate, high) mixed analysis of variance was performed on the number of thought segments listed by each participant. A two-way interaction was predicted such that participants high in the need to evaluate would generate more evaluative thought segments, but equal or fewer nonevaluative thought segments than participants low in the need to evaluate. Obtaining this interaction was important to show that the increase in evaluative thought as a function of the need to evaluate was not simply due to an increase in all types of thought.

Figure 4 presents the mean evaluative and nonevaluative thought segment scores as a function of scores on the NES. First, the NES score did not reliably influence the overall number of thought segments listed by participants, F(2, 90) = 1.44, n.s. (Ms = 12.41, 10.65, 11.47, for low, moderate, and high NES scores respectively). Second, the predicted two-way interaction between the need to evaluate and thought segment type was obtained, F(2, 90) = 4.49, p < .02. Planned comparisons between the obtained means revealed that high-NES participants wrote nearly twice as many evaluative thought segments (M = 2.50) as low-NES participants (M = 1.28), F(1, 90) = 5.92, p < .05. Furthermore, participants high in the need to evaluate wrote fewer nonevaluative thought segments (M = 8.97) than those who were low (M = 11.13), F(1, 90) = 4.91, p < .05. Thus, it appears that the evaluative thought segment scores obtained by high-NES participants again were not obtained due to a global increase in all types of thought.

Regression Analyses

Both stepwise and hierarchical regression analyses were performed analyzing participants' evaluative thought segment scores as a function of gender (scored dichotomously as 0 or 1) and scores on the NES and the Need for Cognition and the Personal Need for Structure Scales. The results of this analysis showed participants' scores on the NES to be the strongest predictor of evaluative thought segments (β = .30, t[93] = 2.99, p < .01). No other variables reliably predicted evaluative responding. The same analysis using nonevaluative thought segments revealed no reliable predictors.

12 We thank Wendi Gardner for suggesting this task.
THE NEED TO EVALUATE

Figure 4. Nonevaluative and evaluative thought segments in autobiographical narratives as a function of individuals' scores on the Need to Evaluate Scale.

A hierarchical regression analysis also was performed on participants' evaluative thought segment scores where all of the control variables were first forced into the equation. Subsequently entering participants' scores on the NES tested whether the need to evaluate had a reliable and unique influence on evaluative responding when all other variables were controlled. The results of this analysis revealed that the need to evaluate exerted a reliable, unique influence on participants' evaluative responses ($\beta = .28, t(88) = 2.48, p < .02$). None of the control variables exerted a reliable influence on the dependent variable. Finally, the interaction product terms were entered between the need to evaluate and each of the control variables. None of the control variables reliably interacted with the NES or with each other to predict evaluative responding.

Summary

Study 5 represents a third successful attempt to demonstrate the predictive validity of the NES. In this study, participants simply were asked to describe the events of their previous day. When these narratives were coded for their evaluative and nonevaluative content, the NES was capable of predicting the prevalence of evaluative thought. Planned comparisons indicated that participants high in the need to evaluate showed greater evaluative thought and less nonevaluative thought than those low in the need to evaluate (see Figure 4). Again, prediction of evaluative thought in this study was unique to the NES, as neither the Need for Cognition Scale nor the Personal Need for Structure Scale accounted for differences in evaluative responding. Although we can only consider Studies 4 and 5 to be preliminary tests of the notion that the NES can assess an implicit motive that operates in spontaneous and unstructured situations, it does appear that the NES is not measuring a strictly self-attributed motive.

General Discussion

The results of the current series of studies offer strong initial support for the view that some people are more prone to engage in evaluative responding than others, and that such an individual difference is measurable. Study 1 demonstrated that the 16-item Need to Evaluate Scale was a reliable measure of a stable individual difference. Supporting the discriminant validity of the NES, Study 2 demonstrated that no more than 13% of the variance in the NES could be accounted for by any of eight previously established and potentially relevant individual difference measures. Supporting the convergent validity of the NES, Study 2 also demonstrated reliable correlations that were consistent with predictions based on functional theories of attitudes. Finally, Study 3 tested the predictive validity of the NES and supported the implication that participants high in the need to evaluate should be more likely than those who are low to have attitudes toward issues they have previously encountered. Studies 4 and 5 further demonstrated the predictive validity of the NES by showing that the extent of participants' evaluative responding could be predicted by their scores on the NES in contexts devoid of explicit cues to engage in evaluation. Additionally, the prediction of evaluative responding was unique to the NES among other individual differences, and reliable prediction was found for both personally relevant stimuli (i.e., participants' own lives), and relatively novel stimuli (i.e., positive, negative, and neutral paintings from various styles and periods). Although there was a small tendency across studies for men to score higher on the NES than women, this effect was reliable only in the largest data set (Study 2). In addition, gender did not influence any of the results we obtained.

Potential Criticisms and Limitations

One potential criticism of the NES is that it, like many other individual difference scales, relies on self-reports. As McClelland et al. (1989) argued, self-attributed traits are often ineffective at predicting behavior unless that aspect of the self-concept is explicitly cued. However, Studies 4 and 5 demonstrated that the NES was capable of predicting evaluative responding when there was nothing explicit in the context to suggest an opportunity to engage in evaluation. Because free thought listings in such situations have been argued to be the
best indicators of implicit motives (i.e., motives independent of the self-concept), it seems reasonable to conclude that the nature of the need to evaluate as measured by the NES is not strictly a self-perception.13

In future research, this question could be examined further by having the same individuals perform several tasks where explicit evaluative motives are salient (e.g., responding to an attitude survey) and several tasks where they are not (e.g., describing one’s previous day). Using covariance structural modelling techniques one could extract the variance common to these multiple tasks and correlate this common variance with participants’ scores on the NES. The variance in common among the nonexplicit evaluative tasks should provide a good indicator of participants’ implicit need to evaluate. One could then determine the extent to which participants’ implicit motivation can be accounted for by their self-attributed need to evaluate (i.e., via the NES). It is possible that the NES will prove to be an adequate measure of behavior on both explicitly and nonexplicitly evaluative tasks. On the other hand, it is possible that there are stable individual differences in the implicit need to evaluate of which participants are unaware and that are not captured by the NES. The need to consider participants’ implicit motives to evaluate is apparent when we consider research supporting the view that many evaluative responses occur in the absence of awareness (e.g., Bargh, 1994; Bargh et al., 1992; Berntson, Boysen, & Cacioppo, 1992; Greenwald & Banaji, 1995). If such nonconscious evaluative processes are variable in terms of frequency or extent, it is not clear whether a self-report measure would correlate at all with such variability. It would be interesting in future research to examine the extent to which the NES can predict nonconscious evaluation since the current research was not designed to address this question.

A potential criticism of our measurement of the need to evaluate might be directed at the frequent use of the term “opinion” rather than “attitude” or “evaluation” in items comprising the NES. One might argue, for example, that an opinion need not be evaluative, and so the scale might be measuring more than just a tendency to engage in evaluative responding. Three responses are offered. First, pilot work led us to believe that our participants use the word “opinion” mostly to refer to their subjective evaluative beliefs, and that the term “attitude” was more ambiguous (see also Boninger, Krosnick, Berent, & Fabrigar, 1995). Second, so long as the majority of opinions are evaluative, then the construct validity of the NES would not be threatened. Third, those items assessing individuals’ need to evaluate that used the term “opinion” were highly related to all items that directly assessed the need to evaluate using terms such as “good,” “bad,” “like,” “dislike,” and “preference,” which would be expected if most opinions were in fact evaluative.

A related potential criticism might be that a number of the items on the NES refer to the strength of one’s evaluations and so the NES might actually be measuring a tendency to engage in extreme evaluation, rather than the tendency to engage in evaluation per se. The reason such items were included was to maximize the variance in participants’ scores on the NES. It is important to remember that although individual differences were predicted in the frequency of evaluative responding, even those who are low in the need to evaluate are likely to evaluate objects on a day to day basis (e.g., due to situational demands). Thus, to simply ask people whether or not they evaluate could result in a scale with limited variance, and identification of those who take evaluation very seriously would be less likely. Finally, the results of the psychometric analyses reported in Study 1 offered strong evidence that the NES was measuring primarily a single construct which would not have been expected if those items using terms such as “opinion” and “strong” were measuring a different construct than the other items. We do believe, however, that it is important in future work to examine whether individuals high in the need to evaluate do form stronger evaluations when the likelihood of evaluative responding is kept constant. In the current research, when participants expressed attitudes about paintings (Study 4) and social issues (Study 3), there were no differences in the extremity of the attitudes reported, though future research might reveal that the need to evaluate is related to other strength properties of attitudes such as their temporal stability or accessibility (see Petty and Krosnick, 1995).

A potential criticism that could be directed at the studies designed to demonstrate the predictive validity of the NES would be that participants were aware of the experimental hypothesis and responded in order to help the researchers (Orne, 1962). Three arguments are offered against the demand effect interpretation. First, in each of the validation studies, the NES was imbedded within a battery of scales which were ostensibly unrelated to the rest of the experimental session. Second, in Studies 4 and 5, no mention was ever made to participants of evaluative responding (or any other particular kind of responding) in the instructions for the thought listings. Thus, even if participants were motivated to confirm the experimenter’s hypothesis, they would not likely have been able to tell which scale was related to the hypothesis. Third, whereas the set of scales was administered at the beginning of the session in Studies 3 and 5, the set of scales was administered to participants after the behavioral validation task in Study 4 (following a lengthy and unrelated memory experiment). Because the results were highly consistent across the three studies, one would have to presume that in Studies 3 and 5, participants identified the NES as the measure of interest and then behaved in accordance with their scores and, in Study 4, behaved as they normally would have but then filled out the NES according to their behavior. This seems unlikely. Nevertheless, it would be beneficial in future work to obtain NES scores at an earlier point in time than the experimental session. It would also be useful to examine the impact of NES on measures of evaluative responding that are more subtle such as in the automatic attitude activation paradigm (Fazio, 1995), or in the use of ARC scores (Adjusted Ratio of Clustering; Roenker, Thompson, & Brown, 1971) or event-related brain potentials (Cacioppo, Crites, Berntson, & Coles, 1993) to assess evaluative categorization.

Implications and Future Directions

The current work, establishing individual differences in evaluative responding, provides a challenge to the prevailing as-

13 One possibility is that the need to evaluate is strictly based in the self-concept, but this particular aspect of the self-concept is chronically accessible. This would allow for its impact in the absence of explicit cues in the environment. We cannot address this possibility with the data we have gathered so far.
sumption that evaluation invariably pervades and dominates the responses of most people in most situations. An interesting direction for future work would be to reexamine the evidence for the pervasiveness of evaluation in light of the need to evaluate construct.

For example, one such direction would be to replicate portions of the factor analytic work of Osgood et al. (1957) to examine whether the dominance of the evaluative dimension of meaning is moderated by participants' need for evaluation. Similarly, it would be interesting to examine whether the need to evaluate moderates the likelihood and extent of the automatic attitude activation effect (e.g., Fazio et al., 1986)—especially for those objects with which participants are not highly familiar. Examination of the role of the need to evaluate in the moderation of established attitudinal phenomena would be an interesting direction for future research as well. For example, as noted previously, in work on the effects of mere thought on attitude polarization (Tesser, 1978; Tesser et al., 1995), the fact that attitudes polarize while being considered seems to depend upon the spontaneous generation of evaluative as opposed to other types of thought. If participants high in the need to evaluate are more likely to engage in such spontaneous evaluative responding, then perhaps the need to evaluate would moderate the effects of mere thought on attitude polarization. Similarly, in the Elaboration Likelihood Model of persuasion (Petty & Cacioppo, 1981, 1986b), attitude change is often thought to be mediated by one's evaluative cognitive responses that are generated when presented with a persuasive communication (Petty, Ostrom, & Brock, 1981). Based on the current research, it might be predicted that high need to evaluate individuals would be more likely to engage in such evaluative thought during exposure to persuasive communications—especially under conditions where the situation does not demand such responding. Although other individual difference variables could make predictions about how that processing would proceed (e.g., argument versus cue-based processing as a function of the need for cognition; Petty & Cacioppo, 1986a), it is plausible that low need to evaluate individuals, rather than engaging in message evaluation during exposure to the message, would be more likely to do so after the message when asked for one's attitude (cf. Haugtvedt & Petty, 1992; Mackie & Asuncion, 1990).

In addition to work on attitudes, we expect that the need to evaluate could play a useful role in a number of lines of social cognitive research examining issues such as person perception, stereotyping, perceptions of in-groups and out-groups, and the self. For example, it would be interesting to pursue whether individuals high in the need to evaluate are more likely to engage in self-evaluation as well as evaluation of issues and other people.

The effects of the need to evaluate are likely to extend beyond these basic areas of research to more applied domains. For example, in assessing responses to social surveys, the need to evaluate construct might be useful in distinguishing individuals who have previously formed evaluations versus those who are providing evaluations on the spot (Petty & Jarvis, 1996). With respect to health implications, recent studies have offered some evidence that people with highly accessible attitudes experience less stress in decision making tasks (e.g., Blascovich, Ernst, Tomaka, Kelsey, Salomon, and Fazio, 1993; Fazio, Blascovich, & Driscoll, 1992), and perhaps are better able to cope with life stressors and negative events (Fazio, 1995). It would be interesting to examine whether the need to evaluate is also associated with better coping in stressful situations as suggested by the attitude accessibility research. Also, the relation of the need to evaluate to individual difference variables such as self-esteem and depression could prove informative. Because these variables are themselves evaluative in nature, it could be predicted that the need to evaluate would play a critical role in their development and maintenance. For example, one might predict that the impact of negative life events on one's level of depression or self-esteem could be moderated by an individual's chronic tendency to engage in evaluation of the events in his or her life (e.g., as suggested by the results of Study 5).

Finally, researchers examining people's evaluations as a dependent variable might be interested in including participants' scores on the NES as a control variable. Because the NES is composed of only 16 items, it represents a convenient way to account for unexplained variance in participants' evaluative responses. Assuming that the NES will account for variance in participants' evaluative responses and will be uncorrelated with the manipulated independent variable, this reduction in error variance could make the test of any independent variable on evaluative responding more sensitive. This benefit would be in addition to any potential for discovering informative interactions between the researcher's independent variables and the need to evaluate.

**Conclusion**

In conclusion, the primary goal of the current research was achieved. The purpose of this work was to examine the hypothesis that some people consistently engage in evaluative responding more than others, and that such an individual difference could be measured. From the evidence reported in five studies, this appears to be the case. The NES was shown to be a reliable measure of a stable individual difference. Both the convergent and discriminant validity of the NES were supported, and the predictive validity of the NES was demonstrated and replicated using a variety of experimental situations and stimuli. People high in the need to evaluate were more likely to describe their previous day in evaluative terms, were more likely to engage in evaluative thought about unfamiliar paintings, and were more likely to express attitudes toward important social and political issues. Although a number of questions remain to be answered regarding the need to evaluate, the current research suggests that further examination of this construct is worthwhile.

**References**


Attitudes, self-esteem, and stereotypes. Psychological Review, 102, 4–27.


Received August 15, 1994
Revision received May 24, 1995
Accepted May 27, 1995

Low Publication Prices for APA Members and Affiliates

Keeping you up-to-date. All APA Fellows, Members, Associates, and Student Affiliates receive—as part of their annual dues—subscriptions to the American Psychologist and APA Monitor. High School Teacher and International Affiliates receive subscriptions to the APA Monitor, and they may subscribe to the American Psychologist at a significantly reduced rate. In addition, all Members and Student Affiliates are eligible for savings of up to 60% (plus a journal credit) on all other APA journals, as well as significant discounts on subscriptions from cooperating societies and publishers (e.g., the American Association for Counseling and Development, Academic Press, and Human Sciences Press).

Essential resources. APA members and affiliates receive special rates for purchases of APA books, including the Publication Manual of the American Psychological Association, and on dozens of topical books each year.

Other benefits of membership. Membership in APA also provides eligibility for competitive insurance plans, continuing education programs, reduced APA convention fees, and specialty divisions.

More information. Write to American Psychological Association, Membership Services, 750 First Street, NE, Washington, DC 20002-4242.