The Measurement of Self-Esteem: Refining Our Methods

David H. Demo

Department of Sociology and Anthropology, Mississippi State University

A review of the literature indicates that (a) very little attention has been devoted to measurement problems plaguing the study of self-esteem and (b) few studies employ more than one type of self-esteem instrument. This study addresses these issues by using eight measures of self-esteem involving self-reports, ratings by others, and a projective instrument. Their intercorrelations are examined to provide preliminary validational evidence; then, confirmatory factor analysis is used to construct measurement models and further assess the validity of the measures. The results suggest that two traditional questionnaires and a personal interview are valid in measuring *experienced self-esteem,* and three measures involving ratings by others are valid in measuring *presented self-esteem.* These findings are consistent with previous multidimensional conceptualizations of self-esteem, indicating that a variety of methods is necessary to adequately measure self-concept.

Self-esteem is a central focus of research examining human personality, and yet the conceptualization and operationalization of this variable have been both haphazard and inconclusive. There is little consensus on a definition; there is a diverse range of measurement procedures; and, in many cases, there are weak or nonexistent correlations among indicators. Hence, various findings relating to self-esteem are not comparable (Wells & Marwell, 1976; Wylie, 1974, 1979). Shavelson, Hubner, and Stanton's (1976) conclusion remains true today; that is, "Self-concept research has addressed itself to substantive problems before problems of definition, measurement, and interpretation have been resolved" (p. 410). Yet studies of the

measurement problems in self-esteem research are rare and inconsequential. A few studies examined the convergent and discriminant validity of self-report measures of self-esteem (Hamilton, 1971; Silber & Tippett, 1965; Van Tuinen & Ramanaiah, 1979); Fleming and Watts (1980) factor analyzed the Janis and Field (1959) Feelings of Inadequacy Scale; Fleming and Courtney (1984) factor analyzed the Self-Rating Scale (a revised version of the Janis & Field Scale); and Marsh, Relich, and Smith (1983) factor analyzed the Self-Description Questionnaire, which is designed to measure seven dimensions of self-concept (Shavelson et al., 1976). The current study is designed to review a variety of instruments that are intended to measure specific dimensions of self-esteem (itself a specific component of self-concept). The objectives and rationale of each measure are presented so that the validity of each can be evaluated.

The proper implementation of this procedure involves across-method triangulation (Denzin, 1970), so that several distinct methodologies can be tested rather than simply comparing scores derived from a few different attitude scales (which all share the survey method). Webb (1970) explained that because every data-gathering method has specific biases, "we should like to converge data from several data classes, as well as converge with multiple variants from within a single class"

(p. 322). It is then necessary to compare various measurement procedures by examining their intercorrelations. By examining convergence or equivalence among measures, one may be able to more easily compare findings across studies and thus construct a nomological network (Cronbach & Meehl, 1955; Golding, 1977) around the concept of self-esteem (Shavelson et al., 1976)—although it is cautioned that cross-method convergence can not be equated with construct validity.

To assess validity, it is also necessary that one carefully examine what it is one is attempting to measure. Researchers adopting the structural perspective (e.g., Coopersmith, 1967; Rosenberg, 1965, 1979) define self-esteem as a *global* positive or negative self-assessment. According to this view, self-esteem is a personality trait characterized by considerable stability from one situation to the next, even from year to year. The vast majority of self-esteem researchers thus employ one-shot questionnaires designed to measure overall or global self-esteem. Many theorists devote attention to the idea of situational variability, but available measurement techniques preclude the possibility of assessing such changes in self-feelings.

Consistent with previous research (Savin-Williams & Demo, 1983), a more processional perspective is assumed in this study. Self-esteem is viewed as a fluctuating self-attitude that most often resembles a baseline or standard self-evaluation, but that also encounters situational fluctuations from this baseline as a function of changing roles, expectations, performances, responses from others, and other situational characteristics. In this manner, individuals may have generally favorable attitudes toward themselves, possess self-respect, and consider themselves persons of worth, but on certain days and in particular situations they may feel better or worse about themselves than is typically the case. This idea is by no means new, dating back at least to James's (1890) simile of self-esteem rising and falling like a barometer, but the empirical measurement of situational variations in self-feelings is rare (see Savin-Williams & Demo, 1983; Savin-William & Jaquish, 1981). Hence, this study uses multiple and repeated measures to obtain "snapshots" of an individual's self-esteem in different social situations.

Over-reliance on traditional questionnaires used to measure global self-esteem has created another problem in that other dimensions of the self-concept have been neglected. Wells and Marwell's (1976) thorough review of self-concept methodologies demonstrated that all instruments have particular biases and "to the extent that self-esteem measurement relies on a single measurement form—orthodox verbal self-ratings—it will be inadequate" (p. 144). One alternative is to involve participant observers and peers for the purpose of exploring a behavioral component of self-esteem (Savin-Williams & Jaquish, 1981).

As traditionally conceived (Goffman, 1956; James, 1890; Rosenberg, 1979), the presented self involves a variety of planned and detailed behavioral routines that are consistent with various role requirements and situational demands, but not necessarily consistent with the actual or the desired self. Measures of *presented self-esteem* are scarce (Wells & Marwell, 1976), however, prompting the construction of a behavior checklist (Savin-Williams & Jaquish, 1981) with which observers can make judgments of others' self-esteem. This instrument is used in the current study, along with two other measures of the presented self: a Q-sort, completed by observers, and peer ratings. These measures are intended to provide information not normally obtained through self-reports. Because ratings by others are based on observation (formal and informal) of an individual's behavior over a period of time, these ratings may be more objective and more generalizable than are self-evaluations. In addition, peers and observers may be better able to assess one's personality characteristics because individuals tend to attribute their own actions and attitudes to situational factors (Jones & Nisbett, 1971). An alternative perspective is offered by Hamilton (1971), who argued that one rationale underlying self-ratings is that they capture vital personal information unavailable to others. In this article we use ratings by others to obtain a few dozen perspectives on a given individual, which we may then compare with self-ratings to identify similarities and discrepancies.

As presently conceptualized, presented self-esteem is distinct from social confidence (Fleming & Courtney, 1984; Fleming & Watts,

1980) or social self-esteem (Van Tuinen & Ramanaiah, 1979) in that the latter represent affective states of self-consciousness and shyness in social situations, whereas the former refers to a self-evaluation that one projects to others more or less intentionally. Van Tuinen and Ramanaiah defined social self-esteem as "a person's sense of adequacy or worth in his [sic] social interaction with people in general" (p. 18). Presented self-esteem, in contrast, focuses on the level of self-regard communicated to others, that is, whether individuals are comfortable with *themselves* rather than with *interactions* per se. Clearly one dimension affects the other, although social self-confidence might be expected to correlate more strongly with presented self-esteem than with self-reported (or *experienced*) self-regard. The three studies discussed above (Fleming & Courtney, 1984; Fleming & Watts, 1980; Van Tuinen & Ramanaiah, 1979) obtain strong correlations between social confidence and global self-esteem. The measures used in the current study are designed to focus on the relation of *experienced* self-esteem to *presented* self-esteem. Specifically, how do others rate an individual's self-regard based on observations of that individual's behavior? Second, how do those ratings correlate with self-ratings?

Four instruments involving traditional self-ratings are designed to measure the privately experienced dimension of self-esteem: two traditional self-report questionnaires (Rosenberg Self-Esteem Scale and Coopersmith Self-Esteem Inventory), a new repeated measures self-report scale, and a personal interview. In addition to others' ratings and self-ratings, a third method is entailed for the eighth measure of self-esteem—a projective instrument. Each of these measures and their objectives are described later. This study is thus exploratory, assuming the position that one gains more by using several measures to understand 25 or 50 individuals (from a properly drawn sample) than by relying on a single questionnaire to provide all the necessary information on several hundred persons.

Previous Research

Wells and Marwell (1976) described the self-esteem literature in general as having an "indeterminant character." Wylie (1961, 1974) was also quite critical of research in this area, arguing that there are far too many instruments used to measure self-esteem and that most are never reevaluated for their adequacy or perceived utility. Gecas's (1982) review confirmed that measurement is still a "serious problem" in self-concept research.

Studies that have examined intercorrelations among measures are also discouraging. For example, Spitzer (1969) found poor intercorrelations among three projective self-evaluation instruments. Another study (Demo & Savin-Williams, 1983) obtained only moderate correlations among three self-report measures. Examining analyses of convergent and discriminant validity, Wylie (1974) found cross-instrument correlations ranging from 0 to .81, with the average being about .4. She concluded the following:

Factor-analytic studies of instruments purporting to measure 'overall' self-esteem, self-acceptance, etc., lead one to believe that either there is no such measurable dimension as overall self-esteem, or at least some of the scales purporting to measure this construct are doing a poor job of it. (p. 101)

Certainly the unexplained variance among the instruments indicates that they are imperfect measures of a unitary concept.

The picture is even bleaker when different types of instruments are compared. Inferred measures (e.g., ratings by others) are susceptible to self-presentation and impression management, which may obscure and distort someone else's perspective of an individual's self-esteem (and other self-attitudes). So should self-reports and inferred measures correlate? The answer is of course affirmative if they claim to be measuring the same thing. Yet many studies (Combs, Soper, & Courson, 1963; Hamilton, 1971; Parker, 1966; Savin-Williams & Jaquish, 1981) report negligible correlations between self-ratings and ratings by others. Coopersmith (1959, 1967), however, found considerable correspondence between the two methods. Wells and Marwell (1976) concluded from their examination of the relevant studies that the two types of measures are distinct and thus will yield different results.

In sum, there are countless self-esteem measures and yet no firm body of evidence

with which to justify them. This research proposes to provide preliminary validational evidence for a range of methods by examining the intercorrelations among measures and using the Linear Structural Relationships computer program (LISREL) estimates of the intercorrelations to construct measurement models.

Method

Participants

The sample consists of 55 adolescents (24 males and 31 females) who were participants in a 6-year longitudinal study of adolescent self-esteem and who were enrolled in the ninth grade at a northeastern school during the 1979–1980 school year. This report is based on data collected during their 9th- and 10th-grade years because these are the years in which the most measures were administered. With the exception of three minority group members, the individuals are Caucasian and represent all socioeconomic classes and major religious identifications.

Due to many difficulties inherent in longitudinal sampling and in the administration of multiple methods, a different but largely overlapping sample exists for each measure. These and other considerations specific to each instrument are described as follows.

Eight Measures of Self-Esteem

Beeper self-reports. Of the eight self-esteem measures employed in this research, the newest and most innovative is the self-report repeated measures technique (Savin-Williams & Jaquish, 1981). The adolescent indicates from a list of adjectives, or *beep sheet* (Appendix A), the words that describe his or her self-feelings at the moment he or she is *beeped,* or signaled to respond. Participants complete the beep sheet six to eight times daily (on a random schedule) for a 1-week period. This method provides a time-sampling technique and is designed to obtain situational snapshots of self-esteem.

The operational definition of self-esteem is derived via subtracting the number of low-self-esteem words that are selected from the number of high-self-esteem words, then dividing this quantity by the total number of words selected (possible range = −1.00 to 1.00). Here, however, we are not concerned with individual beep sheets for each adolescent, but rather with the average self-esteem score obtained for each person across all contexts. This score is then compared and contrasted with other scores for the same individual obtained through separate methods.

This method represents a modification of a technique developed by Csikszentmihalyi and his colleagues (Csikszentmihalyi, Graef, & Larson, 1979; Csikszentmihalyi, Larson, & Prescott, 1977) at the University of Chicago. The objective is to measure self-feelings in naturalistic settings, removing respondents from experimental and testing situations. Savin-Williams and Jaquish (1981) argued that

> What is needed to assess self-regard more accurately are measures that tap a variety of situations or contexts

in which individuals find themselves. Such measures allow for context-specific assessment as well as an overall 'score' (which is simply some derivative of the sum of context-specific scores). (p. 331)

The subsample for the beeper method consists of 51 ninth graders. Twenty-nine of these adolescents completed beep sheets in the 10th grade. They averaged an 81% response rate to the beeps, producing a mean of 48 sheets per individual.

Self-report scales. Two traditional paper-and-pencil measures of global self-esteem are employed in this research: the Rosenberg Self-Esteem Scale (RSE) and the Coopersmith Self-Esteem Inventory (SEI). These measures involve a subsample of 41 participants (16 males and 25 females) in the ninth grade; all but 6 were included the following year. The RSE is a 10-item scale that Rosenberg (1965, 1979) reported to have good reproducibility and scalability. Such information is sample-specific, however, and therefore may not hold true on other data sets.[1]

The SEI consists of 54 items, which Taylor and Reitz (1968) found to have .90 split-half reliability, .88 test–retest reliability over 5 weeks, and .70 test–retest reliability over 3 years. Further, Robinson and Shaver (1973) reported good convergent, discriminant, and predictive validity. However, Wylie (1974) questioned its discriminant validity.

Ratings by others. Three measures of self-esteem used in this study involved others' judgments of each adolescent's self-regard: peer ratings and two forms of observer ratings. Peers rated each other by selecting a number from 1 to 5 (1 = *low self-esteem* and 5 = *high self-esteem*). A peer-based score was obtained for each participant by computing the mean of all ratings given to that individual by his or her peers. This measure involved 53 ninth graders (23 males and 30 females).

Observer ratings of adolescent self-esteem were obtained via two techniques, behavioral checklists and Q-sorts, both of which were completed by undergraduate observers

---

[1] The RSE is scored according to the Likert format in this and other studies (see Rosenberg, 1979, pp. 291–295; Wylie, 1974, pp. 180–189; Wells & Marwell, 1976). Reliability aside, many internal factor analyses (Carmines & Zeller, 1974, 1979; Kaplan & Pokorny, 1969; Kohn, 1977) have revealed two separate factors within the supposedly unidimensional RSE: Kohn identified "self-confidence" and "self-deprecation" factors, whereas Carmines and Zeller referred to the separate factors as "positive self-esteem" and "negative self-esteem." The latter research does suggest, however, that the two factors tap the same theoretical dimension of self-esteem. This conclusion is based on strikingly similar correlations between each of the 2 self-esteem factors and 16 external variables. Carmines and Zeller claimed that because "the items which load higher on the positive self-esteem factor are all worded in a positive direction while those loading on the negative self-esteem factor are all worded in a negative direction" (p. 66), it may be that response set is confounding the unidimensionality of the scale. Following Wylie's (1974) suggestion, researchers isolating separate factors should incorporate those components into multi-trait–multimethod matrices in order to assess their convergent and discriminant validity.

("big brothers/sisters") who met weekly with their same-sex adolescent. The pairs spent several hours together on each occasion, engaged in whatever activities they desired, such as eating, going to movies, and playing athletics. After each occasion together, the observer completed a behavior checklist (Savin-Williams & Jaquish, 1981), which consists of 20 behavioral descriptions (see Table 1) that obtained the highest interobserver reliability from an original list of 48 behaviors. Ten items on the checklist measure high self-esteem (e.g., sits with others during social activities, maintains eye contact, expresses opinions), and 10 behaviors measure low self-esteem (e.g., avoids physical contact, assumes a submissive stance, expresses self-deprecation).

Each checklist produced a self-esteem score by subtracting the number of low self-esteem items from the number of high self-esteem items, then dividing by 10. The resultant proportion scores (range = -1.00 to 1.00) for each checklist for each adolescent were summed, and the mean of those proportion scores provided the behavioral self-esteem score for that individual. This phase of the measurement process spanned 4 months each year

Table 1

Reliability Coefficients of Behaviors on the Behavior Checklist

Indicator	r
Positive	
1. Gives others directives or commands	.72
2. Voice quality is appropriate for situation	.69
3. Expresses opinions	.65
4. Sits with others during social activities	.54
5. Works cooperatively in a group	.76
6. Faces others when speaking or being spoken to	.84
7. Maintains eye contact during conversation	.84
8. Initiates friendly contact with others	.66
9. Maintains comfortable space between self and others	.65
10. Little hesitation in speech, speaks fluently	.69
Negative	
1. Puts down others by teasing, name calling or gossiping	.74
2. Gestures are dramatic or out of context	.87
3. Inappropriate touching or avoids physical contact	.93
4. Gives excuses for failures	.79
5. Glances around to monitor others	.57
6. Brags excessively about achievements, skills, appearance	.91
7. Verbally puts self down; self-deprecation	.81
8. Speaks too loudly, abruptly or in a dogmatic tone	.88
9. Does not express views or opinions, especially when asked	.75
10. Assumes a submissive stance	.68
Overall *r*	74.6

and involved 43 participants (21 males and 22 females) in ninth grade. Twenty-nine of these adolescents were included the following year as 10th graders.

In addition, at the end of the 12-week observation period, each observer assessed the personality characteristics of his or her adolescent using the Q-sort (Block, 1961).[2] The resultant sorting (34 ninth graders, 24 tenth graders) was then correlated with a template for ideal adolescent self-esteem.[3] Wylie (1974) and Wells and Marwell (1976) found the Q-sort particularly appropriate for conceptualizations that compare the self and ideal self. Further, the ipsative, idiographic nature of the instrument enables the observer to describe the target individual in terms of the salience of each trait for that individual rather than by comparing traits across people. It is thus a person-centered more than variable-centered approach, allowing a more interactive and dynamic perspective on the individual's personality in general and self-esteem in particular. Wylie (1974) and Wells and Marwell (1976) also indicated, however, that the validity of this technique is uncertain.[4]

Interview. The seventh measure of self-esteem used in this research involved personal interviews conducted by the author (10th grade only). As Wylie (1974) and Wells and Marwell (1976) illustrated, interviews are rarely used in self-esteem measurement (except clinically); so, little is known about their utility or validity. Each adolescent was asked 20 Likert-format questions (Appendix B), 14 of which measure self-esteem and 6 of which measure dominance.[5] The questions were asked in random order, and unclear or ambiguous answers were probed by the interviewer in order to gain clarification. The

---

[2] The California Q-Set (Form III) involves a specified 9-point distribution: 5, 8, 12, 16, 18, 16, 12, 8, 5. Wylie (1974) cited research (e.g., Livson & Nichols, 1956) that suggests that higher test–retest reliability is obtained under conditions of free sorting or sorting into a rectangular distribution than is the case for bell-shaped sorting. Others (e.g., Nunnally, 1967) claimed that without specifications on category size, the Q-sort becomes just another single-stimulus classification system, freeing the sorter from considering the structure among the elements.

[3] The Q-sort for ideal adolescent self-esteem is based on a template formed by the author and two other self-esteem researchers. High correlations with this configuration indicate high self-esteem, and low correlations represent low-self esteem.

[4] For example, Wells and Marwell (1976) reported that the most frequent criticism of the Q-sort is that its constraints on means and total scores negates the possibility of comparing across individuals or groups. On other grounds, Wylie (1974) found that "consideration of reliability problems and the output of reliability information concerning Q-sorts have been badly neglected" (p. 136). She further reported that no known multitrait–multimethod analyses have incorporated Q-sort scores, a finding that is consistent with my review and indicates uncertainty regarding the instrument's convergent and discriminant validity.

[5] The self-esteem questions were constructed by the author. The data on dominance were collected for other purposes and are not analyzed here.

interview method used here shares many features with self-reports inasmuch as the participants make their own self-descriptive judgments, but the method is also distinct in that a one-to-one social situation exists during the interview.[6] The projective instrument was administered after each interview so that these methods share the same subsample of 34 adolescents (13 males and 21 females).

Projective instrument. The eighth and final measure of self-esteem was used to explore facets of personality and cognitive dynamics that are ignored by more standardized methods. Two pictures from a modified version (Henry & Sims, 1970) of Murray's (1938) Thematic Apperception Test (TAT) were given to 34 tenth-grade participants. The stimulus for boys was a picture of a teenage boy looking into a mirror, and the stimulus for girls was a picture of a young woman looking into a mirror. The obvious assumption on which these measures rest is that the respondents unconsciously identify with the person in the picture; for this reason the stories in response to the mirror pictures were coded for self-esteem.[7]

Classification of Measures

Four of the instruments involve self-ratings: beeper self-reports, RSE, SEI, and interviewee self-descriptions. On this basis alone it is justifiable to consider these four measures as sharing one method: self-report methodology. With the exception of the beeper measure, these techniques rely on the assumption that the respondents, in one measurement situation, can accurately report a variety of self-feelings. All four self-report instruments rest on a second, and possibly more untenable, assumption that individuals are able to, and willing to, honestly state their feelings. Savin-Williams and Jaquish (1981) asserted that "Results from self-report measures might be unduly influenced by the individual's awareness, unconscious defenses, current emotional state, need for social acceptance, or to meet social desirability standards" (p. 333). Self-reports also assume that the respondent attaches equal importance to each scale item and that those items or characteristics are in fact the ones used in evaluating him or herself (Gordon, 1969). For all these reasons we expect the four self-report measures to share variance attributable both to method and to method–trait interaction.

Whereas self-report instruments purport to measure the *experienced* self, ratings by others purport to measure the *presented* self, that is, the self that is displayed in social interaction and is therefore observable (Savin-Williams & Jaquish, 1981). Three measures utilized in this research are classified as ratings by others and thus constitute the second method: peer ratings, observer checklists, and observer Q-sorts. These measures vary in the degree to which they are retrospective. The observer checklists involve measurements immediately subsequent to each observation period (which should make them more responsive to individual change than the other measures), whereas the Q-sort entails one evaluation at the end of the 12-week period, and the peer ratings rely on judgments made over the history of the particular relationship.

The third method, used in 10th grade only, is the projective instrument. This measure assumes that the

participant is unconsciously identifying with the character in the picture, so the respondent is not aware that he or she is revealing self-feelings; the threat of social desirability is thus minimized. This instrument purports to measure unconscious feelings and attitudes, and it is used here to measure unconscious self-esteem. The dissimilarities between this procedure and those described earlier justify consideration of the projective technique as an independent method.

The ninth-grade correlation matrix shown in Table 2 represents a classification of two methods and six measures. The principal criterion used in evaluating the measures is *convergent validity:* strong correlations among different measures of the same trait (Campbell & Fiske, 1959).

Findings

Examination of the ninth-grade correlation matrix provides a preliminary assessment of convergent validity. These findings should be regarded as suggestive, rather than definitive, but they are useful in making preliminary determinations and in structuring subsequent analyses, that is, measurement models.

It is evident in Table 2 that only one correlation (RSE–peer ratings) is noticeably different in the pairwise and listwise matrices.[8] Examination of the pairwise correlations re-

---

[6] This format could be detrimental to validity efforts because social desirability effects may become more pronounced than they are in the group administration of questionnaires. Because the interviewer did not know the interviewees beforehand, however, it may be that these effects are less serious. A clear advantage to this method is that all questions are answered and participants have the opportunity to clarify any items they find confusing or do not understand.

[7] Three judges read each participant's response to the mirror picture. Statements such as "Kid just got a haircut and he's proud of it" were coded high on self-esteem, whereas a common response reflecting low self-esteem was "He's sad and depressed and he thinks life's all over for him." Scores were assigned on a 1 to 5 scale, with *low self-esteem* (1) and *high self-esteem* (5). All three judges gave the same score for 17 of the 34 subjects, and 2 of the 3 judges agreed on an additional 15 responses, resulting in an intercoder reliability of .79. Data on verbal fluency were also collected using the projective stimuli; this information is described in an earlier report (Demo, 1981).

[8] A comparison of the cases included in the pairwise matrix but deleted in the listwise matrix for ninth grade shows that the means and standard deviations on each self-esteem measure are not significantly different (employing a .05 level) from those cases included in the listwise analysis. The same comparison in 10th grade yielded the same conclusion.

Table 2
Pairwise and Listwise Correlations Among Self-Esteem Measures in Ninth Grade

Method	Self-report method			Ratings by others		
	Beeper	RSE	SEI	Peer	Checklist	Q-sort
Self-report						
Beeper						
RSE	.20		.58**	.04	−.05	.18
SEI	.44**	.55***		.37*	.10	.33*
Ratings by others						
Peer ratings	.12	.32*	.41**		.51**	.52**
Observer checklist	−.25	.15	.18	.39**		.67***
Observer Q-sort	−.28	.19	.33*	.38*	.61***	

Note. Correlations below the diagonal are based on pairwise deletion (minimum $n = 22$; average $n = 35$; maximum $n = 51$). Correlations above the diagonal are based on listwise deletion, $n = 26$, with the beeper measure excluded for reasons explained in the text. The correlation between these two matrices is .58, $p = .04$. RSE = Rosenberg Self-Esteem Scale; SEI = Coopersmith Self-Esteem Inventory.
* $p < .05$. ** $p < .01$. *** $p < .001$.

veals that the beeper self-report correlates significantly with only one other measure and correlates negatively with two measures of self-esteem, indicating no convergent validity.

Each of the remaining measures correlates significantly with at least two other measures of self-esteem. Two self-report instruments (RSE and SEI) intercorrelate significantly, but RSE only exhibits one significant correlation across methods (.32 with peer ratings), indicating modest convergent validity for the RSE. The SEI correlates significantly with every other measure except the observer checklist, providing strong evidence of convergent validity.

The peer-based self-esteem scores correlate significantly with the other two measures sharing the observer method and with the RSE and SEI, indicating impressive convergent validity. The observer checklists show minimal evidence of validity, correlating significantly with the other measures sharing its method but not with any of the self-report scales. The Q-sort demonstrates strong convergent validity, intercorrelating significantly with all measures except the RSE.

LISREL

To more thoroughly test these findings, the variance–covariance matrix for the self-esteem measures was read into the LISREL computer program, and measurement models were constructed congruent with the findings just presented. The Linear Structural Relationships computer program is used because it employs maximum likelihood confirmatory factor analysis, widely accepted as the most useful technique of its kind.[9] The objective of these analyses is to isolate systematic variance—such as method variance attributable to RSE and SEI self-report questionnaires—thereby enabling a finer assessment of validity in our self-esteem measurements. This research thus focuses on directly observed variables, as well as on unmeasured, or latent, variables. Observational or measurement errors, for example, create errors in the observed variables (e.g., a measure of self-esteem), and LISREL estimates these values in separate matrices for residual covariance and measurement error covariance. The goodness-of-fit statistics obtained through this procedure determine whether the null hypothesis (e.g., the model fits) is rejected or accepted. A high chi-square value and low significance level indicate a poor fit, whereas a low chi-square value and high significance level indicate an acceptable fit between the model and the data for the corresponding degrees of freedom.

[9] Maximum likelihood estimates of the population parameters are advantageous because they are robust, asymptotic, and have a small variance. These properties are particularly important for small samples because both the standard errors of the estimates and the chi-square test of the model's fit may be imprecise.

Table 3
Factor Loadings and Unexplained Variance for Ninth Grade Measures of Self-Esteem

Measure	Loading		T value	h^2	U^2
	Factor 1	Factor 2			
RSE	1.000		0.0		
SEI	.612		2.551*	.374	.626**
Peer ratings		.627	3.177**	.393	.607**
Observer checklist		.791	4.053**	.626	.374
Observer Q-sort		.852	4.384**	.726	.274

Note. $\chi^2(5) = 7.88$, $p = .16$. RSE = Rosenberg Self-Esteem Scale; SEI = Coopersmith Self-Esteem Inventory.
* $p < .05$. ** $p < .01$.

Ninth-Grade Null Model

Because listwise matrices are more appropriate for factor analysis and because it has been demonstrated that the likelihood ratio is still meaningful with a sample size such as ours (Geweke & Singleton, 1980), we used the listwise matrix to estimate various models.[10] The first (or null) model was constructed to determine the factor loadings for five self-esteem measures on one latent variable. The beeper self-report was excluded because it demonstrated little validity and because an earlier analysis (Demo, 1981) indicated that the beeper method had a remarkably high proportion of unexplained variance. Together these findings suggest that the beeper instrument measures a construct other than self-esteem. The null model did not fit well, however, producing a chi-square value of 15.25 with 5 degrees of freedom and a probability level of 0.

Ninth-Grade Model With Two Factors

Recent research suggested separate but interacting dimensions of self-esteem. Gecas (1971) identified "self-worth" and "self-power" components; Franks and Marolla (1976) referred to "inner" and "outer" dimensions, and Savin-Williams and Jaquish (1981) posited "experienced" and "presented" selves. It was this latter finding that we wished to test. The implication is that individuals *sense* or *experience* a level of self-regard that may or may not correspond with the level of self-regard *presented* to others in social interaction. Furthermore, there is a relation between the two dimensions such that a change in one is associated with a change in the other.

An individual whose confidence and self-esteem are bolstered by some personal experience or action may be expected in subsequent situations to present himself or herself in a more self-respecting manner.

Thus, Model 2 allows for two factors, with the RSE and SEI loading on the first latent variable; this construct represents the self-report method and experienced dimension of self-esteem. Peer ratings, the observer checklist, and observer Q-sort all load on the second factor, which represents ratings by others and the presented dimension of self-esteem. The two factors are allowed to correlate because they are conceptualized as interacting dimensions of self-regard.

Table 3 shows that this model produced a significantly better fit ($p < .01$) than the null model, reducing the chi-square value by 7.37, with the same degrees of freedom, and increasing the probability level to .16. More important, however, this model attained a chi-square to degrees of freedom ratio of better than 2:1, indicating a good fit with the data.

Although the RSE had to be constrained equal to 1.0 in order to estimate the model, the four factor loadings we were able to estimate were all significant ($p < .05$). The correlation between experienced and presented factors was .12, suggesting two clearly distinct dimensions of self-esteem. Two measures (SEI and peer ratings) had significant measurement error, and the high correlation

[10] Also, we have demonstrated earlier that there are no significant differences in means and standard deviations for the pairwise and listwise matrices, and the correlation between matrices is significant ($p = .04$).

Table 4
Pairwise and Listwise Correlations Among Measures of Self-Esteem in Tenth Grade

Method	Self-report			Ratings by others	
	RSE	SEI	INT	Checklist	Q-Sort
Self-report					
RSE		.66***	.57**	.16	.41*
SEI	.65***		.48*	.30	.68***
Interview	.42*	.50***		.34	.25
Ratings by others					
Observer Checklist	.12	.13	.14		.51**
Observer Q-Sort	.45*	.63***	.25		.49**
Projective					
TAT	−.14	−.24	−.04	−.42*	−.40*

Note. Correlations below the diagonal are based on pairwise deletion (minimum $n = 20$, average $n = 29$, maximum $n = 38$). Correlations above the diagonal are based on listwise deletion, $N = 21$, with TAT excluded for reasons explained in the text. The correlation between these two matrices is .91, $p = .00$. RSE = Rosenberg Self-Esteem Scale; SEI = Coopersmith Self-Esteem Inventory; INT = interview; TAT = Thematic Apperception Test.
* $p < .05$. ** $p < .01$. *** $p < .001$.

between RSE and SEI (Table 2) suggests that these scales shared method variance that we were unable to estimate.

Reexamining Model 2: Validity and Reliability

Model 2 produced a good fit with the data, providing a second mechanism for assessing the validity of the six measures involved. The correlations between the observed scores and latent variables, heretofore referred to as factor loadings, represent the validities of the measures. The squared correlations between the observed scores and the latent variables represent the reliabilities of each measure. These values are shown in Table 3: validities corresponding to factor loadings and reliabilities indicated by h^2.

Summarizing the ninth-grade findings, the two observer measures fared very well in terms of both validity and reliability, whereas the SEI and peer ratings demonstrated strong validity but lower reliability as measures of self-esteem.

Tenth-Grade Null Model

To further examine the validity of the measures and to test the two-dimensional factor structure, another combination of measures was analyzed involving the 10th-grade data. The correlations shown in Table 4

indicate that the projective instrument lacks any semblance of convergent validity as a measure of self-esteem, so it is excluded from further consideration. The null model for the remaining five measures obtained a chi-square of 9.99, 5 degrees of freedom, and a probability level of .07. This suggests an acceptable fit for a one-factor model, but a two-factor model was estimated to determine if a better fit would result.

Tenth-Grade Model With Two Factors

Consistent with the ninth-grade results, a good model was obtained via positing a self-report or experienced factor, an observer or presented factor, and an interaction between the two dimensions of self-esteem. This model (Table 5) reduced the chi-square by 6.94 with 2 less degrees of freedom, indicating a significantly ($p < .05$) better fit than did the one-factor model.

All five factor loadings were significant, and the correlation between factors increased to .57 ($p < .05$) in the 10th grade. One correlated error is also incorporated into this model, suggested by the largest correlation in Table 4: SEI and Q-sort. These two instruments are substantially longer than any of the other measures and may be tapping additional traits beyond self-esteem. This error was not significant, however, and only the

Table 5
Factor Loadings and Amount of Unexplained Variance for Tenth-Grade Measures of Self-Esteem

| Measure | Loading | | *T* value | h^2 | U^2 |
	Factor 1	Factor 2			
RSE	.814		3.848**	.662	.338
SEI	.780		3.651**	.609	.391
Interview	.692		3.203**	.479	.521*
Observer checklist		.571	2.065*	.326	.674*
Observer Q-sort		.836	2.794**	.699	.301
Correlated Error					
SEI–Q-sort			1.719		.293

Note. $\chi^2(3) = 3.05$, $p = .38$. RSE = Rosenberg Self-Esteem Scale; SEI = Coopersmith Self-Esteem Inventory.
* $p < .05$. ** $p < .01$.

observer checklist and interview had significant measurement error ($p < .05$). In sum, all five measures exhibited validity, but the reliability of the interview and observer checklist was considerably weaker than that of the other instruments.

The 9th- and 10th-grade data are consistent, then, in supporting a two-dimensional model of self-esteem, but the correlation between the two dimensions is vastly different for the 2 years. The data also support the validity of three self-report measures (RSE, SEI, interview) and three inferred measures (peer ratings, observer checklist, and observer Q-sort).

Discussion

This study has explored the validity of self-esteem measurement and compiled rather consistent evidence for the eight measures employed in this research. The beeper measure and TAT each obtained only one significant correlation with other measures of self-esteem, indicating poor convergent validity for both measures. The TAT may be measuring a distinct trait such as nonphenomenal self-regard (which might not be expected to correlate with phenomenal measures; see Wylie, 1974), but it is more likely tapping an unrelated construct such as imagination. The beeper instrument is more curious, capitalizing on repeated measures but apparently measuring a separate variable (e.g., nonevaluative self-feelings). It is suggested here that the poor validity of this measure is attributable to the beep sheet (Appendix A) and the

selection of situational feelings (e.g., happy, relaxed, depressed, or frustrated) that may be independent of self-regard; one need not be happy in order to like oneself. Attaching a different measure (e.g., RSE or SEI) to the beeper method may facilitate the important and necessary task of examining self-conception within a temporal framework.

Confirmatory factor analyses substantiated the validity of the RSE and SEI, two traditional self-report inventories, as measures of experienced self-esteem. The interview also fared well in this regard, suggesting that this method is underused. If indeed there are separate dimensions of self-esteem that require distinct methods for their measurement, the opportunity to probe respondents and clarify information broadens our perspective on the experienced dimension beyond that obtained through traditional questionnaire data. Employing a distinct methodology offers the advantage of permitting statistical analyses of intercorrelations among measures of the same construct, as well as improved estimates of method variance.

Two new measures, peer ratings and observer checklist, demonstrated clear validity as measures of presented self-regard. Furthermore, the observer Q-sort (an established instrument used in a new format) was the strongest measure of this dimension, suggesting that there are many possible techniques for assessing this and other dimensions of self-conception that are as yet unexplored. Certainly the utility of ratings by others, advocated by Savin-Williams and Jaquish

(1981), has been supported in this research, suggesting a solid alternative to traditional paper-and-pencil measurement. Furthermore, these new measures should be used to clarify the fine distinctions between such dimensions as experienced self-esteem, presented self-esteem, and social self-esteem.

As we have indicated, the LISREL findings corroborated two separate dimensions of self-esteem: experienced self-regard measured by self-report and presented self-regard measured by specific others. However, the correlation between these factors was weak in 9th grade but strong in 10th grade, again illustrating that further research is necessary if we are to adequately understand the complexities of self-conception. One implication is that the line of research initiated by Gecas (1971, 1972) and Franks and Marolla (1976) should be pursued, enabling a fuller understanding of inner self-esteem or SE power and outer self-esteem or SE-worth, as well as the relation of these dimensions to experienced and presented self-esteem. For example, in self-report situations are we more inclined to report inner self-esteem (i.e., personal feelings of competence and effectiveness), whereas ratings by others are more dependent on reputational or outer criteria? Measurement considerations of this type will become increasingly important as researchers move away from unidimensional conceptualizations of the self to explore specific components, such as social confidence (Fleming & Courtney, 1984; Fleming & Watts, 1980) or self-efficacy (Gecas, 1982; Gecas & Schwalbe, 1983), and to assess the developmental change and stability of various dimensions (Savin-Williams & Demo, 1984).

At present, however, we lack a sufficiently clear and comprehensive conceptual framework for understanding self-conception, or even self-esteem, and thus we cannot simply equate cross-method convergence with construct validity (Golding, 1977). A large amount of method variance remains undefined, necessitating further correlational research (see Marsh et al., 1983 for an excellent illustration), as well as the type of logical and experimental work outlined by Shavelson et al. (1976). We must also keep in mind that self-esteem measures fail to tap many other dimensions of self-conception and that every

measurement approach must be evaluated in terms of the particular aspect of self it purports to capture (Gordon, 1969).

In sum, the validity of six measures of self-esteem has been supported. In light of the present findings and Wells and Marwell's (1976) observation that interviews are rarely used in self-esteem measurement, further use of this method is certainly warranted. Traditional scales such as the RSE and SEI have also been validated, though their assumptions and limitations should be understood. Lastly, the construct validity of ratings by others has been suggested, providing reasonable measures to supplement, rather than replace, more orthodox procedures for assessing self-esteem.

References

Block, J. (1961). *The Q-sort method in personality assessment and pyschiatric research.* Springfield, IL: Charles C Thomas.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Carmines, E. G., & Zeller, R. A. (1974). On establishing the empirical dimensionality of theoretical terms: An analytic example. *Political Methodology, 1,* 75–96.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment: Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-017.* Beverly Hills, CA: Sage.

Combs, A. W., Soper, D. W., & Courson, C. C. (1963). The measurement of self-concept and self-report. *Educational and Psychological Measurement, 23,* 493–500.

Coopersmith, S. (1959). A method for determining types of self-esteem. *Journal of Abnormal and Social Psychology, 59,* 87–94.

Coopersmith, S. (1967). *The antecedents of self-esteem.* San Francisco: Freeman.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Csikszentmihalyi, M., Graef, R., & Larson, R. (1979, September) *Age differences in the quality of subjective experience.* Paper presented at the meeting of the American Psychological Association, New York City.

Csikszentmihalyi, M., Larson, R., & Prescott, S. (1977). The ecology of adolescent activity and experience. *Journal of Youth and Adolescence, 6,* 281–294.

Demo, D. H. (1981). *Theoretical and methodological considerations in the study of self-esteem.* Unpublished doctoral dissertation, Cornell University.

Demo, D. H., & Savin-Williams, R. C. (1983). Early adolescent self-esteem as a function of social class: Rosenberg and Pearlin revisited. *American Journal of Sociology, 88,* 763–774.

Denzin, N. K. (1970). *The research act in sociology.* Chicago: Aldine.

Fleming, J. S., & Courtney, B. E. (1984). The dimensionality of self-esteem: II. Hierarchical facet model for revised measurement scales. *Journal of Personality and Social Psychology, 46,* 404–421.

Fleming, J. S., & Watts, W. A. (1980). The dimensionality of self-esteem: Some results for a college sample. *Journal of Personality and Social Psychology, 39,* 921–929.

Franks, D. D., & Marolla, J. (1976). Efficacious action and social approval as interacting dimensions of self-esteem: A tentative formulation through construct validation. *Sociometry, 39,* 324–341.

Gecas, V. (1971). Parental behavior and dimensions of adolescent self-evaluation. *Sociometry, 34,* 466–482.

Gecas, V. (1972). Parental behavior and contextual variations in adolescent self-esteem. *Sociometry, 35,* 332–345.

Gecas, V. (1982). The self-concept. *Annual Review of Sociology, 8,* 1–33.

Gecas, V., & Schwalbe, M. L. (1983). Beyond the looking-glass self: Social structure and efficacy-based self-esteem. *Social Psychology Quarterly, 46,* 77–88.

Gewecke, J. F., & Singleton, K. J. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. *Journal of the American Statistical Association, 75,* 133–137.

Goffman, E. (1956). *The presentation of self in everyday life.* Edinburgh, Scotland: University of Edinburgh.

Golding, S. L. (1977). Method variance, inadequate constructs, or things that go bump in the night. *Multivariate Behavioral Research, 12,* 88–98.

Gordon, C. (1969). Self-conceptions methodologies. *Journal of Nervous and Mental Disease, 148,* 328–364.

Hamilton, D. L. (1971). A comparative study of five methods of assessing self-esteem, dominance, and dogmatism. *Educational and Psychological Measurement, 31,* 441–452.

Henry, W. E., & Sims, J. H. (1970). Actor's search for self. *Transaction, 7,* 57–62.

James, W. (1980). *Principles of psychology.* New York: Holt.

Janis, I. L., & Field, P. B. (1959). A behavioral assessment of persuasibility: Consistency of individual differences. In C. I. Hovland & I. L. Janis (Eds.), *Personality and persuasibility* (pp. 29–54). New Haven, CT: Yale University Press.

Jones, E. E., & Nisbett, R. E. (1971). *The actor and the observer: Divergent perceptions of the causes of behavior.* Morristown, NJ: General Learning Press.

Kaplan, H. B., & Pokorny, A. D. (1969). Self-derogation and psychosocial adjustment. *Journal of Nervous and Mental Disease, 149,* 421–434.

Kohn, M. (1977). *Class and conformity: A study in values.* Chicago: University of Chicago Press.

Livson, N. H., & Nichols, T. F. (1956). Discrimination and reliability in Q-sort personality descriptions. *Journal of Abnormal and Social Psychology, 52,* 159–165.

Marsh, H. W., Relich, J. D., & Smith, I. D. (1983). Self-

concept: The construct validity of interpretations based upon the SDQ. *Journal of Personality and Social Psychology, 45,* 173–187.

Murray, H. (1938). *Explorations in personality.* New York: Oxford University Press.

Nunnally, J. (1967). *Psychometric theory.* New York: McGraw-Hill.

Parker, J. (1966). The relationship of self report to inferred self concept. *Educational and Psychological Measurement, 26,* 691–700.

Robinson, J., & Shaver, P. (1973). *Measures of social psychological attitudes.* Ann Arbor, MI: Institute for Social Research.

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Rosenberg, M. (1979). *Conceiving the self.* New York: Basic Books.

Savin-Williams, R. C., & Demo, D. H. (1983). Situational and transsituational determinants of adolescent self-feelings. *Journal of Personality and Social Psychology, 44,* 824–833.

Savin-Williams, R. C., & Demo, D. H. (1984). Developmental change and stability in adolescent self-concept. *Developmental Psychology, 20,* 1100–1110.

Savin-Williams, R. C., & Jaquish, G. A. (1981). The assessment of adolescent self-esteem: A comparison of methods. *Journal of Personality, 49,* 324–336.

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research, 46,* 407–441.

Silber, E., & Tippett, J. S. (1965). Self-esteem: Clinical assessment and measurement validation. *Psychological Reports, 16,* 1017–1071.

Spitzer, S. (1969). Test equivalence of unstructured self-evaluation instruments. *Sociological Quarterly, 10,* 204–215.

Taylor, J., & Reitz, W. (1968). *The three faces of self-esteem* (Research Bulletin No. 80). London, Ontario, Canada: University of Western Ontario, Department of Psychology.

Van Tuinen, M., & Ramanaiah, N. V. (1979). A multimethod analysis of selected self-esteem measures. *Journal of Research in Personality, 13,* 16–24.

Webb, E. J. (1970). Unconventionality, triangulation, and inference. In N. K. Denzin (Ed.), *Sociological methods: A sourcebook* (pp. 322–328). New York: McGraw-Hill.

Wells, L. E., & Marwell, G. (1976). *Self-esteem: Its conceptualization and measurement.* Beverly Hills, CA: Sage.

Wylie, R. C. (1961). The self-concept: A critical survey of pertinent research literature. Lincoln: University of Nebraska Press.

Wylie, R. C. (1974). *The self-concept: Vol. 1. A review of methodological considerations and measuring instruments* (rev. ed.). Lincoln: University of Nebraska Press.

Wylie, R. C. (1979). *The self-concept: Vol. 2. Theory and research on selected topics* (rev. ed.). Lincoln: University of Nebraska Press.

(Appendixes follow on next page)

Appendix A

Table 1A
Beep Sheet

Date: Time: am/pm

As you were beeped
  Where were you?
  What things were you doing?
  What were you thinking about?
  Wish you were doing something else? What?
  Who were you with? Give the number, age, sex, and relationship (parents, sister, friend, acquaintance, stranger, etc.) of those present.
Time is passing (circle appropriate dot):
  Fast . . . . . Slow
How would you describe yourself at the moment beeped? Circle as many words below as are appropriate:

Inhibited	Happy	Skilled	Left behind
Clear	Relaxed	Productive	Exposed
Consistent	Free	Unloved	Fussy
Tense	Sluggish	Useless	Loved
Confident	Lonely	Growing	Bored
Unprepared	Powerful	Overwhelmed	Unsure
Belonging	Empty	Affectionate	On time
Weak	Ashamed	Depressed	In control
Safe	Proud	Needed	Conforming
Spontaneous	Secure	Frustrated	Manipulated

Other feelings about yourself:

Appendix B

Table 1B
Self-Esteem/Dominance Interview

Question	Question
1. How does it feel to be 15?	11. Do you feel like an important person in the school?
2. Do you feel down, or depressed, very much? What is it like?	12. Do you speak up when you have something to say in school or at home?
3. How often do you feel really good? Can you describe it?	13. Do you think you can meet the challenges ahead of you?
4. Do you like yourself? How much? Why not?	14. Do you think you're a good leader?
5. In what situations do you feel best about yourself? Worst?	15. Do you enjoy having authority over other people?
6. Would you say your friends like you? How much?	16. Do you enjoy planning things and deciding what each person in a group should do?
7. Can your friends depend on you?	17. Is it easy for people to win arguments with you?
8. Do you think your life so far has been a success or a failure?	18. Do you have a natural talent for influencing people?
9. If you could change something(s) about yourself, what would it/they be?	19. Do you like to give orders and get things moving?
10. Overall, do you like the way other people treat you?	20. Is there anything else about yourself you'd like to say?

Note. Questions were asked in random order, except that the interview always started with Question 1 and ended with 20. Individual items numbered 1–13 and 20 were scored *high self-esteem* (5) and *low self-esteem* (1), yielding scale scores of 14–70. Items numbered 14–19 were scored *high dominance* (5) and *low dominance* (1), yielding scale scores of 6–30.