DOCUMENT RESUME

ED 242 772 TM 840 189

AUTHOR Marsh, Herbert W.
TITLE The Bias of Negatively Worded Items in Ratings Scales for Preadolescent Children: A Cognitive-Developmental Phenomenon.
PUB DATE 20 Feb 84
NOTE 28p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Cognitive Development; Correlation; Elementary Education; Factor Structure; *Negative Forms (Language); Preadolescents; Rating Scales; Reading Achievement; *Reading Difficulties; *Self Concept Measures; Test Bias; *Test Construction; *Test Items

ABSTRACT

Negative item bias is produced by the inability of preadolescent children to respond appropriately to negatively worded items on rating scales, and is hypothesized to be a cognitive-developmental phenomenon. The effect is examined with responses to the Self Description-Questionnaire (SDQ), a multifactor self-concept instrument. In study 1, response to positive and negative items were uncorrelated in grade 2 but were substantially correlated by grade 5. In study 2, confirmatory factor analysis of response by grade 5 students demonstrated that the negative items contributed both to the scale they were designed to measure and to a "negative item" factor. The negative item factor was nearly uncorrelated with any of the self-concept factors, but was substantially correlated with reading achievement. The two studies demonstrate that younger children and children with poorer reading skills are less able to respond appropriately to negatively-worded items, and that this effect produces a bias in their response to the SDQ. This supports the contention that the effect is a cognitive-developmental phenomenon. (Author/PN)

The Bias of Negatively Worded Items In Ratings Scales For Preadolescent Children: A Cognitive-Developmental Phenomenon

Herbert W. Marsh
University of Sydney, Australia

20 February, 1984

Running Head: Negative Items

The Bias of Negatively Worded Items In Ratings Scales For Preadolescent Children: A Cognitive-Developmental Phenomenon

ABSTRACT

The negative item bias is produced by the inability of preadolescent children to respond appropriately to negatively worded items on rating scales, and is hypothesized to be a cognitive-developmental phenomenon. The effect is examined with responses to the Self Description Questionnaire (SDQ), a multifactor self-concept instrument whose factor structure, reliability, and validity have been clearly demonstrated in numerous other studies. In study 1, responses to positively and negatively worded items were compared for children (n = 658) in grades 2 - 5. Particularly in grade 2, children frequently responded "true" to negative items, indicating a very poor self-concept, even when their other responses indicated a very positive self-concept. Responses to positive and negative items were uncorrelated (-0.02) in grade 2, but were substantially correlated by grade 5 (0.60). In study 2 confirmatory factor analyses of responses by year 5 students (n = 559) demonstrated that the negative items contributed both to the scale they were designed to measure and to a "negative item" factor. The negative item factor was nearly uncorrelated with any of the self-concept factors, but was substantially correlated with reading achievement (0.42). Taken together, the two studies demonstrate that younger children and children with poorer reading skills are less able to respond appropriately to negatively worded items and that this effect produces a bias in their response to the SDQ. This supports the contention that the effect is a cognitive-developmental phenomenon.

Test construction specialists argue for the use of some negatively worded items on personality, attitude and other rating scale instruments in order to disrupt response sets such as responding to all items with the same response category. This procedure is particularly useful for single-scale instruments where all items are designed to measure the same construct. For multiscale instruments the practice seems less useful, and the confirmation of the scales through procedures such as factor analysis provides a test for such a response set. The use of negatively worded items assumes that they measure the same construct as positively worded items. However, this assumption is rarely tested and its validity seems questionable when respondents are preadolescent children. In order to respond appropriately to negatively worded items, respondents often have to invoke a double negative logic that requires a higher level of verbal reasoning than do positively worded items. For example, the item "I am NOT a good student" requires a response of "False" to indicate that "I am a good student". If this logic is not appropriately employed, respondents may give an answer which has exactly the opposite meaning to their intended response. For purposes of this study a negative item bias is defined to be when a child responds inappropriately by saying "true" to a negative statement when their responses to positive items have consistently indicated that the opposite response would be more appropriate, or vice versa. Such an effect will create a method/halo bias that is specific to the negative items.

Development of the SDQ.

The Self Description Questionnaire (SDQ) is a multifactor self-concept instrument for preadolescent children. Its factor structure, reliability, and validity have been clearly demonstrated in numerous studies. The SDQ is designed to measure seven factors of self-concept derived from the Shavelson model (Shavelson & Bolus, 1982; Shavelson , Hubner & Stanton, 1976) and six independent factor analyses of responses by disparate groups have each identified these factors (Marsh, Barnes, Cairns & Tidman, in press; Marsh, Relich & Smith, 1983; Marsh, Smith & Barnes, 1983b). The SDQ scales are reliable (coefficient alphas in the .80's and .90's), moderately correlated with measures of the corresponding academic abilities (r's from 0.3 to 0.7 -- see Marsh, Parker & Smith, 1983; Marsh & Parker, in press; Marsh, Smith & Barnes, 1983a), in agreement with self-concepts inferred by primary school teachers (see Marsh, Parker & Smith, 1982;

Marsh, Smith, Barnes & Butler, 1983) and reasonably stable over time (Marsh, Smith, Barnes & Butler, 1983).

In the development of the SDQ, unlike the SDQ II and the SDQ III which are designed for older subjects, negative items were found to be ineffective in defining the different areas of self-concept they were designed to measure. Preliminary analyses indicated that negatively worded items contributed less to the internal consistency of the scales, and exploratory factor analyses sometimes revealed a negative item factor (i.e., a factor on which only negatively worded items loaded). Younger children in particular often responded "true" to negative items, indicating a very poor self-concept, when their responses to positive items consistently indicated a positive self-concept. This suggested that the problem might be a cognitive-developmental phenomenon. In subsequent revisions considerable care was taken in the wording of the negative items so that they were clearly negative and avoided the problem of double negative reasoning as much as possible. Thus, an item like "I do not like mathematics" was changed to "I hate mathematics." However, numerous attempts to revise the negative items failed to solve the problem and led to the recommendation that these items should not be included when scoring the SDQ (Marsh, Barnes, Cairns & Tidman, in press). The purposes of this study are to examine more carefully the effect as a cognitive-developmental phenomenon, and to explore its relationship to other theoretical and methodological perspectives.

Theoretical and Methodological Perspectives

A wide range of observations from disparate areas of research appear to be related to the negative item bias. Theoretical findings in developmental psychology and psycholinguistics may provide a basis for understanding the effect, while methodological approaches and findings from personality and achievement testing may provide research designs helpful in the study the phenomenon. A review of the relevant research in each of these areas is beyond the scope of this study, but it is important to delineate these areas.

A Developmental/Psycholinguistic Perspective. Slobin (1971), and Klima and Bellugi-Klima (1971) indicate that the concept of negativity develops very early as is evident in primitive two-word sentences (e.g., not hungry), but they point out that for complex sentences the negative transformation of an affirmative sentence becomes more difficult since the negative element cannot just be placed at the start or end of the phrase. Braine and Rumain (1983) examined 17 inference schemas of reasoning and the ages at which they are

exhibited. The schema most relevant to the negative item bias is exemplified by: it is false that there is not a "W", therefore there is a "W". (p. 278). In their review, Braine and Rumain found limited developmental research on this schema but reported that 20% of kindergartners and 90% of 10-year-olds could appropriately apply this type of inverse reasoning, suggesting that "canceling a negative develops in the early school years" (p. 285). Researchers have also identified cognitive-developmental stages in children's ability to apply other forms of inverse reasoning. Attribution researchers (Kun, 1977; Nicholls, 1978; also see Marsh, Cairns, Relich, Barnes & Debus, in press) have found that children as young as five understand that ability and effort each contribute to the likelihood of success, but it is not until age 10 or later that children understand that less effort is required to achieve success if the subject is more able. This research indicates that while the concept of negation develops very early, the inverse reasoning needed to correctly respond to negative r[illegible]ng items probably develops during early school years.

Personality Research. The tendency for subjects to respond to personality rating items independently of the content has been variously referred to as response set, response bias, response style, or a method/halo effect, and different approaches emphasize the nonsubstantive or substantively irrelevant components of responses to structured items (see Wiggins, 1973 for a review). Jackson (1967; Jackson & Messick, 1958, 1961) argues that content is what is left over after sources of style and method have been removed through approaches such as regression and factor analysis. Most response style research considers the effects of response tendencies such as social desirability, where subjects attribute to themselves socially desirable characteristics, or acquiescence, where subjects tend to agree to items as self-descriptive independent of the item content. In a study of acquiescence, Trott and Jackson (1967) suggested that the influence of style increased when subjects were given less time to study each item and when each item was more clearly related to the content dimension that it was designed to measure, but that it was uncorrelated with verbal ability for university students. While many possible causes of response styles have been considered, they are generally not considered as a cognitive-developmental phenomenon as they are here. Nevertheless, the negatively worded item bias does qualify as a response set as conceptualized by Wiggins, and correlational approachs similar to those described by Jackson are

employed in the present investigation.

Positive and negative items designed to measure the same construct are sometimes found to define two separate factors when examined with empirical procedures such as factor analysis. Naylor reported (1978) that responses by university students to a state-anxiety inventory produced two factors representing respectively positive and reverse scored items. Androgeny research (e.g., Spence, Helmreich & Holahan, 1979; Antill, Cunningham, Russell & Thompson, 1981) has found that items designed to measure masculinity and femininity actually define four factors; masculinity and femininity are each defined by two separate factors representing positive-valued and negative-valued items. In instances such as these, it is not clear whether the difference between positive-item and negative-item factors is substantive, nonsubstantive, or substantively irrelevant. Nevertheless, it seems that these examples differ from the phenomenon examined here in that they occur with subjects who have the cognitive-developmental ability to respond appropriately to the negative items.

Responses to a personality test, particularly by preadolescent children, may measure a different construct than that which the test was intended to measure. For example, Bridgeman and Shipman (1978) reported that preschool self-concept was significantly correlated to year 3 scores in reading and math but not to year 3 self-concept. This led the authors to speculate that the preschool measure of self-concept was measuring some construct besides self-concept that was correlated with achievement. Even though their preschool self-concept measure did not require reading, it probably required a level of verbal reasoning that was difficult for many preschoolers. Hence, the preschool responses may have had a substantial verbal component that biased the interpretation of self-concept, but was predictive of subsequent reading performance. Ironically, such a bias would inappropriately make responses to a self-concept instrument seem to be more valid when assessed against reading achievement or related measures of academic achievement.

Achievement Testing. Achievement tests are designed to measure mastery of a particular body of knowledge or proficiency in specific skills. Cronbach (1971; 1980) and others argue that it is not only important that a test measure what it is supposed to measure, but also that it not measure what it is not supposed to measure. As an example, Cronbach (1980, p. 106) described a content specific test where the content "is all too often a Sleeping Beauty screened off from the student by tangled clauses and thorny (but pointless!)

jargon." He suggests that any item that is as highly correlated with scores on a reading comprehension test as the total score on the specific achievement test has serious invalidity for evaluating the content. Cronbach (1971) describes factor analytic and correlational techniques for separating content-specific variance from that due to other causes, in much the same way as personality researchers examine the effect of response styles.

The Present Investigation.

The purposes of the present research are to determine how the negatively worded items are related to the SDQ scales as defined by positively worded items, to grade level, and to reading achievement. Data in study 1 of the present investigation come from previous research designed to examine the effect of age and sex on self-concept (Marsh, Barnes, Cairns & Tidman, in press). In that study it was demonstrated that: 1) separate exploratory factor analyses of responses to the positively worded items from each of four age levels clearly identified the SDQ scales; 2) a linear, negative relationship existed between age and most of the self-concept scales; 3) student sex affected several scales in a manner consistent with sex stereotypes, but that was independent of age; and 4) the SDQ scales became more distinct with age. Confirmatory factor analyses, using LISREL, were subsequently performed, supporting conclusions 1 and 4 (Marsh & Hocevar, 1984; Marsh & Shavelson, 1983). In study 1 the responses to negatively worded items are added to those analyzed previously to examine the negative item bias and its relationship to age. New data are collected for study 2 where tests are made of confirmatory factor analytic models in which a negative item factor was explicitly defined. Verbal ability measures are incorporated into these models to determine how responses to the negative items are related to reading ability.

STUDY 1

METHOD.

Samples and Procedures . Two independent samples were used in study 1. The first sample consisted of the 170 second grade (primarily seven year olds) and the 251 fifth grade children (primarily 10 year olds) who attended one of four public coeducational schools in Sydney, Australia. Communities served by these schools varied in social economic status from lower and lower-middle class to middle and upper-middle class. Across all the children in this sample, academic abilities tended to be about average. The second

sample in this study consisted of the 103 third grade children (primarily 8 year olds) and 134 fourth grade children (primarily 9 year olds) who attended one of two public coeducational schools in Sydney, Australia. Neither of these schools was the same as in sample 1. Children in this second sample were somewhat below average in terms of academic ability, and tended to come from families in the lower, lower-middle, and middle social classes.

In study 1 the two samples are not equivalent, but the the grade levels within each of the samples were selected so as to provide a strong control against linear age effects being the result of nonequivalent samples. Since the youngest and oldest children in study 1 come from the same sample, any differences due to nonequivalent samples would produce a nonlinear age effect where the results for children in grades 2 and 5 would differ systematically from those in grades 3 and 4. Thus, while this design is biased against the demonstration of linear age effects, it provides a stronger control against such an effect being the result of nonequivalent groups than is typically available (see Marsh, Barnes, Cairns & Tidman, in press, for further discussion).

In both samples, the SDQ was administered during a regular class session approximately one third of the way through the school year, and was the first measure to be administered as part of a more extensive battery of tests. The SDQ was administered by one of the authors of that study according to standardized procedures developed in previous research. Students responded to each item along a five point scale which varies from "1 - False" to "5 - True". The SDQ was read aloud to children to minimize reading difficulties, and they responded to several examples before any of the SDQ items were presented. The children were specifically instructed not to say their responses aloud or talk to other pupils. As a consequence of earlier research the SDQ was read aloud at a fairly rapid pace, and the whole questionnaire required approximately 8 minutes to administer (not including time for the instructions and examples), though children were given time at the end of the administration to go back to any items that they had left blank.

Analysis in this study is based upon student responses to 60 items designed to measure seven SDQ factors. A total of 56 items (eight per factor) are positively worded, while the remaining 10 are negatively worded. A brief description of the seven SDQ factors is as follows:

Physical Abilities/Sports (PHYS) -- student ratings of their ability and enjoyment of physical activities, sports, and games.

Physical Appearance (APPR) -- student ratings of their own attractiveness, how their appearance compares with others, and how others think they look.

Relationship With Peers (PEER) --student ratings of how easily they make friends, their popularity, and whether others want them as a friend.

Relationship With Parents -- student ratings of how well they get along with their parents, whether parents are easy to talk to, whether their parents like them, and whether they like their parents.

Reading (READ) -- student ratings of their ability in and their enjoyment/interest in reading.

Mathematics (MATH) -- student ratings of their ability and enjoyment/interest in mathematics.

School Subjects (SCHL) -- student ratings of their ability and enjoyment/interest in "all school subjects".

Statistical Analysis. All the statistical analyses described in study 1 were conducted with the commercially available SPSS program (Hull & Nie, 1981; Nie, et al., 1975). Before any analyses were performed, the responses to the negatively worded items were reflected so that all items varied along a scale where 1 represented the lowest level of self-concept and 5 the highest. Then a value of 4.0, the average response, was substituted for all missing responses (less than 1/4 of 1%).

RESULTS and DISCUSSION.

The purpose of the first set of analyses is to confirm that negatively worded items are less consistent with other items in the scale they are designed to measure than are positively worded items. A series of item analyses (Hull & Nie, 1981) were conducted for the total sample and separately for each grade level. For the total sample, the coefficient alphas for every scale and the average correlation among items within each scale were higher when the negative items were excluded (see Table 1). This replicates findings from earlier research. However, examination of the results for the different grade levels demonstrates that this effect depends upon age. For the younger children, the exclusion of the negatively worded items consistently produces the largest improvement in the coefficient alphas. Also, the negative items form a scale with reasonable internal consistency -- particularly for the youngest pupils.

Insert Table 1 About Here

For the total sample, the coefficient alpha for responses to the set of positive items is .93 and for negative items is .73. Thus, to the extent that the two sets of items are measuring the same construct then they should correlate approximately .8 or higher with each other (i.e., within the limits of the reliabilities of the two means). For

the total sample, the correlation between responses to the two sets of items is only .27, indicating that they are measuring different constructs (see Table 2). Furthermore, the results illustrate a dramatic developmental effect. For the youngest children the two sets of responses are uncorrelated (r = -.02), while the correlations are much larger for the oldest children (r= .60). Thus, for the youngest children the negative items are measuring a construct that is unrelated to self-concept, while for the oldest children the negative item responses are substantially related to positive item responses but still contain considerable variance that is reliable and unique. These results clearly justify the decision to exclude the negatively worded items in scoring the SDQ, but they also suggest that the method effect is developmentally related to the age of the subjects.

Insert Table 2 About Here

The self-concepts scores of preadolescent children have been quite high in all studies with the SDQ. Consistent with those results, the average response to positively worded items is about 4 on a five-response scale (i.e., a response of "Mostly True" to positive statements). If children are responding appropriately to the negative items, then the average response to them should also be quite high (after responses to the negative items have been reflected). However, if some children are responding inappropriately by saying "True" or "Mostly True" to negative items when their intended meaning is the opposite, then the means for the negative items should be much lower (i.e., indicate a poorer self-concept) and the standard deviations higher.

For positively worded items the mean response across all scales shows a consistent and marked decline with age ($r = -.20$, $p < .001$). Marsh, Barnes, Cairns and Tidman (in press) demonstrated that this effect is consistent across most of the SDQ scales and is primarily a linear effect. In marked contrast, the average response to the negatively worded items shows a marked increase with age ($r = .23$, $p < .001$). For the youngest children, responses to the negatively worded items are much lower than to the positive items (see Table 2). It is only for the oldest children that the mean response to positive and negative items is approximately the same. This suggests that some children are inappropriately giving responses of "True" and "Mostly True" to negative items when in fact they have positive self-concepts. Also consistent with this conclusion are the larger standard deviations for responses to negative items by younger children,

suggesting that some children are responding inappropriately while others are not.

In summary, some children at each grade level seem to respond inappropriately to negatively worded items. The phenomenon is clearly age related and occurs more frequently with younger children. Since this bias is systematic rather than constant or random, it is particularly serious. These findings support the decision not to include responses from negatively worded items in the scores derived from the SDQ, but they also have important implications for other rating scales designed for use by children and for the further study of this effect as a cognitive-developmental phenomenon.

STUDY 2

The results of study 1 show that responses to negatively worded items are influenced by a method/halo effect and that this effect varies with age. The negative items apparently require a higher level of verbal reasoning in order to respond appropriately, and this is why the effect is larger for the younger children. Despite the intuitive appeal of this explanation, study 1 suffers important weaknesses which limit the strength of the conclusions. The use of exploratory factor analyses in the original research (Marsh, Barnes, Cairns & Tidman, in press) precluded a test of whether negatively worded items contributed to a "negative item factor", to the appropriate scale which the item was designed to measure, or to both. The suggestion that the negative item bias is systematically related to verbal reasoning or reading ability could not be tested directly, since reading scores were not available. Instead, this inference was based upon the finding that the negative item effect varied for different age groups and that the younger children have poorer verbal skills.

The purpose of study 2 is to further examine these issues with procedures which overcome the weaknesses. A new sample of fifth grade pupils completed the SDQ and two verbal achievement tests, and were ranked in terms of their reading ability by their teachers. Results of study 1 showed that the negative item bias was weaker for fifth grade students compared with younger children, but it was still evident. For study 2 confirmatory factor analytic (CFA) models were tested which required that negative items load on the factor which they were designed to measure, on a separate negative item factor, or on both. The verbal ability measures were also incorporated into these models in such a way that the relationship between the negative item bias and verbal ability could be tested. Since students from only one grade level were considered, the effect of age must be minimal and any

effect of reading achievement must be relatively independent of age.

METHOD.

Sample and Procedures. Pupils in study 2 were a new sample of 559 fifth grade students (mostly 10 year olds) enrolled in 19 fifth grade classes in one of seven private Catholic schools in Sydney, Australia. None of these schools were the same as employed in study 1. Most of the students attended single-sex classes (18 of the 19 classes) Children in the sample came from families which varied in socioeconomic status from lower-middle to upper-middle class. Across all the children in study 2 the academic abilities were about average. Data considered in study 2 are part of a larger project which is described in more detail by Marsh, Smith & Barnes (1983a). For purposes of this analysis, consideration is limited to pupil responses to the SDQ, results from two verbal ability tests, and teacher ratings of each pupil's reading ability.

The SDQ was administered in the same manner as described in study 1, but a slightly revised version of the SDQ was employed in study 2. This version of the SDQ contained 76 items -- the additional 10 items were designed to measure general self-concept or self-esteem. Thus, the current version of the SDQ is designed to measure eight factors -- the seven described earlier and a general-self scale. Of the 76 items, a total of 12 were negatively worded (two for each of the three academic scales and the general self scale, and one each for the four nonacademic scales).

The two achievement tests of verbal ability were the Comprehension test and the Word Knowledge test of the Primary Reading Survey Tests (ACER, 1976). The Word Knowledge test consists of 40 multiple choice synonym items and takes 20 minutes to answer. The Comprehension test consists of 34 multiple-choice items and takes 30 minutes. In addition, teachers were asked to judge the reading ability of each child along a scale that varied from "1 - Very Poor" to "9 - Very Good", thus providing a third measure of reading ability.

The achievement tests were distributed to the schools by the researchers, but were actually administered by the classroom teachers during a regular class session before the administration of the SDQ. The tests were then scored by the researchers with the understanding that feedback would be given to the schools after completion of the study. Two of the schools declined to participate in the achievement testing, though they did agree to the administration of the SDQ and to complete teacher ratings. The SDQ was administered along with other

materials during a regularly scheduled class while teachers were asked to complete a teacher rating form for each child. Some teachers did not actually complete the forms until later, and one teacher eventually declined to complete the forms at all.

Statistical Analyses. The CFA in study 2 were performed with the commercially available LISREL V program (Joreskog & Sorbom, 1981). With LISREL V the researcher is able to define alternative factor solutions designed to test different hypotheses, and to compare the ability of competing models to fit the original data (see Joreskog & Sorbom, 1981; Long, 1983). The LISREL V program, after testing for identification, attempts to minimize a maximum likelihood function which is based upon differences between the original and reproduced covariance matrix, and provides an overall chi-square goodness-of-fit test (Joreskog & Sorbom, 1981; Maruyama & McGarvey, 1980). For large complex problems with large sample sizes, the observed chi-square will nearly always be statistically significant, and alternative indications of goodness-of-fit are required. The most commonly used alternative is the ratio of the chi-square to the degrees-of-freedom (df) in the model. However, this value is still directly related to the sample size such that the same solution will lead to a much larger ratio when based upon more cases. Other indices have been developed which are not affected by sample size. LISREL V presents the root mean square residual (RMS) which is based upon the residual covariances -- the difference between the original and reproduced correlations in this example. Bentler & Bonnett (1980) developed an index called coefficient d which scales the observed chi-square along a scale which varies from zero to 1.0. The zero point represents the chi-square obtained from a null model (normally one which results in a reproduced covariance matrix which is diagonal) and 1.0 represents an ideal fit. Thus, it is like an estimate of the variance which can be explained by a given model.

In preliminary analyses, the factor structure underlying the 64 positively worded items from the SDQ (i.e., 8 items from each of eight scales) was examined. For purposes of this and subsequent analyses, each scale was defined by four variables representing the total response to a pair of items. Within each scale, the first two items which were positively worded defined the first item pair, the next two the second item pair, and so forth. This is the same procedure used by Marsh, Barnes, Cairns and Tidman (in press) and other SDQ research (see Marsh & O'Neill, in press, for further discussion). In the next series of analyses the 12 negatively worded items were included,

only one negative item. A series of different CFA models were defined in which each negative item was required to load only on the factor which it was designed to measure, only on a ninth, "negative item" factor, or on both. The ability of each of these models to fit the data was tested. In the final set of analyses the three reading scores were used to define a reading ability factor, and this factor was related to the self-concept factors and to the negative item factor.

All the analyses were based upon a 47 x 47 correlation matrix representing the 32 positively worded item pairs, the 12 negatively worded items, and the three reading scores. For the self-concept responses there was almost no missing data (less than 1/10 of 1% of the responses) and the mean response was substituted for the few missing values. However, for the teacher ratings of reading ability there were 36 missing values (6%), representing primarily students from one class where the teacher did not complete the ratings, and 142 missing values (25%) for the reading tests, representing primarily students from two schools which did not administer the achievement tests. For purposes of this study pair-wise deletion of missing data was used in the determination of the correlation matrix (see Nie, et al., 1976). However, a similar correlation matrix based upon only those cases which had no missing data for the three reading measures was virtually the same as the one which was actually used. Thus, while the large number of missing values for the reading scores does require that the results be interpreted cautiously, it is unlikely to have any substantial effect.

RESULTS.

CFA of the Positively worded Item Pairs. In CFA (confirmatory factor analysis) alternative models are specified by fixing or constraining elements in three matrices which are conceptually similar to matrices resulting from common factor analysis. These are:

1) LAMBDA Y, a matrix of factor loadings;

2) PSI, a factor correlation matrix which represents the relationships among the factors; and

3) THETA EPSILON, a diagonal matrix of error/uniqueness terms that are conceptually similar to one minus the communality estimates in exploratory factor analysis.

The results of the CFA (see Table 3) illustrate the pattern of parameters to be estimated in these three matrices, using only the positive item pairs. All coefficients with a value of "0" or "1" are fixed (i.e., predetermined) and not estimated as part of the analysis,

while other parameters are free and estimated in the analysis. For this problem 32 measured variables are used to define eight factors. The free parameters consist of 32 factor loadings in LAMBDA Y, the 28 correlations among the eight factors in PSI, and the 32 error/uniquenesses in THETA. This factor pattern is very restrictive in that it allows each variable to load on one and only one factor, and represents an ideal of "simple structure." The parameter estimates (see Table 3) indicate that each of the eight self-concept factors is well-defined. The goodness-of-fit indices (see Table 4) indicate that the model adequately explains the data. Despite the large sample size, the chi-square/df ratio is only slightly larger than 2, while the values for RMS and coefficient d each indicate that the fit is good.

Insert Tables 3 & 4 About Here

CFA of Positive & Negative Items. In the second set of CFA models, the 12 negatively worded items are added to the variables shown in Table 3. In models 2.1 - 2.3 each negative item is required to load only on the self-concept factor that is was designed to measure (model 2.1), or only on a ninth, negative item factor (model 2.2), or on both the self-concept factor and the negative item factor (model 2.3 -- see Table 5). Inspection of the goodness-of-fit indices (see Table 4) indicates that model 2.3 provides the best fit to the data. Thus, variance in responses to the negative items represents both the factors which the items were designed to measure and a method/halo effect.

Insert Table 5 About Here

The parameter estimates for model 2.3 (see Table 5) indicate that the inclusion of the negative items has virtually no effect on the parameter estimates for the positively worded item pairs; the parameter estimates in Table 5 are nearly the same as in Table 3. Factor loadings on the self-concept factors are smaller for the negatively worded items than for the positively worded items, but all of the loadings are statistically significant ($p < .01$). The loadings for the negative items are somewhat smaller on the negative item factor than for the self-concept factors, but 11 of the 12 loadings on the negative item factor are also statistically significant ($p < .01$). Correlations between the negative item factor and the self-concept factors (in the PSI matrix) are all close to zero and only the correlation with Reading self-concept ($r = .15$) reaches significance at the .01 level (the correlation with Math self-concept reaches

significance at $p < .05$). This demonstrates that the negative item bias is nearly uncorrelated with any of the self-concept scales.

Model 2.4 differs from 2.3 only in that the 8 correlations (in the PSI matrix) between the negative item factor and the self-concept factors were fixed to be zero. Inspection of the goodness-of-fit indices demonstrates that this model fits the data nearly as well as model 2.3 (in which each of these correlations were estimated but were observed to be close to zero). Despite the large sample size the difference in chi-squares between models 2.3 and 2.4 fails to reach statistical significance at $p < .01$, though it is significant at $p < .05$ (chi-square difference = 16, df = 8, $p < .05$).

CFA With Reading Measures. In the third set of analyses, the three reading measures were added to the variables described in models 2.0 - 2.4. In each instance the three reading measures were used to define an additional factor called reading ability. The three reading measures were free to load on this additional factor, but not on any other factors. Again, model 3.3, where the negative items were allowed to load on both the self-concept factors and the negative item factor, was able to explain the data substantially better than models in which the negative items loaded only on the self-concept factors or only on the negative item factor. Also, model 3.4, where correlations between the negative item factor and the self-concept factors were fixed to be zero, was nearly indistinguishable from model 3.3.

Inspection of the parameter estimates for model 3.3 (see Table 6) shows that for the self-concept variables -- both the positively and negatively worded items -- the estimates are nearly the same as for model 2.3. The Reading ability factor is well defined in that each of the three variables designed to define it loads substantially on that factor. The Reading Ability factor correlates substantially with Reading self-concept ($r = .43$), but not with any of the other self-concept factors. The Reading Ability factor is also substantially correlated with the Negative Item factor ($r = .42$).

The correlations between the Reading Ability factor and the other factors in model 3.3 are particularly important for this study. The negative item factor represents a method/halo bias, and these results show that this bias is substantially correlated with reading ability. Children with poorer reading skills are more likely to respond "True" to negatively worded items rather than to respond in a manner consistent with their responses to positive items. The finding that reading ability is only correlated with Reading self-concept, but not with other self-concept factors, further demonstrates the

distinctiveness of the different self-concept factors. In summary these findings demonstrate that negative items contribute significantly to both the scale they were designed to measure and a to negative item bias. The negative item bias is nearly uncorrelated with the self-concept factors but is substantially correlated with reading achievement.

DISCUSSION & OVERVIEW

In each study, the results suggest that negatively worded items are often responded to inappr riately by preadolescent children. When forced to use the more difficult reasoning required by the negatively worded items, children often respond "True" or "Mostly True", implying a poor self concept, even though their responses to positively worded items indicate that they have favorable self-concepts (see footnote 1). This phenomenon is more likely to occur for younger children and for children with poorer reading ability. Since most children have high self-concepts (i.e., the average response is 4 on a 5 point response scale), children who are younger and/or who have poorer reading skills will inappropriately appear to have systematically lower self-concepts than other children merely as an artifact of the negative item bias. The demonstration of the substantial correlation between reading achievement and the negative item bias in a single year group indicates that the effect of reading on the bias is relatively independent of age. The negative item effect will bias interpretations of self-concept scores so that they erroneously appear to be more highly correlated to reading achievement and other academic achievement scores that are frequently used to validate self-concept measures, and so that comparisons across age groups are invalid.

While the results of these two studies clearly justify the decision to exclude responses from the negatively worded items when scoring the SDQ, several features of the present investigation may limit the generalizability of the conclusions. Trott and Jackson (1967) found that a method effect varied with the amount of time subjects had to study each item, and so if the SDQ items were presented at a slower pace the negative item bias might be smaller. Furthermore, the complications involved in using a five-point response scale may have exacerbated the negative item bias. However, Marsh and Smith (1982) identified a substantial negative item factor in responses by fifth and sixth grade students to the Coopersmith Self Esteem Instrument. On the Coopersmith, half the items are negatively worded, subjects respond to each item with either a "Like me" or "Not

Like Me" response, and students were given longer to respond to each item. Hence, the results from the Coopersmith instrument indicate that the negative item effect may generalize to instruments in which a larger proportion of the items are negatively worded, the response scale has only two categories, and the pace of item presentation is slower.

This investigation is based on responses to a self-concept instrument, but it is likely that a similar phenomenon occurs with other rating instruments as well. The double negative logic required to answer negative items appropriately is not limited to self-concept items, and the negative items on the SDQ were more carefully constructed to avoid this problem than is typically the case with other rating scales. Also, while the findings of this study are limited to the responses of preadolescent children, a similar phenomenon may occur with the responses of older subjects. Thus, the type of analysis described here -- particularly the CFA with the inclusion of verbal ability measures -- provides a model for other studies to examine the operation of such halo/method effects.

The focus of this study has been on the effect of negatively worded items as a bias to rating instruments that are used by preadolescent children. However, the contention that the effect is a cognitive-developmental phenomenon was strongly supported, and further research into the substantive aspects of this effect should prove valuable. The results of study 1 show that there is a dramatic developmental shift during early school years in the ability of preadolescent children to respond appropriately to this type of rating item. These results correspond with the conclusion by Braine and Rumain (1983) about the age at which children can appropriately use inference schemas of reasoning that require double negative logic. The results of study 2 show that within a single grade level, there were substantial individual differences in the size of the effect and these are related to verbal achievement. Hence, the substantial effect of verbal achievement in study 2 is relatively independent of age, even though the age effect in study 1 was confounded by differences in verbal achievement. Further research is clearly needed to relate this cognitive-developmental effect to cognitive stages of early development considered in other research.

FOOTNOTES

1 - Preadolescent children typically use the most favorable three response categories when responding to self-concept items on a five-point response scale, indicating positive self-concepts. This makes it relatively easy to recognize when children with the most favorable self-concepts are responding inappropriately to negatively worded items, since the appropriate and inappropriate responses are at opposite ends of the response scale. However, it is more difficult to recognize when children with the least favorable self-concepts are responding inappropriately, since both appropriate and inappropriate responses would tend to be near the middle of the response scale. Consequently, attempts to estimate the frequency of ocurrenc[of the negative item bias are likely to underestimate its actual occurrence.

REFERENCES

Antill, J. K., Cunningham, J. D., Russell, G., & Thompson, N. L. (1981). An Australian sex-role scale. Australian Journal of Psychology, 33, 169-183.

Australian Council for Educational Research (ACER). (1976). Primary Reading Survey Tests A - D. Hawthorn, Victoria, Australia, ACER.

Bentler, P. M. & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin,88, 588-606.

Bridgeman, B. & Shipman, V. C. (1978). Preschool measures of self-esteem and achievement motivation as predictors of third-grade achievement. Journal of Educational Psychology, 70, 17-28.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? In W. B. Schrader (Ed.), Measuring Achievement: Progress Over a Decade (p. 99-108). New directions for testing and measurement. San Francisco: Jossey-Bass.

Cronbach, L. J. (1971). Test Validation. In R. L. Thorndike (Ed.), Educational Measurement (2nd Ed., p. 443-507). Washington D.C.: American Council on Education.

Hull, C. H. & Nie, N. H. (1981). SPSS Update 7 - 9. New York, McGraw-Hill.

Jackson, D. N. (1967). Acquiescence response styles: Problems in identification and control. In I. A. Berg (Ed.), Response set in personality assessment (p. 71-111). Chicago: Aldine.

Jackson, D. N. & Messick, S. (1958). Content and style in personality assessment. Psychological Bulletin, 55, 243-252.

Jackson, D. N. & Messick, S. (1961). Acquiescence and desirability as response determinants on the MMPI. Educational and Psychological Measurement, 21, 771-790.

Joreskog, K. G. & Sorbom, D. (1981). LISREL V: Analysis of Linear Structural Relations By the Method of Maximum Likelihood. Chicago: International Educational Services.

Klima, E. S. & Bellugi-Klima, U. (1971). Syntactic regularities in the speech of children. In A. Bar-Adon & W. F. Leopold, Child language: A book of readings (p. 412-424). Englewood Cliffs, NJ: Prentice-Hall.

Kun, A. (1977). Development of the magnitude-covariation and compensatory schema in ability and effort attributions of performance. Child Development, 48, 862-873.

Long, J. S. (1983). Confirmatory Factor Analysis. Beverly Hills: Sage.

Marsh, H. W., Barnes, J., Cairns, L. & Tidman, M. (in press). The Self Description Questionnaire (SDQ): Age effects in the structure and level of self-concept for preadolescent children. Journal of Educational Psychology. (in press).

Marsh, H. W., Cairns, L., Relich, J., Barnes, J., & Debus, R. L. (1984). The relationship between dimensions of self-attribution and dimensions of self-concept. Journal of Educational Psychology, (in press).

Marsh, H. W. & Hocevar, D. (1984). The Application of confirmatory factor analysis to the study of self-concept: First and higher order factor structures and their invariance across age groups. School of Education, University of Southern California. (in review)

Marsh, H. W. & Parker, J. W. (in press). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well. Journal of Personality and Social Psychology.

Marsh, H. W., Parker, J. W. & Smith, I. D. (1983). Preadolescent self-concept: Its relation to self-concept as inferred by teachers and to academic ability. British Journal of Educational Psychology, 53, 60-78.

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, 1983, 45, 173-187.

Marsh, H. W. & Smith, I. D. (1982). Multitrait-multimethod analyses of two self-concept instruments. Journal of Educational Psychology, 74, 430-440.

Marsh, H. W., Smith, I. D. & Barnes, J. (1983a). Multidimensional Self-concepts: Relationships with sex and academic achievement. Department of Education, University of Sydney. (in review).

Marsh, H. W., Smith, I. D. & Barnes, J. (1983b). Multitrait-multimethod analyses of the Self Description Questionnaire: Student-teacher agreement on multidimensional ratings of student self-concept. American Educational Research Journal, 20, 333-357.

Marsh, H. W., Smith, I. D., Barnes, J. & Butler, S. Self-concept: Reliability, dimensionality, validity, and the measurement of change. Journal of Educational Psychology, 75, 772-790.

Martin, D. S. & Rumain, B. (1983). Logical Reasoning. In J. H. Flavell & E. M. Markman (Volume Eds.), P. H. Mussen (Ed.), Handbook of child psychology: Cognitive development (Vol. III, p. 263-340). New York: Wiley.

Maruyama, G. & McGarvey, B. (1980). Evaluating causal models: An application of maximum likelihood analysis of structural equations. Psychological Bulletin, 87, 502-512.

Naylor, F. D. (1978). Success and failure experiences and the factor structure of the State-Trait Anxiety Inventory. Australian Journal of Psychology, 30, 217-226.

Nicholls, J. (1978). The development of the concept of effort and ability, perceptions of academic attainment, and the understanding that difficult tasks require more ability. Child Development, 49, 800-814.

Nie, N. H., Hull, C. H., Jenkins, J. G. Steinbrenner, K. & Bent, D. H. (1975). Statistical Package for the Social Sciences. New York: McGraw-Hill.

Shavelson, R. J. & Bolus, R. (1982). Self-concept: The interplay of theory and methods. Journal of Educational Psychology, 74, 3-17.

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

Spence, J. T., Helmreich, R. L., & Holahan, C. K. (1979). Negative and positive components of psychological masculinity and femininity and their relationships to self-reports of neurotic and acting out behaviors. Journal of Personality and Social Psychology, 37, 1673-1682.

Trott, D. M. & Jackson, D. N. (1967). An experimental analysis of acquiescence. Journal of Experimental Research in Personality, 2, 278-288.

Wiggins, J. S. (1973). Personality and prediction: Principles of personality assessment. Menlo Park, CA:Addison-Wesley.

Wylie, R. C. The self-concept (Rev. ed., Vol. 1). (1974). Lincoln: University of Nebraska Press.

Wylie, R. C. The self-concept (Vol. 2). (1979). Lincoln: University of Nebraska Press.

TABLE 1

Coefficient Alphas and Average Item Intercorrelations for Scales With and Without Negatively Worded Items

Scale	Grade 2		Grade 3		Grade 4		Grade 5		Total	
	With	Without	With	Without	With	Without	With	Without	With	Without
PHYS										
alpha	.66	.78	.65	.71	.76	.80	.75	.78	.69	.77
avg r	.22	.31	.19	.23	.26	.32	.26	.30	.23	.30
APPR										
alpha	.83	.85	.80	.81	.89	.90	.88	.87	.86	.87
avg r	.36	.41	.32	.35	.48	.52	.46	.46	.42	.45
PEER										
alpha	.74	.83	.70	.72	.84	.87	.80	.81	.78	.82
avg r	.28	.38	.21	.24	.38	.45	.32	.45	.30	.36
PRNT										
alpha	.64	.80	.58	.66	.69	.77	.76	.79	.69	.76
avg r	.24	.33	.16	.20	.24	.29	.29	.33	.24	.30
Total Non-Academic										
alpha	.88	.92	.84	.5	.91	.92	.89	.88	.88	.90
avg r	.20	.27	.13	.15	.23	.26	.18	.20	.18	.22
READ										
alpha	.75	.84	.74	.74	.86	.86	.90	.90	.84	.86
avg r	.26	.39	.23	.26	.40	.41	.49	.53	.36	.44
MATH										
alpha	.79	.82	.83	.87	.90	.90	.90	.91	.86	.89
avg r	.30	.46	.33	.45	.47	.54	.49	.55	.40	.51
SCHL										
alpha	.76	.82	.83	.83	.83	.83	.85	.85	.81	.84
avg r	.25	.36	.32	.38	.34	.38	.36	.41	.31	.39
Total Academic										
alpha	.90	.93	.92	.92	.91	.90	.93	.93	.91	.92
avg r	.24	.35	.27	.32	.26	.28	.31	.35	.27	.34
Total Self										
alpha	.93	.95	.91	.92	.94	.94	.93	.93	.93	.93
avg r	.19	.26	.15	.17	.18	.21	.17	.19	.17	.18
Negative Items										
alpha	.79		.65		.67		.63		.73	
avg r	.26		.16		.17		.15		.21	

TABLE 2

Means and Standard Deviations For Positively and Negatively Worded Items; and Correlations Between the Two Sets of Items

Grade Level	Positively Worded Items (N = 56 Items) Mean	SD	Negatively Worded Items (N = 10 Items) Mean	SD	Correlations Between Two Item Sets
2	4.24	0.69	3.43	1.10	- .02
3	4.12	0.54	3.88	0.75	+ .42 *
4	4.03	0.57	3.91	0.67	+ .60 *
5	3.97	0.56	4.00	0.63	+ .59 *
Total	4.02	0.59	3.84	0.77	+ .27 *

* p < .01

TABLE 3
LISREL Maximum Likelihood Estimates For Parameters in Model 1.1: 8 self-concept Factors (positively worded items only)

Factor Loading Matrix (LAMBDA)

Variables	PHYS	APPR	PEER	PRNT	READ	MATH	SCHL	GENL	Uniqueness/ error
Phys1	79*	0	0	0	0	0	0	0	37*
Phys2	84*	0	0	0	0	0	0	0	30*
Phys3	81*	0	0	0	0	0	0	0	34*
Phys4	80*	0	0	0	0	0	0	0	35*
Appr1	0	70*	0	0	0	0	0	0	52*
Appr2	0	63*	0	0	0	0	0	0	61*
Appr3	0	85*	0	0	0	0	0	0	28*
Appr4	0	81*	0	0	0	0	0	0	34*
Peer1	0	0	76*	0	0	0	0	0	42*
Peer2	0	0	78*	0	0	0	0	0	39*
Peer3	0	0	75*	0	0	0	0	0	44*
Peer4	0	0	82*	0	0	0	0	0	32*
Prnt1	0	0	0	48*	0	0	0	0	77*
Prnt2	0	0	0	52*	0	0	0	0	72*
Prnt3	0	0	0	81*	0	0	0	0	35*
Prnt4	0	0	0	84*	0	0	0	0	30*
Read1	0	0	0	0	86*	0	0	0	26*
Read2	0	0	0	0	87*	0	0	0	24*
Read3	0	0	0	0	84*	0	0	0	29*
Read4	0	0	0	0	84*	0	0	0	30*
Math1	0	0	0	0	0	84*	0	0	30*
Math2	0	0	0	0	0	88*	0	0	23*
Math3	0	0	0	0	0	89*	0	0	22*
Math4	0	0	0	0	0	91*	0	0	16*
Schl1	0	0	0	0	0	0	78*	0	40*
Schl2	0	0	0	0	0	0	66*	0	56*
Schl3	0	0	0	0	0	0	78*	0	39*
Schl4	0	0	0	0	0	0	85*	0	28*
Genl1	0	0	0	0	0	0	0	66*	57*
Genl2	0	0	0	0	0	0	0	75*	44*
Genl3	0	0	0	0	0	0	0	80*	36*
Genl4	0	0	0	0	0	0	0	71*	49*

Correlations Among Factors (PSI)

Factors	PHYS	APPR	PEER	PRNT	READ	MATH	SCHL	GENL
PHYS	1							
APPR	42*	1						
PEER	66*	49*	1					
PRNT	40*	27*	47*	1				
READ	16*	08	17*	06	1			
MATH	28*	21*	26*	24*	13*	1		
SCHL	40*	26*	34*	31*	43*	74*	1	
GENL	75*	49*	80*	53*	27*	40*	54*	1

* p < .01

Note: Parameters with values of 0 and 1 were fixed and not estimated as part of the analysis. The four measured variables designed to measure each factor are the sum of responses to pairs of positively worded items.

TABLE 4

Summaries of Goodness of Fit Indices for the CFA Models Containing Self-concept (SC), Negative Item, and Reading Ability factors

Model Description	chi-square	df	chi-sq/df ratio	RMS	Coeff d
1) Positive Items Only					
1.0 Null Model	11,263	496	22.70	.305	.00
1.1 Full Model (see Table 3)	1,020	436	2.34	.044	.91
2) Positive & Negative Items					
2.0 Null Model	14,163	946	14.97	.263	.00
2.1 8 SC factors with Neg items on SC factors only	2,250	874	2.57	.056	.84
2.2 8 SC factors & 1 Neg item factor with Neg items on Neg item factor only	2,808	866	3.24	.077	.80
2.3 8 SC factors & 1 neg item factor with neg items on both SC & neg item factors (see Table 5)	1,822	854	2.13	.046	.87
2.4 Same as 2.3 with corr's between SC factors & neg item factor set to 0	1,838	862	2.13	.048	.87
3) Positive & Negative Items, and Reading Ability Factor					
3.0 Null Model	15,415	1081	14.26	.253	.00
3.1 Model 2.1 with reading ability factor	2,715	998	2.72	.062	.82
3.2 Model 2.2 with reading ability factor	3,224	989	3.26	.077	.79
3.3 Model 2.3 with reading ability factor (see Table 6)	2,234	977	2.29	.050	.86
3.4 Model 2.4 plus reading ability factor	2,251	985	2.29	.052	.85