Child Development, August 1998, Volume 69, Number 4, Pages 1030–1053

Structure, Stability, and Development of Young Children's Self-Concepts: A Multicohort–Multioccasion Study

Herbert W. Marsh, Rhonda Craven, and Raymond Debus

A new, individual administration procedure for assessing multiple dimensions of self-concept for young children 5–8 years of age (Marsh, Craven, & Debus) was the basis of this study. We expanded this application in a multicohort-multioccasion (MCMO) study that provides simultaneous multicohort comparisons (cross-sectional comparisons of different age cohorts) and longitudinal comparisons of the same children on multiple occasions. There was reasonable support for predictions that reliability, stability, factor structure, and the distinctiveness of the SDQ factors would improve with age (a between-group age cohort comparison) and from 1 year to the next (a longitudinal comparison), and that small gender differences were reasonably stable over age. Consistent with the proposal that children's self-perceptions become more realistic with age, Time 1 (T1) teacher ratings were more highly correlated with student self-ratings at T2 than T1 and contributed to the prediction of T2 self-concept beyond effects mediated by T1 self-concepts. The results support and expand the surprisingly good support for the multidimensionality of self-concept responses for very young children using this procedure.

INTRODUCTION

Research on self-concept development continues to be focused predominantly on middle childhood and adolescent years (e.g., Dusek & Flaherty, 1981; Harter, 1985, 1986; Hattie, 1992; Stipek, 1981; Stipek & MacIver, 1989; Wigfield, 1994; Wigfield, Eccles, MacIver, Reuman, & Midgley, 1991), but there has been increasing attention given to the structure and development of self-concept in very young children (Byrne, 1996; Eccles, Wigfield, Harold, & Blumenfeld, 1993; Harter, 1983, 1985, 1986; Harter & Pike, 1984; Marsh, Craven, & Debus, 1991). A major focus of this research has been to clarify the emergence and progressive differentiation of more specific facets of self-concept in the early childhood years. Innovative approaches in developing measures and procedures for their administration to young children below the age of 8 years (Byrne, 1996; Harter & Pike, 1984; Marsh et al., 1991) have contributed to changing perspectives on the emergence of a more differentiated structure of self-concept.

Predictions about how self-concept and its factorial structure evolve with age have been proposed from a variety of theoretical perspectives. Shavelson, Hubner, and Stanton (1976) hypothesized that self-concept becomes more differentiated with age. Marsh (1985, 1990; Marsh, Barnes, Cairns, & Tidman, 1984), expanding on the Shavelson et al. hypothesis, proposed that self-concepts of very young children are consistently high but that with increasing life experience children learn their relative strengths and weaknesses so that with increasing levels of age, mean levels of self-concept decline, individual self-concept becomes more differentiated, and self-concept becomes more highly correlated with external indicators of competence (e.g., skills, accomplishments, and self-concepts inferred by significant others). Markus and Wurf (1987) noted that the structure of self depends on both the information available to an individual and the cognitive ability to process this information. Reflecting changes in the cognitive ability to process self-relevant information, Harter (1983, 1985) proposed that self-concept becomes increasingly abstract with age, shifting from concrete descriptions of behavior in early childhood, to trait-like psychological constructs (e.g., popular, smart, good looking) in middle childhood, to more abstract constructs during adolescence. According to her conceptual model, the concept of a global self-worth did not evolve before the age of about 8. Eccles et al. (1993; also see Wigfield & Eccles, 1992) developed an expectancy-value model of academic choice in which self-perceptions of competency and task value ratings in different domains form the basis of subsequent academic choice and have recently examined the domain specificity of responses by very young children. Based on earlier research (e.g., Nicholls, 1979; Parsons & Ruble, 1977; Stipek & MacIver, 1989), Eccles et al. proposed that the declines in mean levels of competency self-ratings reflected an optimistic bias for very young children and increased accuracy in responses by young children as they grow older.

For older children, there have been considerable advances in the quality of self-concept research due to stronger theoretical models, and the development of multidimensional measurement instruments based on theoretical models (see Byrne, 1984, 1996; Harter, 1983, 1985, 1986; Marsh, 1990, 1993a; Marsh & Craven, 1997; Marsh & Hattie, 1996). These advances have not, however, been fully applied to research with young children in part because psychometrically strong, multidimensional instruments have not been developed for young children (see Harter, 1983; Marsh & Craven, 1997; Marsh et al., 1991; Stipek & MacIver, 1989; Wylie, 1989). As proposed by Harter (1983, 1985; Harter & Pike, 1984), the effective measurement of self-concept with very young children may require simplified item contents or pictorial representations, simplified response formats, and individually based interviews instead of conventional paper-and-pencil tests that are group administered. Perhaps, as appears to have been the case for research with older children, progress in theory, research, and practice for very young children will be stimulated by the development of better multidimensional measurement instruments.

THE MULTIDIMENSIONALITY OF SELF-CONCEPT RESPONSES BY YOUNG CHILDREN

In this section, three issues are discussed that form the basis of this study: (1) how best to measure self-concept for young children to study the development of a differentiated factor structure of self-concept, (2) substantive developmental issues such as age differences and the development of gender differences in self-concept responses by young children, and (3) the accuracy of self-concept inferred by significant others and how this varies with age.

Measurement of Young Children's Self-Concepts

Harter and Pike (1984) developed an instrument to measure self-concept scales (physical, cognitive, peers, and maternal) using items represented by parallel verbal statements and pictures. They found that below age 8 children either did not understand general self-worth items or did not provide reliable responses, prompting them to exclude this scale from their instrument. Their exploratory factor analyses supported only two scales: competence (incorporating the physical and cognitive scales) and social acceptance (incorporating the peer and maternal scales). The authors noted that the factor structure was less differentiated than typically found for older children and that there was no differentiation among competencies in specific areas, thus supporting the frequently noted assumption that the structure becomes more differentiated with age. Their failure to support their a priori four-factor structure, however, provides a weak basis of inference about the structure of self-concept, particularly when they did not use confirmatory factor analysis (CFA) that allows the researcher to specify the model to be tested (e.g., Marsh & Hocevar, 1985). Marsh et al. (1991) suggested that the failure to separate even the physical and academic components that are so robust in responses by slightly older students was surprising. They noted, for example, that correlations among the physical and academic scales reported by Harter and Pike (*r*s of .43–.56) did not approach 1.0 even after correction for unreliability and that there was support for their differentiation in relation to other criteria reported by Harter and Pike (e.g., teacher ratings, choice behavior, being held back a grade). Finally, even support for their conclusion that self-concept becomes more differentiated with age was not strong in that Harter and Pike neither reported administering their instrument to older children nor offered any evidence that self-concept became more differentiated within the 4–7 age range that they considered. In summary, the Harter and Pike interpretation of their results may be unduly pessimistic about the ability of very young children to differentiate among multiple dimensions of self-concept, and it may be premature to conclude that children at this age level can identify only two broad components of self.

Marsh et al. (1991) described a new, adaptive procedure for assessing multiple dimensions of self-concept for kindergarten, first- and second-year students aged 5–8. They explored pictorial self-concept instruments like that developed by Harter and Pike (1984) but found that the juxtaposition of the pictures and verbal explanations seemed more confusing to young students than the verbal presentations alone. The critical component of their study was the individualized interview format used to collect SDQ-I responses. There was an initial concern that the 64-item SDQ-I instrument would be too long for these very young children, but items near the end were more effective than earlier items. Apparently children learned to respond appropriately so that they were responding more appropriately to items at the end of the instrument than to items at the beginning of the instrument. This observation has important implications for the typically short instruments used with young children. Based on CFAs, Marsh et al. found support for all eight SDQ-I scales, including the Esteem scale at each year level. However, with increasing age the SDQ-I factors became more differentiated as inferred from the decreasing size of factor correlations. They also reported gender differences for this very young group of children that were largely consistent with extrapolations from earlier research with older children.

Based on an expectancy-value model, Eccles and her colleagues (Eccles, Adler, & Meece, 1984; Eccles et al., 1993; Wigfield et al., 1997) focused on how expectancy (self-perceptions of competency, expectations of future success, and self-efficacy) combine with task value to influence academic choice. Their competency-related beliefs in specific activities (how good they are, how good they are relative to others, how good they are relative to other activities, expectations of future success, ability to master new skills) were closely related to the self-concept construct. Eccles et al. (1993) provided further support for the multidimensionality of self-concept for young children, finding that children in grades 1, 2, and 4 differentiated math, reading, music, and sports self-concepts (self-perceived competency ratings). In exploratory factor analyses of responses to items from all four domains and both the competency and task value components, four domain-specific factors were clearly evident for each age group. The CFAs of competency and task value ratings conducted separately for each domain provided supported the separation of task and competency ratings, although no CFAs of competency ratings from different domains were reported. Consistent with research with older children (e.g., Eccles & Wigfield, 1995), there was no evidence that children distinguished between different competency-related beliefs (e.g., self-perceived skills and expectations for future success).

Wigfield et al. (1997), using a cohort-sequential longitudinal design (see Baltes & Nesselroade, 1979), assessed three age cohorts (students in grades 1, 2, and 4) in each of three successive years. This allowed them to compare age differences due to cohort differences with true longitudinal comparisons and to evaluate stability over time for responses by the same children. Consistent with previous research (e.g., Marsh, 1989) they found that perceived competency declined with age, and these effects were reasonably consistent over domain and across cohort and longitudinal comparisons. There was also a consistent pattern in test-retest stability coefficients in which stability over time was low for the youngest children and grew steadily with age, although these comparisons were complicated to some extent by age-related differences in reliability. Wigfield et al. reported predicted gender differences (favoring boys in math and sport, favoring girls in music and reading), but found few interactions with gender and either cohort or time of measurement. Instead, consistent with Marsh (1989) and Marsh et al. (1991), they found that gender differences emerged early and were reasonably consistent over age. Finally, teacher and parent ratings of competence collected at each time of measurement showed little systematic relation to self-ratings for the youngest children but were substantially related to self-perceptions for the oldest children. For example, only teacher ratings of sport were significantly related to self-perceptions in Years 1 and 2, whereas teacher ratings in all four domains were significantly related to self-perceptions in Years 5 and 6. Interestingly, in contrast to children's self-perceptions, teacher and parent ratings of competence did not vary as a function of age. Wigfield et al. concluded that very young children have more optimistic self-perceptions of their own competence whereas older children have more realistic self-perceptions.

In the last decade there were two major reviews of self-concept measures (Byrne, 1996; Wylie, 1989) for different age groups. Both reviewers noted a large number of different instruments available for very young children but concluded that only two were sufficiently developed to warrant consideration; the Harter and Pike (1984) instrument discussed earlier and the Joseph (1979) instrument that relies on items from a variety of different domains to infer a global, undifferentiated self-concept and so is of limited relevance to our concern about the multidimensional structure of self-concept. Both reviewers noted that there was a paucity of psychometric evidence (reliability, stability, factor structure) for either of these instruments and the evidence that was available was not particularly encouraging. Essentially no further development of either instrument was reported by Byrne beyond that described 7 years earlier by Wylie (1989). Although Byrne (1996) did not consider the Marsh et al. (1991) study in her chapter on measures for very young children, she did review the SDQ-I in another chapter devoted to measures for somewhat older, preadolescent children, concluding that (Byrne, 1996, p. 117) "there is absolutely no doubt that the SDQ-I is clearly the most validated self-concept instrument available for use with preadolescent children." Byrne specifically noted the effectiveness of the Marsh et al. (1991) study in adapting the SDQ-I for use with very young children, emphasizing that the psychometric properties based on this single study were stronger than those based on any of the instruments specifically designed for very young children.

Age and Gender Differences in Responses by Very Young Children

Age differences. In her classic review, Wylie (1979) reported that age differences in overall or total self-concept were small and inconsistent. In his subsequent review and empirical analyses 12,000 responses from the SDQ normative archives, Marsh (1989) reported that there was a reasonably consistent pattern of self-concepts declining from a young age through early adolescence, leveling out, and then increasing at least through early adulthood. These age differences were, however, small and differed somewhat depending on the particular scale. For example, Marsh et al. (1984) reported that the initially very high parents self-concept ratings remained high across the early preadolescent period but fell sharply during the adolescent period. More recently, Chapman and Tunmer (1995) also reported that reading self-concept declined with age based on a cross-sectional study of very young children as did the Marsh et al. (1991) and Wigfield et al. (1997) studies described earlier. Marsh and Craven (1997) argued that whereas young children have extremely high self-concepts, they develop more realistic appraisals of their relative strengths and weaknesses with age and this added experience is apparently incorporated into their self-concepts. Hence, with increasing life experience, self-concept in specific domains should become more differentiated, more accurately reflecting the child's relative strengths and weaknesses.

Crain (1996) recently reviewed development and age differences in self-concept but concluded that the typical poor quality of measurement instruments used in studies of very young children undermined extrapolations from that research. Crain, however, specifically highlighted the important contributions of the Marsh et al. (1991, p. 403) study, emphasizing that this research cast doubt on previous research for this age group but concluded that "certainly, further examination of very young children's multidimensional self-concept is a fruitful topic for further research." In particular, she highlighted the need for longitudinal research to elucidate changes in the structure, development, and stability of young children's self-concepts over time.

Whereas most research on age and development effects in self-concept has focused on mean differences, Shavelson et al. (1976) proposed age and developmental differences in the structure of self-concept. They hypothesized that self-concept becomes more differentiated with age but offered no clear rationale for evaluating this hypothesis. Marsh (1989; Marsh et al., 1991) operationalized this hypothesis to mean that correlations among SDQ scales become smaller with age and tested this hypothesis with responses to the three SDQ instruments. For responses by very young children, correlations decreased in size from kindergarten to Year 1, and from Year 1 to Year 2. For preadolescent responses, correlations decreased from Year 2 to Year 3, and to lesser extents between Years 3 and 4, and between Years 4 and 5. For responses by older preadolescents, adolescents, and late adolescents, no further declines in the correlations were found. Thus, these data support the Shavelson et al. hypothesis that self-concept becomes differentiated with age for very young children, but not for responses by older children. Perhaps differentiation in self-concept reaches some optimal point at which social comparison processes and cognitive abilities are adequately developed rather than after a cumulative period of life experiences. However, as in most studies of self-concept development, researchers relied on cross-sectional age comparisons rather than stronger tests of differences over time for the same cohort or more sophisticated designs that allow simultaneous cross-sectional and longitudinal comparisons.

Gender differences. Historically, studies of gender differences in self-concept have focused primarily on global or total scores. Early reviews (e.g., Maccoby & Jacklin, 1974; Wylie, 1979) reported little or no gender differences, but Feingold (1994) compared results from three meta-analyses of gender differences in personality variables including self-esteem that each demonstrated small differences (effect sizes of .10 to .16) favoring males.

Consistent with Wylie's (1979) suggestion, Marsh (1989) reported that these reasonably small differences in total scores reflect larger, counterbalancing gender differences in specific components of self-concept. Based on normative archive responses to the three SDQ instruments (covering the preadolescent to young adult age range), he reported statistically significant but small gender differences in most SDQ scales, some favoring girls but more favoring boys. Total self-concept scores favored boys, although gender explained only 1% of the variance. The gender differences in specific scales tended to be consistent with traditional gender stereotypes: (1) boys had higher self-concepts for Physical Ability, Appearance, Math, Emotional Stability, Problem Solving, and Esteem; (2) girls tended to have higher self-concepts for Verbal/Reading, School, Honesty/Trustworthiness, and Religion/Spiritual Values; and (3) there were no gender differences for the Parents scale in any of the three data sets. Gender differences in the social scales, however, were not fully consistent with traditional gender stereotypes favoring girls. These gender differences in self-concept are broadly consistent with gender stereotypes, but the small effects suggest, perhaps, that this influence on self-concept is diminishing. Marsh (1989) also found that age × gender interactions were typically small, except for Appearance (very young girls had higher Appearance self-concepts than boys, but older girls had much lower Appearance self-concepts). Apparently, gender stereotypes have already affected self-concepts by preadolescence, and these effects are relatively stable from preadolescence to at least early adulthood. Subsequent research with even younger children (Eccles et al., 1993; Marsh et al., 1991) further supports the suggestion that gender stereotypes affect self-concepts of children at very young ages. In a recent review of gender differences in self-concept, Crain (1996, p. 412) also concluded that there are "differences in domain-specific self-concepts of boys and girls that tend to run along gender-stereotypic lines." However, she argued that the gender differences were typically not of sufficient size to be of substantive significance and that societal changes in the role of women may alter the pattern of gender differences in multiple dimensions of self-concept.

Inferred Self-Concept Ratings

Self-concept ratings inferred by others are used to determine how accurately self-concept can be assessed by external observers and to validate self-concept responses. For teachers, being able to infer self-concept accurately is particularly important for understanding and responding to their students. When multiple dimensions of self-concept are represented by both self-ratings and inferred ratings, multitrait-multimethod (MTMM) analyses (Campbell & Fiske, 1959; Marsh & Grayson, 1995) provide tests of the construct validity of the responses. *Convergent validity* is inferred from substantial correlations between self-ratings and inferred-ratings on matching self-concept traits. *Discriminant validity* provides a test of the distinctiveness of self-other agreement and of the multidimensionality of the self-concept facets; it is inferred from the lack of correlation between nonmatching traits.

In eight MTMM studies, Marsh (1988, 1990) demonstrated significant agreement between multiple self-concepts inferred by primary school teachers and student responses to the SDQ-I. The mean correlation between student and teacher ratings across all scales was r = .30, but agreement was strongest where the teachers could most readily make relevant observations (math, .37; reading, .37; school, .33; physical ability, .38; and, perhaps, peer relations .29). Student-teacher agreement was lower on Relations with Parents (.17) and Physical Appearance (.16). Marsh and Craven (1991) extended this research in a comparison of the abilities of elementary school teachers, mothers, and fathers to infer multiple self-concepts of preadolescent children. Responses by mothers and by fathers were slightly more accurate than those by teachers, but all three groups were more accurate in their inferences about Physical Ability, Reading, Mathematics, and School self-concepts than other specific scales or Esteem. Self-other agreement in this study tended to be better than has been found in other research, but this apparently reflects the fact that children and significant others all completed the complete SDQ-I instrument (in contrast to studies described earlier in which teachers infer self-concepts of all students in their class using a single-item summary rating of each SDQ-I scale whereas students complete multi-item scales). Nevertheless, the self-other agreement reported by Marsh and Craven was still much less than values reported in studies with older participants. For example, in MTMM studies of university students and "person in the world who knew them best" (Marsh, Barnes, & Hocevar, 1985; Marsh & Byrne, 1993), self-other agreement was very high (mean r = .57), and four of the scales had self-other correlations over .75. The authors speculated that self-other agreement was so good because (1) the participants were older and thus knew themselves better and based their self-responses on more objective, observable criteria; (2) both participants and significant others made their responses on the same well-developed instrument based on multi-item scales; and (3) the significant others in these studies knew the participants better than the observers in most research. These results imply that external observers are best able to infer self-concepts when respondents are older and the multidimensionality of self-concept is taken into account.

The juxtaposition of the self-other agreement studies for different age groups also provides at least indirect support for the proposal that with increasing age and life experience children learn their relative strengths and weaknesses so that their self-concepts became more differentiated and more highly correlated with external indicators, including self-concepts inferred by significant others (also see Wigfield et al., 1997). A fuller evaluation of this proposal, however, requires CFA models based on longitudinal data for very young children. For example, particularly strong support for this proposal would be the demonstration that ratings by an external observer at T1 contribute directly to the prediction of self-concept ratings of the child at T2 beyond what can be predicted by the child's self-concept ratings at T1. In this sense, ratings by the external observer are able to predict changes in the child's self-concept over time because the child's self-concept is likely to change in the direction of being more consistent with external criteria used by the observer at T1 (i.e., becoming more "realistic").

The Present Investigation

Marsh et al. (1991) provided a promising advance in the measurement of very young children's self-concepts and in clarifying the emergence and progressive differentiation of specific facets of self-concept for this age group. Due in part to limitations in self-concept research with very young children, reviewers (e.g., Byrne, 1996; Crain, 1996) noted the important contributions of this study but also emphasized the need to follow up this research with longitudinal studies more appropriate for evaluating the development of self-concept. In response to such concerns, the present investigation is based on the Marsh et al. (1991) study but expands on the empirical and theoretical implications of that earlier research in a number of ways. In particular, the earlier study was based on a single wave of data from three age cohorts so that developmental implications relied primarily on cross-sectional comparisons. Here we used a multicohort-multioccasion (MCMO) design with two waves of data collected 1 year apart with the same children in each of three age cohorts. Based on these data, we contrasted cross-sectional (multiple age cohort) comparisons with true longitudinal (multiple occasion) comparisons. This provides a much stronger basis for evaluating age-related differences in reliability, dimensionality, and gender differences that were the focus of the earlier research, and for evaluating stability over time. In addition, self-concept ratings inferred by teachers are included in the present investigation. This additional source of information allows the evaluation of the accuracy of teacher's inferred self-concept ratings, an examination of how relations between inferred and actual ratings vary with age cross-sectionally and longitudinally, and an evaluation of the proposal that young children's self-concept becomes more predictable with age. More specifically, in addition to building on the earlier research by providing further psychometric support for the use of the individually administered SDQ-I with young children, the present investigation is designed to (1) test the Shavelson et al. (1976) hypothesis that self-concept becomes more differentiated with age and to provide more specific data on how the factor structure of self-concept varies over time for children aged 5–8, (2) evaluate the stability of young children's self-concepts over time, (3) evaluate gender and age differences for young children longitudinally as well as cross-sectionally, and (4) evaluate how teacher ratings of the self-concepts of their students relate to students' own self-concept ratings and how these relations vary with age. From a practical perspective the ability to measure the self-concepts of young children and elucidate developments in self-concept over time enables early childhood practitioners to understand young children better, to identify an accurate basis for assessment, and to provide an outcome measure for a variety of interventions. Hence this research has the potential to advance self-concept theory, research, and practice.

METHOD

Sample

The sample considered in this study is a total of 396 students who at T1 were enrolled in kindergarten (n = 127, M age = 5.4, SD = .4), Year 1 (n = 139, M age = 6.3, SD = .4), and Year 2 (n = 130, M age = 7.4, SD = .5). The participants came primarily from middle-class families and attended one of three schools in suburban metropolitan Sydney, Australia, that agreed to participate in the study. Although these students were a year older and enrolled in Years 1, 2, and 3 at T2, we refer to the age cohorts as kindergarten, Year 1, and Year 2. Because of the focus of this study on longitudinal comparisons, responses from 98 children who had data at T1 but not T2 were excluded. This attrition was due to normal absences on the day the materials were administered and the typical mobility of families in this region of metropolitan Sydney.

Instrument: The SDQ-I

The SDQ-I (Marsh, 1988, 1990) is based on the Shavelson et al. (1976) model and is designed for preadolescents (SDQ-I). Research (see Marsh, 1988, 1990, 1993a; Marsh & Craven, 1997) has shown that (1) factor analyses have consistently identified each a priori SDQ-I factor; (2) the reliability of each scale is generally in the .80s and .90s whereas correlations among the factors are quite small (median *r*s less than .20); (3) the self-concept responses are consistently correlated with external validity criteria (e.g., self-concepts in matching areas inferred by significant others, academic achievement indicators, age, gender, locus of control, self-attributions for the causes of academic successes and failures, physical fitness and participation in sports, and self-concept enhancement interventions). This research provides strong support for the construct validity of responses to the SDQ instruments for children as young as 10 or, perhaps, 8 (see reviews by Byrne, 1996; Hattie, 1992; Wylie, 1989).

The SDQ-I (Marsh, 1988) is designed to measure eight self-concept scales that are summarized in the Appendix. Three total scores can also be formed on the basis of these scales: academic self-concept (the average of reading, mathematics, and school self-concept scales), nonacademic self-concept (the average of physical, appearance, peer, and parent relations self-concept scales), and total self (the average of academic and nonacademic total scales). Each of the eight SDQ-I scales was defined by responses to eight positively worded items. On the standard SDQ-I there are an additional 12 negatively worded items. Because previous research has shown that children have trouble responding appropriately to the negatively worded items, they are not included in the scores derived from the SDQ-I (Marsh, 1988). For purposes of the individually administered SDQ-I used here, the negatively worded items were excluded altogether. As described below, the response scale typically used with the SDQ-I was also altered for purposes of just the individually administered responses.

As part of the study, teachers at T1 rated each of their students only on the eight SDQ-I scales, based on a single summary item representing each scale instead of the multiple items representing each scale actually completed by students. Teachers were given a single page of instructions containing a brief definition of each SDQ-I scale, a list of all the students in their class with eight columns next to each student's name corresponding to the eight SDQ-I scales, and instructions about how to complete the survey. Teachers were instructed to make judgments of each child's self-concept based on the student's own feeling about himself or herself (i.e., to infer their students' self-concepts) using a 9 point (1 = poor to 9 = high) response scale.

Procedures

Procedures for the administration of the standard SDQ-I (see Marsh, 1988) were adjusted to enable the modified SDQ-I to be administered as an individual interview and are described in greater detail by Marsh et al. (1991). The interviewers were 110 university students in a primary teacher education program who already had experience working with young children. All interviewers were given a 2-hour training program and subsequently tested children from each of the three age groups as part of a class assignment. At each participating school a group of interviewers simultaneously conducted interviews with all students from a particular class. The testing was conducted using an individual, one-on-one interview style format. Each testing session began with a brief set of instructions assuring participants of the confidentiality of their responses and presenting four example items. After reading each example item the interviewer asked the child if he or she understood the sentence. If the child did not understand the sentence, the interviewer explained the sentence further, paraphrasing any words the child did not understand, ascertained that the child understood the sentence, re-read the sentence, and requested a response. The interviewer initially asked the child to respond "yes" or "no" to the sentence to indicate whether the sentence was true or false as a description of the child. If the child initially responded "yes," the interviewer then asked the child if he or she meant "yes always" or "yes sometimes." If the child initially responded "no," the interviewer then asked the child if he or she meant "no always" or "no sometimes." The second response probe was stated for every response even when it was answered in the initial response (e.g., the child said "yes always" instead of "yes"), thus providing a check on the accuracy of the child's initial response. After the child successfully responded to example items and any questions were answered, the interviewers then read aloud each of the 64 positively worded SDQ-I items. The child was encouraged to seek clarification of any item they did not understand. If the child stated that the item was not understood, the interviewer explained the meaning of the item further and ascertained that the child understood the sentence before readministering the item. If the child indicated that he or she understood the sentence but could not decide whether to respond yes or no, the interviewer recorded a response of 3, halfway between the responses of "no sometimes" and "yes sometimes." Because this occurred infrequently and children were not told of this option, this middle category was used very infrequently. Halfway through the administration of the SDQ-I items the interviewer asked the child to do some physical activities for a brief period before proceeding to administer the remaining 32 items. This procedure was intended to cater to young children with short attention spans.

Statistical Analyses

The statistical analyses consisted of an evaluation of the psychometric properties (internal consistency reliability, stability over time, and factor structure) of the self-concept responses, of gender and age differences in the self-concept ratings, and of relations between self-concept ratings and self-concept inferred by teachers. All analyses were conducted with SPSS (Norusis, 1993), including the CFAs that were conducted with the SPSS version of LISREL (Jöreskog & Sörbom, 1989).

As in other SDQ research (e.g., Marsh, 1988; Marsh & Hocevar, 1985) factor analyses were conducted on item-pair scores (or parcels) in which the first two items in each scale are averaged to form the first item pair, the next two items are used to form the second pair, and so forth. Analysis of 32 item pairs instead of 64 individual items is advantageous because this strategy substantially reduces the number of measured variables and because the responses to item pairs tend to be more reliable, to be more normally distributed, and to have less idiosyncratic variance than do individual items. The CFAs were conducted with the SPSS version of LISREL 7 (Jöreskog & Sörbom, 1989) using maximum likelihood estimates derived from covariance matrices based on pairwise deletion for missing data. A detailed description of CFA is beyond the scope of the present investigation and is available elsewhere (e.g., Bollen, 1989; Byrne, 1989; Jöreskog & Sörbom, 1989; Pedhazur & Schmelkin, 1991). Following Marsh, Balla, and Hau (1996) and McDonald and Marsh (1990), we emphasize the relative noncentrality index (RNI) to evaluate goodness of fit but also present the chi-square test statistic and *df* that allow the calculation of other indexes of fit. Whereas there are no precise standards for what values of indices such as these are needed for an "acceptable" fit, typical guidelines are that the RNI should be greater than .9. However, model comparison is also facilitated by positing a partially nested ordering of models in which the parameter estimates for a more restrictive model are a proper subset of those in a more general model (for further discussion, see Bentler, 1990). In the present application, for example, a model in which factor loadings are constrained to be invariant across the three age cohorts is nested under a model in which there are no such invariance constraints of the factor loadings. The fit indices for alternative models, however, can be compared whether or not the particular models are nested. Whereas tests of statistical significance and indices of fit aid in the evaluation of the fit of a model, there is ultimately a degree of subjectivity and professional judgment in the selection of a "best" model.

Although a variety of different CFAs is considered in the results, the SDQ-I self-concept factors are always inferred from multiple indicators of the latent construct. For gender and for teacher-inferred self-concept ratings for each SDQ-I scale, however, there is only a single indicator of each latent construct. In each case, single-indicator latent variables were assumed to be measured without error. Whereas this strategy is reasonable for gender, there is likely to be error in the teacher ratings. Whereas it is reasonable to incorporate a plausible estimate of measurement error into the analysis that would automatically increase correlations between teacher ratings and other constructs, the present strategy is conservative in relation to showing that teacher ratings are related to student responses.

In most applications of CFA, a priori models typically assume that the residual variance (uniqueness plus random error, hereafter referred to as uniquenesses) associated with each measured variable is independent of residual variances associated with other measured variables. However, when the same items are administered to the same participants on multiple occasions, it is likely that the uniquenesses associated with the matching measured variables are correlated. If there are substantial correlated uniquenesses that are not included in the model, then the estimated correlations between the corresponding latent constructs will be positively biased. In the present application, for example, this would result in a positively biased estimate of the test-retest stability coefficient relating responses to the same latent variable on two occasions. This situation is not specific to CFA studies, and Marsh (1993b) demonstrated that stability coefficients can exceed 1.0 when disattenuated for measurement error. However, the inclusion of correlated uniquenesses in CFAs provides a test for these correlated uniquenesses and a control for what would otherwise be a positive bias. Because of this problem, test-retest stability correlations are likely to be negatively biased because they do not take into account measurement error and, perhaps, positively biased because they do not control for correlated uniquenesses. This complexity is likely to be compounded when, as in the present investigation, comparisons are made between different age cohorts where the size of measurement errors and, perhaps, the correlated uniquenesses are likely to vary with age. Because of these complications, a potentially important contribution of the present investigation is to evaluate age cohort differences in test-retest stability coefficients with latent variable models that incorporate appropriate control for measurement error and test for correlated uniquenesses.

In preliminary analyses, the inclusion of these correlated uniquenesses was supported by modestly better fits to the data and, in particular, because their exclusion would positively bias the test-retest stability coefficients. Whereas stability coefficients were smaller when correlated uniquenesses were included (Table 4), the differences were typically small. Across all 32 sets of analyses (eight SDQ-I scales for total, kindergarten, Year 1, and Year 2), the difference between the two stability correlation estimates never exceeded .06 and typically was much smaller, suggesting that the inclusion of correlated uniquenesses in this study was not a critical issue. To facilitate the substantive import of the results, only the models with correlated uniquenesses are presented along with the more conservative estimates of test-retest correlations based on these models.

RESULTS

Internal Consistency, Stability, and Distinctiveness

We begin with a preliminary overview of psychometric results based on traditional approaches that also serve as an advance organizer to more sophisticated CFA approaches. Because our major thrust is on the CFA results, these preliminary results are summarized only briefly.

Internal consistency. Coefficient α estimates of reliability tended to increase with age, based on both cross-sectional and longitudinal comparisons (Table 1). For example, the median α varied from .74 (T1 K) to .85 (T2 Year 2) for the three age cohorts (Table 1). This pattern was evident in the mean and median αs for the cross-sectional, between-group comparison of different age cohorts and particularly for T1 to T2 longitudinal comparisons within each age cohort and for the total sample. Thus, for example, median α was lowest for kindergarten responses (.74 for T1, .78 for T2), followed by Year 1 (.80 for T1, .79 for T2), and largest for the Year 2 responses (.82 for T1, .85 for T2). This pattern is also reasonably consistent for the individual scales, although αs for T1 and T2 responses at Year 2 do not differ systematically. Interestingly, a major exception to this pattern was for Esteem, where reliability estimates did not seem to be consistently related to either cross-sectional or longitudinal age differences.

In contrast to the scale scores, the total (total academic, total nonacademic, and total self) scores were more reliable (αs from .83 to .95) and these αs did not vary systematically as a function of cross-sectional or longitudinal age comparisons. The total scores were more reliable than the individual scores because they were based on so many more responses. However, the lack of consistent age differences in αs for total scores was in marked contrast to the individual scores (except for the Esteem scale, which may be more like the total scores in that its intent was to infer an overall evaluation of self). This pattern of results, increasing internal consistency estimates of reliability with age for specific self-concept scales but not for global and total scores, implied that older children were more clearly differentiating among the specific components of self-concept. Because this issue was of central concern to the present investigation, it was the focus of subsequent analyses.

Stability over time. As expected, test-retest stability correlations based on scale scores were all statistically significant and tended to increase with age (see correlations between T1 and T2 measures, labeled T12 in Table 1). Across the eight scales, the stability coefficients for the oldest, Year 2 students (mean r = .47) were higher than the coefficient based on the total sample (mean r = .37). However, stability coefficients for the kindergarten (mean r = .32) and Year 1 (mean r = .32) samples were similar.

Discriminant validity and distinctiveness. In Table 2 all correlations between T1 and T2 scale scores are presented for each year group to evaluate the discriminant validity of the self-concept responses. Adapting the terminology of MTMM analyses (Campbell & Fiske, 1959; Marsh & Grayson, 1995), the stability coefficients (coefficients in bold in the main diagonal of Table 2) were viewed as convergent validities, and the different occasions were taken to be the multiple methods. From this perspective, one indication of discriminant validity was to compare each stability coefficient (convergent validities) with the other 14 coefficients in the same row and column for that age cohort. Across the eight scales there were a total of 8 × 14 = 112 comparisons each for the kindergarten, Year 1, Year 2, and total samples. The stability coefficients were larger than the comparison coefficient for all 112 comparisons based on the total sample, and for 109, 109, and 91 of 112 comparisons based on responses for children from Year 2, Year 1, and kindergarten, respectively. For each year group considered separately, there was at least one failure (i.e., one comparison coefficient that was larger than the convergent validity) for one scale (Esteem) for Year 2 responses, two scales (Esteem, Peers) for Year 1 responses, and four scales (Esteem, Peers, Reading,

Table 1 Coefficient Alpha Estimates of Reliability at Time 1 (al) and Time 2 (al), T1/T2 Stability Correlation, Correlations among All and Selected Scales, and SDs of the Scale Scores for Each Grade Level and the Total Sample

	Kind (n = 127)			Year 1 (n = 139)			Year 2 (n = 130)			Total (n = 396)		
Scores	T1	T2	T12	T1	T2	T12	T1	T2	T12	T1	T2	T12
Reliability and stability:												
Total scores:												
Nonacademic	.83	.88	.43	.88	.88	.35	.86	.89	.54	.86	.88	.44
Academic	.89	.91	.36	.92	.89	.29	.91	.91	.54	.91	.91	.40
Total	.93	.94	.45	.95	.93	.29	.94	.93	.57	.94	.93	.43
Scale scores:												
Physical	.52	.70	.49	.71	.72	.43	.69	.78	.49	.65	.73	.46
Appearance	.74	.82	.36	.83	.87	.40	.86	.88	.43	.82	.86	.41
Peers	.77	.75	.27	.79	.77	.27	.80	.84	.52	.78	.79	.36
Parents	.66	.71	.32	.72	.69	.39	.69	.74	.50	.70	.72	.39
Read	.78	.85	.26	.85	.80	.24	.83	.88	.59	.82	.85	.37
Math	.78	.81	.20	.85	.83	.28	.85	.90	.45	.83	.85	.32
School	.70	.81	.39	.82	.85	.28	.84	.85	.46	.79	.84	.38
Esteem	.75	.73	.27	.79	.71	.25	.74	.75	.33	.76	.72	.28
Median scale score	.74	.78	.29	.80	.79	.28	.82	.85	.48	.79	.82	.38
Mean scale score	.71	.77	.32	.80	.78	.32	.79	.83	.47	.77	.80	.37
Distinctiveness of scales:												
Mean of all *r*s	.45	.45	.22	.50	.40	.13	.39	.31	.21	.44	.38	.18
Mean of seven select *r*s	.45	.36	.23	.48	.36	.14	.30	.19	.13	.41	.30	.16
SD of eight SDQ scales	.38	.43		.42	.45		.49	.55		.43	.48	

Note: Presented under each column are the coefficient alpha estimates of reliability at Time 1 (T1) and Time 2 (T2), and the test-retest correlation between matching scales at T1 and T2 (T12). Under the heading "distinctiveness," the mean of all *r*s refers to the mean of all off-diagonal correlations in the 8 × 8 matrix of correlations at T1, T2, or T12 (i.e., nonmatching correlations between T1 and T2 responses). Mean of seven selected *r*s refers to the mean of seven correlations predicted a priori to be smallest. The *SD*s of the eight SDQ scales reflect the extent to which children give the similar (lower *SD*s) or different (higher *SD*s) mean responses to the different SDQ scales.

Math) for kindergarten responses. Overall, these results provide strong support for the discriminant validity of the self-concept responses but also support the prediction that discriminant validity increases with age.

Marsh (1989) proposed an alternative approach to evaluating how the distinctiveness of the self-concept traits vary with age. He argued that some of the scales (e.g., Reading and School) should be substantially correlated whereas others (e.g., Physical and Reading) should not. Based on previous research and theory, he selected seven correlations that he predicted to be the lowest. Based on the assumption that self-concept becomes more differentiated with age, he reasoned that the difference in the mean of the selected correlations and mean of all correlations should increase with age. Here we extended this logic to evaluations of longitudinal differences in correlations for the same age cohort on different occasions as well as cross-sectional differences between age cohorts like those considered by Marsh (1989). At T1, the mean of the selected *r*s compared to the mean of all *r*s (see Table 1) did not differ for kindergarten students, was slightly lower for Year 1 students, and was clearly lower for Year 2 students. Although both the means of all *r*s and selected *r*s tended to decrease over time, the decrease was larger for the mean of *r*s selected a priori to be lower. For each age cohort, the difference between the mean of the selected *r*s and the mean of all *r*s was larger at T2 than T1. Hence, these comparisons based on the cross-sectional and particularly the longitudinal comparisons supported the hypothesis that self-concept becomes more differentiated with age.

Alternatively, distinctiveness can be operationalized as the extent to which children give the same or similar mean responses to all scales (lower distinctiveness) or give different mean responses to different scales (higher distinctiveness). Following Marsh (1989), this was operationalized as the *SD* of the eight SDQ scale scores, computed separately for each of the 3 (age cohort) × 2 (time) combinations. The results (Table 1) provided a clear pattern in which responses became more differentiated (larger *SD*s) with age based on both cross-sectional age-cohort comparisons and longitudinal comparisons. These results provided clear support for the proposal that SDQ responses become more differentiated with age.

Table 2 Correlations between Time 1 and Time 2 Responses for Kindergarten (K), Year 1, and Year 2 Students and for the Total Sample

Variables	Physical	Appearance	Peers	Parents	Reading	Math	School	General
Physical:								
K	**.49***	.10	.13	.05	.23*	.20*	.26*	.14
1	**.43***	.09	.11	.11	.14	.20*	.23*	.11
2	**.49***	.13	.21*	.21*	.14	.10	.16	.14
T	**.46***	.12*	.15*	.11*	.16*	.17*	.22*	.13*
Appearance:								
K	.24*	**.36***	.11	.26*	.31*	.16	.23*	.27*
1	.17*	**.40***	.21*	.10	.13	.02	.08	.21*
2	.16	**.43***	.38*	.27*	.12	.12	.21*	.27*
T	.18*	**.41***	.24*	.18*	.17*	.10	.18*	.24*
Peers:								
K	.30*	.21*	**.27***	.25*	.23*	.20*	.24*	.33*
1	.10	.00	**.27***	.30*	.02	.06	.06	.17*
2	.11	.07	**.52***	.23*	.21*	.10	.11	.19*
T	.16*	.09	**.36***	.25*	.15*	.12*	.13*	.23*
Parents:								
K	.28*	.24*	.23*	**.32***	.31*	.28*	.29*	.31*
1	.09	.08	.22*	**.39***	.11	.10	.14	.21*
2	.01	.18*	.31*	**.50***	.37*	.25*	.38*	.18*
T	.13*	.15*	.25*	**.39***	.27*	.21*	.26*	.24*
Reading:								
K	.12	.29*	.08	.12	**.26***	.14	.29*	.30*
1	.12	.14	.17	.09	**.24***	.20*	.16	.14
2	.07	−.06	.24*	.20*	**.59***	.10	.24*	.09
T	.09	.11*	.17*	.10*	**.37***	.14*	.24*	.17*
Math:								
K	.25*	.20*	.11	.18*	.33*	**.20***	.33*	.21*
1	.30*	.09	.03	−.01	.08	**.28***	.21*	.05
2	.13	.17*	.28*	.23*	.22*	**.45***	.27*	.10
T	.22*	.16*	.15*	.11*	.21*	**.32***	.27*	.12*
School:								
K	.24*	.33*	.16	.32*	.34*	.20*	**.39***	.35*
1	.18*	−.01	.09	.05	.09	.20*	**.28***	.14
2	.15	.17	.29*	.32*	.46*	.24*	**.46***	.23*
T	.18*	.17*	.19*	.19*	.29*	.21*	**.38***	.24*
General:								
K	.28*	.18*	.16	.24*	.31*	.11	.25*	**.27***
1	.25*	.14	.16	.26*	.14	.13	.18*	**.25***
2	.15	.20*	.47*	.42*	.34*	.21*	.26*	**.33***
T	.23*	.17*	.25*	.28*	.26*	.15*	.23*	**.28***

Note: Stability coefficients, correlations between T1 and T2 responses to the same SDQI scale, are in bold.
* $p < .05$.

Factor Structure

Confirmatory factor analysis provides a particularly powerful tool for evaluating the factor structure underlying responses by these young children. Results from separate CFAs conducted for responses at T1 and T2 were summarized in Table 3. For each CFA, a very restrictive a priori model was posited in which each measured variable was allowed to load on only one factor and uniquenesses associated each variable were assumed to be uncorrelated. The factor solutions were well defined in that both solutions were fully proper, goodness of fit was reasonable (RNI = .916 and .901 for T1 and T2, respectively), and all factor loadings were statistically significant and substantial (varying from .45 to .85). Although the factor correlations were also substantial, varying from .25 to .81, none approached 1.0 even though they have been corrected for measurement error. Of particular interest was the comparison of the parameter values for the T1 and T2 solutions. Factor loadings tended to be larger for T2 than T1 (22 were larger, 8 were smaller, and 2 were the same in Table 3) whereas factor correlations tended to be smaller at

Table 3 Separate Factor Solutions for Responses at Time 1 (T1) and Time 2 (T2)

	Physical		Appearance		Peers		Parents		Reading		Math		School		General		Uniquenesses	
Scale	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
Factor loadings:																		
Physical:																		
1	.48	.58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.76	.67
2	.46	.68	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.79	.54
3	.65	.71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.58	.49
4	.65	.62	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.58	.61
Appearance:																		
1	0	0	.60	.70	0	0	0	0	0	0	0	0	0	0	0	0	.64	.52
2	0	0	.67	.79	0	0	0	0	0	0	0	0	0	0	0	0	.56	.37
3	0	0	.74	.79	0	0	0	0	0	0	0	0	0	0	0	0	.45	.38
4	0	0	.74	.71	0	0	0	0	0	0	0	0	0	0	0	0	.46	.50
Peers:																		
1	0	0	0	0	.62	.67	0	0	0	0	0	0	0	0	0	0	.62	.55
2	0	0	0	0	.71	.70	0	0	0	0	0	0	0	0	0	0	.50	.51
3	0	0	0	0	.68	.66	0	0	0	0	0	0	0	0	0	0	.54	.57
4	0	0	0	0	.73	.66	0	0	0	0	0	0	0	0	0	0	.46	.57
Parents:																		
1	0	0	0	0	0	0	.53	.57	0	0	0	0	0	0	0	0	.71	.68
2	0	0	0	0	0	0	.43	.42	0	0	0	0	0	0	0	0	.82	.83
3	0	0	0	0	0	0	.70	.71	0	0	0	0	0	0	0	0	.51	.50
4	0	0	0	0	0	0	.72	.80	0	0	0	0	0	0	0	0	.49	.36
Reading:																		
1	0	0	0	0	0	0	0	0	.72	.74	0	0	0	0	0	0	.49	.45
2	0	0	0	0	0	0	0	0	.76	.85	0	0	0	0	0	0	.42	.28
3	0	0	0	0	0	0	0	0	.78	.82	0	0	0	0	0	0	.40	.33
4	0	0	0	0	0	0	0	0	.80	.79	0	0	0	0	0	0	.37	.38
Math:																		
1	0	0	0	0	0	0	0	0	0	0	.58	.65	0	0	0	0	.66	.58
2	0	0	0	0	0	0	0	0	0	0	.70	.74	0	0	0	0	.51	.45
3	0	0	0	0	0	0	0	0	0	0	.83	.84	0	0	0	0	.32	.29
4	0	0	0	0	0	0	0	0	0	0	.79	.82	0	0	0	0	.38	.32
School:																		
1	0	0	0	0	0	0	0	0	0	0	0	0	.58	.72	0	0	.66	.48
2	0	0	0	0	0	0	0	0	0	0	0	0	.65	.71	0	0	.57	.50
3	0	0	0	0	0	0	0	0	0	0	0	0	.70	.70	0	0	.51	.52
4	0	0	0	0	0	0	0	0	0	0	0	0	.78	.76	0	0	.39	.42
General:																		
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.45	.55	.70	.80
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.62	.65	.58	.61
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.73	.74	.45	.47
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	.70	.70	.51	.51
Factor correlations:																		
Physical																		
Appearance	.41	.41																
Peers	.55	.48	.59	.55														
Parents	.40	.35	.46	.47	.53	.58												
Reading	.52	.27	.41	.25	.58	.40	.49	.33										
Math	.52	.48	.47	.36	.58	.32	.40	.34	.57	.38								
School	.57	.41	.61	.40	.69	.51	.56	.39	.75	.61	.78	.68						
General	.57	.58	.75	.69	.80	.80	.65	.66	.63	.47	.58	.55	.81	.68				

Note: Separate factor analyses of responses for T1 and T2 are summarized in completely standardized form. All parameters with the value of 0 were fixed a priori and not estimated as part of the analysis.

T2 than T1 (22 were smaller, 4 were larger, and 2 were the same in Table 3). These higher factor loadings were consistent with earlier findings that T2 responses were more reliable, whereas the lower factor correlations were consistent with earlier findings that T2 scales were better differentiated.

Marsh et al. (1991) were also concerned with a possible fatigue effect in asking very young children to respond to so many self-concept items. They found, however, that factor loadings tended to be larger for items near the end of the SDQ than those near the start. They interpreted this to mean that there was a warmup effect, whereby young children learned to respond appropriately, rather than a fatigue effect. There was support for this conclusion based on both T1 and T2 factor solutions (Table 3). The factor loadings were larger for the last measured variable than the first measured variable for all eight factors in the T1 solution and for seven of eight factors in the T2 solution. For both T1 and T2 solutions the median factor loadings increased steadily for indicators 1 (.58 and .66 for T1 and T2, respectively), 2 (.66 and .70), 3 (.70 and .72), and 4 (.74 and .74). These results suggest a substantial warmup effect that was particularly evident at T1 but that was still evident at T2. Whereas the factor loadings were systematically larger for the T2 solution, the difference was primarily due to the higher loadings for measured variables from the first half of the SDQI.

Summarized in Table 4 are a series of four models fit separately for each SDQ scale. Separate one-factor congeneric factor models were used to assess the unidimensionality of responses for each scale at T1 and T2, respectively (those labeled T1 and T2). For the total sample, these one-factor models provided a very good fit for T1 and T2 responses (RNIs vary from .96 to 1.0). The RNIs were also high for analyses of each age cohort considered separately, with the possible exception of kindergarten responses to the Physical and Appearance scales (RNIs of .86 and .84, respectively).

Also summarized in Table 4 are two-factor models fit to the T1 and T2 responses for each scale (those labeled T1/T2 in Table 4). These models were used to assess whether two factors—one for T1 responses and one for T2 responses—adequately fit the data and to define an optimal estimate of the correlation between the two factors—the T1/T2 stability coefficient. These stability coefficients in Table 4 are all substantially larger than those based on scale scores considered earlier (Table 1). This follows because the correlations in Table 4 were corrected for measurement error (and the inclusion of correlated uniquenesses had only a small effect). However, the pattern of differences observed in Table 1 was evident in Table 4 as well. Thus, for example, stability coefficients were reasonably similar for kindergarten (median r = .32) and Year 1 (median r = .34) students, but those for Year 2 were substantially larger (median r = .55).

Multiple Group Comparisons: Invariance over Age Cohorts

In analyses summarized in this section, structural equation path models relating gender and teacher inferred ratings (of their students' self-concepts) to T1 and T2 self-concept ratings were evaluated for each SDQ-I scale. Critical issues were the influence of gender and teacher ratings (collected at T1) on T1 self-concept responses and whether these variables had an additional direct effect on T2 self-concept ratings beyond the effects that were mediated by T1 self-concept. If gender influenced T2 self-concept directly, then there would be evidence that the gender differences were changing with age. If T1 teacher ratings directly affected T2 self-concept, then there would be support for the proposal that students' self-concepts became more predictable with age.

In CFA studies with multiple groups, it is possible to test the invariance (equality) of any one, any set, or all parameter estimates across the multiple groups. Here we evaluated the invariance of various sets of parameters across the multiple age groups (Table 5). In the least restrictive Model 1 (Table 6), no parameters were constrained to be equal across the three age cohorts, and this model provided a good fit for each of the SDQ scales. In Model 2 the factor loadings relating the T1 and T2 self-concept ratings to their latent construct were constrained to be invariant across the three age cohorts. Although this model resulted in a significantly poorer fit in a strict statistical sense for a few scales, all the RNIs were .92 or greater. In Model 3 all parameter estimates were constrained to be invariant across the three age cohorts, and this model resulted in significantly poorer fits for all of the SDQ scales. In Model 4, the invariance constraints on the uniquenesses were relaxed so that the uniquenesses were estimated separately for each age cohort. This resulted in a substantially improved fit relative to the model with all parameters constrained to be invariant. Whereas the fit of Model 4 was statistically poorer than that of Model 1 (with no invariance constraints) for several of the scales, all of the RNIs were .94 or greater.

In Models 5–7, the invariance of selected structural parameters of particular interest were evaluated. In each model, the uniquenesses were independently estimated in each age cohort (as in Model 4) along

Table 4 Psychometric Properties (Goodness of Fit, Reliability, Test-Retest Correlation) for Each Scale: Total Sample and Each Year Group Separately

	Total (n = 393)				K (n = 127)				Year 1 (n = 139)				Year 2 (n = 130)			
Scales	χ^2	*df*	RNI	Stabil Corr	χ^2	*df*	RNI	Stabil Corr	χ^2	*df*	RNI	Stabil Corr	χ^2	*df*	RNI	Stabil Corr
Physical:																
T1	5.63	2	.98	. . .	5.93	2	.86	. . .	.28	2	1.02	. . .	6.00	2	.97	. . .
T2	12.86	2	.97	. . .	4.53	2	.97	. . .	8.94	2	.94	. . .	3.37	2	.99	. . .
T1/T2	42.14	15	.96	.601	34.18	15	.87	.798	20.54	15	.98	.523	16.57	15	.99	.573
Appearance:																
T1	33.82	2	.93	. . .	16.04	2	.84	. . .	2.31	2	1.00	. . .	9.97	2	.96	. . .
T2	24.29	2	.96	. . .	17.81	2	.89	. . .	8.47	2	.97	. . .	1.06	2	1.00	. . .
T1/T2	73.46	15	.95	.494	52.75	15	.85	.484	29.56	15	.97	.433	16.30	15	1.00	.492
Peers:																
T1	2.55	2	1.00	. . .	4.43	2	.98	. . .	.13	2	1.01	. . .	.72	2	1.01	. . .
T2	12.62	2	.97	. . .	1.93	2	1.00	. . .	5.62	2	.97	. . .	5.94	2	.98	. . .
T1/T2	30.40	15	.98	.451	13.15	15	1.01	.295	22.56	15	.98	.350	17.76	15	.99	.651
Parent:																
T1	8.82	2	.97	. . .	3.81	2	.97	. . .	.16	2	1.02	. . .	10.46	2	.92	. . .
T2	14.41	2	.96	. . .	5.68	2	.96	. . .	12.17	2	.91	. . .	.53	2	1.01	. . .
T1/T2	37.18	15	.97	.459	15.94	15	.99	.401	29.59	15	.94	.436	19.53	15	.98	.587
Reading:																
T1	18.19	2	.98	. . .	9.73	2	.95	. . .	2.69	2	1.00	. . .	5.13	2	.99	. . .
T2	11.23	2	.99	. . .	3.18	2	1.00	. . .	3.05	2	.99	. . .	3.67	2	1.00	. . .
T1/T2	57.49	15	.97	.417	27.98	15	.97	.312	24.65	15	.98	.247	27.20	15	.98	.637
Math:																
T1	7.85	2	.99	. . .	2.53	2	1.00	. . .	2.12	2	1.00	. . .	7.82	2	.98	. . .
T2	12.31	2	.98	. . .	15.47	2	.92	. . .	4.79	2	.99	. . .	12.66	2	.97	. . .
T1/T2	30.05	15	.99	.356	29.70	15	.95	.244	13.27	15	1.00	.307	32.10	15	.97	.462
School:																
T1	.42	2	1.00	. . .	.78	2	1.02	. . .	.38	2	1.01	. . .	.13	2	1.01	. . .
T2	25.33	2	.96	. . .	17.57	2	.88	. . .	4.59	2	.99	. . .	6.83	2	.98	. . .
T1/T2	39.51	15	.98	.483	31.63	15	.92	.551	8.77	15	1.02	.337	27.05	15	.97	.542
Esteem:																
T1	1.07	2	1.00	. . .	.01	2	1.02	. . .	1.26	2	1.00	. . .	3.56	2	.99	. . .
T2	3.66	2	.99	. . .	.08	2	1.02	. . .	2.94	2	.99	. . .	2.97	2	.99	. . .
T1/T2	33.93	15	.97	.299	15.81	15	1.00	.317	33.31	15	.94	.219	14.85	15	1.00	.417

Note: T1 = time 1, T2 = time 2, *df* = degrees of freedom, RNI = relative noncentrality index, Stability Corr = test-retest correlation for analyses of T1/T2 responses. Congeneric one-factor models were conducted for T1 and T2 responses for the total sample and for each year group (kindergarten, Year 1, Year 2) and evaluated in relation to RNI. Two-factor models (with correlated uniquenesses relating responses to the same measured variables administered at T1 and T2) were then conducted for T1/T2 responses, and these are summarized by RNIs and test-retest correlations.

with an additional set of parameters. Tests of statistical significance were used to evaluate whether freeing the additional parameters led to an improved fit to the data compared to Model 4.

In Model 5, the constraint requiring the stability coefficients leading from T1 self-concept to T2 self-concept to be invariant across the three age-cohort groups was relaxed. This led to a statistically significant ($p < .05$) improvement in fit for two scales (Parents, Reading) and marginally improved fits ($p < .10$) in two other scales. For all four of these scales, the stability coefficient increased with age.

In Model 6, the invariance constraints on path coefficients leading from teacher ratings and gender to T1 and T2 self-concept were relaxed. However, the change in chi-square was not even marginally significant for any of the SDQ scales. Finally, in Model 7, the invariance constraint on the correlation between teacher ratings and gender was relaxed. Here again, however, these constraints were not even marginally significant for any of the SDQ scales.

Based on these results summarized in Table 5, Model 4 (with all parameters invariant across the age cohorts except for the uniquenesses) was selected as the best fitting model and selected parameter estimates from this model are presented in Table 6. For each SDQ scale, correlations between gender, teacher ratings, T1 self-concept, and T2 self-concept are pre-

Table 5 Invariance Tests Conducted on Path Model

	Model 1 All Free			Model 2 Fact Load Invar			Model 3 All Inv			Model 4 Uniq (U) Free			Model 5 U & Stabil Free			Model 6 U & Paths Fr			Model 7 U & TR/Sex Corr Free		
Scale	χ^2	TLI	RNI	χ^2	TLI	RNI	χ^2	TLI	RNI	χ^2	TLI	RNI	χ^2	TLI	RNI	χ^2	TLI	RNI	χ^2	TLI	RNI
df	93			105 (13)^a			141 (48)^a			125 (32)^a			123 (2)^b			117 (8)^b			123 (2)^b		
Phys	159.6	.88	.92	169.9	.90	.92	234.5**	.89	.88	189.6**	.91	.92	189.3	.91	.92	182.4	.91	.92	189.4	.91	.92
Appr	167.3	.91	.94	196.2**	.90	.92	256.9**	.90	.90	216.0**	.92	.92	215.9	.91	.92	208.5	.91	.92	214.5	.91	.92
Peer	118.7	.96	.97	139.8**	.95	.96	190.9**	.95	.94	162.1**	.96	.96	156.3*	.96	.96	155.9	.95	.96	161.1	.95	.96
Prnt	118.9	.94	.96	137.9*	.93	.94	311.8**	.72	.71	164.2**	.93	.93	156.9*	.94	.94	156.7	.92	.93	163.0	.93	.93
Read	141.3	.95	.97	152.0	.96	.97	236.9**	.94	.94	192.9**	.95	.96	170.8*	.97	.97	180.5	.95	.96	192.7	.95	.95
Math	145.9	.94	.96	155.7	.95	.96	257.4**	.92	.91	175.9**	.96	.96	170.3*	.96	.97	170.4	.95	.96	175.6	.96	.96
Schl	127.8	.95	.97	155.6**	.94	.95	214.3**	.93	.93	173.1**	.95	.95	170.6	.95	.95	163.9	.95	.95	172.9	.95	.95
Estm	118.8	.95	.97	138.0	.94	.96	209.9**	.91	.91	157.7**	.95	.96	157.0	.95	.95	150.7	.95	.95	156.8	.95	.95

Note: Phys = physical, Appr = appearance, Peer = peers, Prnt = parents, Read = reading, Math = mathematics, Schl = school, Estm = esteem. Alternative models were specified such that all parameters were free (i.e., no invariance constraints were imposed; Model 1); factor loadings were invariant across groups (Model 2); all parameters were invariant across groups (Model 3; total invariance); only uniquenesses were free (Model 4); uniqueness and self-concept stability coefficients were free (Model 5); uniquenesses and paths leading from teacher ratings and gender to T1 and T2 self-concept were free (Model 6); and uniquenesses and correlations between teacher ratings and gender were free (Model 7).

[a] Tests of statistical significance compared this model with no invariance constraints (TOT FR), and the value in parentheses is the difference in *df* for the two tests.

[b] Tests of statistical significance compared this model with U FR, and the value in parentheses is the difference in *df* for the two tests.

* $p < .10$; ** $p < .05$.

Table 6 Path Coefficients (PC) and Correlations (*r*) Relating Time 1 and Time 2 Self-Concept (T1, T2), Teacher Ratings of Self-Concept (TR), and Gender (Gen) for Each Self-Concept Factor

Variables		T1	T2	TR	Gender
Physical:					
T1	*r*				
T2	*r*	.65**			
TR	*r*	.14**	.31**		
Gend	*r*	−.42**	−.48**	−.09*	
T1	pc	0	0	.10	.41**
T2	pc	.52**	0	.22**	.25*
Appearance:					
T1	*r*				
T2	*r*	.48**			
TR	*r*	.04	.08		
Gend	*r*	.11**	.02	.11*	
T1	pc	0	0	.03	.11**
T2	pc	.48**	0	.06	−.04
Peers:					
T1	*r*				
T2	*r*	.46**			
TR	*r*	.14**	.30**		
Gend	*r*	.05	.06	−.03	
T1	pc	0	0	.14**	.05
T2	pc	.42**	0	.24**	.05
Parents:					
T1	*r*				
T2	*r*	.47**			
TR	*r*	−.06	.13**		
Gend	*r*	.08	.08	.01	
T1	pc	0	0	−.06	.08
T2	pc	.48**	0	.16**	.04
Reading:					
T1	*r*				
T2	*r*	.44**			
TR	*r*	.23**	.16**		
Gend	*r*	.09*	.08*	.07	
T1	pc	0	0	.23**	.08
T2	pc	.42	0	.06	.04
Math:					
T1	*r*				
T2	*r*	.37**			
TR	*r*	.14**	.16**		
Gend	*r*	−.05	−.12**	.00	
T1	pc	0	0	.14**	−.04
T2	pc	.34**	0	.11**	−.10**
School:					
T1	*r*				
T2	*r*	.48**			
TR	*r*	.13**	.19**		
Gend	*r*	.06	.05	.11**	
T1	pc	0	0	.12**	.05
T2	pc	.46**	0	.13**	.01
Esteem:					
T1	*r*				
T2	*r*	.34**			
TR	*r*	.12**	.19**		
Gend	*r*	.09	−.01	.09*	
T1	pc	0	0	.12**	.08
T2	pc	.32**	0	.16**	−.05

Note: Path coefficients (pc) and correlations relating the major constructs based on the path model in which uniquenesses were freely estimated for each age group, but all other coefficients (including those presented here) were constrained to be invariant across age groups (see Table 3).

* $p < .10$; ** $p < .05$.

sented first. These are followed by path coefficients leading from gender and teacher ratings to T1 and T2 self-concept, and from T1 self-concept to T2 self-concept. The stability coefficients leading from T1 self-concept to T2 self-concept tended to be somewhat smaller than those in Table 4 because the effects of gender and teacher ratings have been removed. However, the differences were not large, and the pattern of results was very similar.

Gender differences. Gender differences (Table 6) were represented as simple correlations between gender and self-concept and as path coefficients based on the path model. Path coefficients leading from gender to T1 and T2 Physical self-concept were both significant, implying that differences favoring boys were increasing with time. Consistent with this observation, the simple correlation between gender and T2 Physical self-concept was larger than the corresponding correlation with T1 Physical self-concept.

For Appearance self-concept, gender was positively correlated with T1 self-concept but was not significantly related to T2 self-concept. Whereas the path from gender to T1 Appearance was significantly positive, the path from gender to T2 Appearance was nonsignificantly negative. Consistent with Marsh (1989), this suggests that the differences in favor of girls at T1 may decline with age.

For Reading self-concept, correlations between gender and self-concept were marginally significant at T1 and T2 ($p < .10$), but neither path coefficient from gender to self-concept was significant. For Math self-concept, the path from gender to T1 self-concept was nonsignificant whereas the path to T2 self-concept was significantly negative. This implied that the gender differences in favor of boys were increasing over time. For all other SDQ scales there are neither significant correlations nor significant path coefficients relating gender to self-concept.

Although not the major focus of the present investigation, it was also interesting to note that the pattern of gender differences in teacher ratings of students' self-concepts were reasonably consistent with those observed in student ratings. Whereas teachers inferred no gender differences in Math self-concept favoring boys at T1, the corresponding gender difference in T1 self-concept ratings was also nonsignificant. Although teachers inferred boys to have higher Physical self-concepts than girls, the size of the gender differences was much smaller than observed in student self-concept ratings.

Teacher inferred self-concept ratings. The most interesting results in Table 6, perhaps, were correlations and path coefficients leading from teacher ratings of students' self-concepts and students' actual self-concepts. For all but one of the SDQ scales, T1 teacher ratings were more highly correlated with T2 self-concept ratings than with T1 self-concept ratings. Consistent with this observation, teacher ratings contributed to the prediction of T2 self-concept ratings beyond the contribution of T1 self-concept ratings for six of eight SDQ scales—all but Appearance and Reading. For Reading self-concept, teacher inferences were more accurate than any other scales at T1 and all of the relation between teacher ratings and T2 Reading self-concept was mediated by T1 self-concept. Teacher inferences of Appearance self-concept were not significantly related to either T1 or T2 self-concept ratings by students. Particularly because T1 teacher ratings and T1 self-concept ratings by students were collected at the same time (near the end of the school year), it is particularly noteworthy that teacher ratings were more highly correlated with T2 student ratings collected a year later when students had a new teacher. These results provided clear support for the hypothesis that student self-concept ratings grew more predictable with time—that they were less likely to make idiosyncratic self-ratings and were more likely to base their self-concept ratings on criteria like those used by external observers.

The Effects of Gender, Age Cohort, and Time on Multiple Dimensions of Self-Concept

Here we simultaneously evaluated age differences with cross-sectional comparisons and longitudinal comparisons of the same age cohort on different occasions based on an MCMO design. The critical comparisons involved cross-sectional comparisons based on the multiple age cohorts, longitudinal comparisons based on responses by the same cohort on the multiple occasions (T1 and T2), and age cohort × occasion interactions that tested the consistency of longitudinal comparisons over the different age cohorts. This MCMO design was operationalized as a 3 (age cohort) × 2 (gender) × 2 (time) design in which time was a within-subjects (repeated-measures) effect whereas age cohort and gender were between-subjects effects (see Table 7). The main effects of age cohort and time provided alternative (cross-sectional and longitudinal) tests of the effect of age. If there were no effect of age, then the main effects of age cohort, time, and their interaction should all be nonsignificant. If the effect of age was linear, then the effects of both age and time should both be significant, but the age cohort × time interaction should be nonsignificant. However, if the effect of age was nonlinear, then there may be main and interaction effects that would require a careful evaluation of the means for each cohort/time combination. In the present investigation the comparison of the cohort and time effects was facilitated because each age cohort differed from the next cohort by 1 year and the time interval in the longitudinal comparisons was also 1 year. The construct validity of interpretations of age differences would be strengthened if these tests provided consistent results. The main effect of gender provided a test of gender differences averaged across age cohorts and time. However, the gender × age cohort interaction and the gender × time interaction each provided alternative tests of the consistency of the gender effects over age.

Particularly in developmental research there is an apparent preference for longitudinal comparisons that also allow researchers to evaluate test-retest stability over time. Ultimately, however, mean differences based on cross-sectional comparisons and longitudinal comparisons are both legitimate approaches to evaluating age differences. Because there are potential strengths and weaknesses in both strategies, the best solution is to combine both types of comparison in the MCMO design. However, particularly when sample sizes are modest, an overreliance on simplistic tests of statistical significance can be counterproductive. Thus, for example, a marginally significant (longitudinal) time effect and a marginally nonsignificant (cross-sectional) age cohort effect may actually reflect the underlying age difference in a very consistent manner. For this reason, it is critical to evaluate the consistency in the pattern, direction, and size of age differences inferred from longitudinal and cross-sectional comparisons, particularly when only one of the comparisons is significant or there is an age cohort × time interaction.

Age differences. The main effects of either age cohort or time were statistically significant for Appearance, Parents, and School, whereas the age cohort × time interaction effect was significant for Reading. For Appearance, there was a statistically significant decline in self-concept that was evident in both the cross-sectional (age cohort) and longitudinal (time) indicators of age differences. Because the time × age cohort interaction was not significant, the decline in Appearance over time did not vary as a function of the age cohort. For School self-concept, there was a decline in self-concept with age cohort and a marginal decline ($p = .07$) with time. Whereas the differences were not large, they were reasonably consistent across the two indicators of age differences.

For Parents self-concept, there was a significant age cohort effect that interacted with time, but no main effect of time. Considering both age cohort and time means, self-concept was stable for kindergarten

Table 7 Effects of Gender, Year in School, and Time on Multiple Dimensions of Self-Concept

	K (N = 127)				Time 1 (N = 139)				Time 2 (N = 130)				p Values for Tests of Significance						
	Time 1		Time 2		Time 1		Time 2		Time 1		Time 2								
Scale and Gender	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	Gend (G)	Year (Y)	G × Y	Time (T)	T × G	T × Y	T × G × Y
Physical:																			
Male	4.50	.43	4.61	.43	4.53	.52	4.59	.49	4.62	.41	4.52	.50	.00	.31	.14	.06	.01	.16	.47
Female	4.27	.47	4.17	.59	4.24	.53	4.10	.55	4.12	.46	3.97	.58							
Total	4.39	.46	4.40	.55	4.37	.54	4.33	.58	4.39	.50	4.27	.61							
Appearance:																			
Male	4.11	.70	4.10	.72	4.01	.82	3.94	.83	3.96	.74	3.89	.79	.28	.00	.52	.01	.16	.91	.73
Female	4.36	.53	4.17	.70	4.12	.61	3.93	.80	3.96	.76	3.86	.74							
Total	4.23	.63	4.13	.71	4.07	.71	3.93	.81	3.96	.75	3.88	.76							
Peers:																			
Male	4.24	.67	4.16	.74	4.20	.71	4.17	.67	4.17	.76	4.16	.70	.23	.58	.85	.25	.93	.70	.97
Female	4.33	.64	4.25	.49	4.29	.63	4.23	.57	4.19	.59	4.19	.71							
Total	4.29	.65	4.20	.63	4.25	.67	4.20	.62	4.18	.69	4.17	.70							
Parents:																			
Male	4.36	.54	4.43	.56	4.39	.64	4.55	.49	4.58	.45	4.50	.59	.04	.02	.82	.24	.55	.03	.05
Female	4.57	.40	4.44	.49	4.47	.40	4.57	.37	4.59	.35	4.66	.33							
Total	4.46	.49	4.44	.53	4.44	.52	4.56	.43	4.58	.40	4.57	.49							
Reading:																			
Male	4.20	.69	4.35	.79	4.40	.65	4.45	.50	4.24	.79	4.01	.91	.03	.14	.06	.74	.93	.00	.31
Female	4.40	.56	4.39	.59	4.32	.67	4.43	.57	4.48	.51	4.33	.62							
Total	4.30	.64	4.37	.70	4.36	.66	4.44	.54	4.35	.68	4.16	.80							
Math:																			
Male	4.26	.66	4.21	.74	4.23	.77	4.34	.57	4.11	.85	4.15	.90	.08	.50	.63	.82	.34	.69	.44
Female	4.10	.75	4.13	.78	4.14	.74	4.10	.72	4.18	.68	4.03	.80							
Total	4.18	.71	4.17	.76	4.18	.75	4.21	.66	4.14	.77	4.09	.85							
School:																			
Male	4.14	.62	4.03	.79	4.13	.71	4.14	.59	3.92	.82	3.84	.84	.18	.04	.28	.07	.79	.44	.76
Female	4.21	.60	4.15	.64	4.10	.69	4.09	.75	4.15	.59	3.98	.64							
Total	4.17	.61	4.09	.72	4.12	.70	4.11	.68	4.02	.73	3.91	.75							
Esteem:																			
Male	4.12	.69	4.24	.57	4.25	.64	4.29	.50	4.23	.64	4.25	.53	.33	.86	.43	.99	.10	.94	.59
Female	4.35	.52	4.26	.62	4.28	.57	4.25	.53	4.27	.44	4.23	.40							
Total	4.23	.62	4.25	.59	4.27	.60	4.27	.51	4.25	.55	4.24	.47							
Total:																			
Male	4.24	.46	4.26	.47	4.27	.53	4.31	.40	4.23	.49	4.17	.48	.81	.26	.61	.17	.17	.38	.80
Female	4.32	.42	4.24	.47	4.24	.45	4.22	.42	4.24	.36	4.15	.36							
Total	4.28	.44	4.25	.47	4.26	.48	4.26	.42	4.24	.43	4.16	.43							

Note: For each self-concept scale and the total score a 2 (time) × 3 (year in school) × 2 (gender) analysis of variance was conducted in which time was a repeated measures (within-subject) variable and gender and year in school were between-subjects variables; p values for each effect are presented.

(4.46) and Year 1 (4.44 for both T2 from the kindergarten cohort and T1 for the Year 1 cohort), increased between Year 1 and Year 2, and then was stable over Year 2 and Year 3 (means of 4.56, 4.58, 4.57 for the T2 cohort of Year 1 and both times for the Year 2 cohort). In this case, there was a reasonably consistent pattern of results when both age cohort and times within each cohort were evaluated simultaneously.

For Reading self-concept, there were no significant effects of either age cohort or time, but there was a significant interaction between these two effects. Considering all six means constituting the three age cohorts and two times, there appeared to be an "inverted u" effect in which self-concept increased across the first two age cohorts and over time within each of these age cohorts, and then decreased from the second age cohort to the third age cohort and over time within the third age cohort. Again, there was a reasonably consistent pattern of age differences when means for each age cohort and times within each age cohort were considered simultaneously.

Gender differences. There were significant main effects ($p < .05$) for Physical self-concept (favoring boys), Parents (favoring girls), and Reading self-concept (favoring girls) and a marginal gender difference ($p = .08$) in Math self-concept (favoring boys). Because some differences favored girls whereas others favored boys, the differences in the total score were not statistically significant. Interestingly, there were no significant differences in Esteem even though research reviewed earlier based on older participants typically reported significant differences favoring males (Marsh, 1989).

Of particular interest was the question of whether gender differences were consistent over cross-sectional and longitudinal age comparisons. For Physical self-concept there was a significant gender × time interaction ($p < .01$) in which gender differences favoring boys are larger at T2 than T1 for each of the three age cohorts. Although not statistically significant ($p = .14$), there was a similar pattern of results for the gender × age cohort interaction. For each time, gender differences favoring boys were smallest in kindergarten, intermediate for Year 1, and largest at Year 2. Whereas no other interaction effects involving gender were statistically significant at the traditional $p < .05$ level, there were marginal effects for Reading (gender × age cohort; $p = .06$). For Reading self-concept, the expected gender difference favoring girls was evident in Year 2 and, to a lesser extent, in kindergarten, but not in Year 1. These differences, however, were consistent over time, suggesting that the marginal effect may have been an idiosyncratic cohort difference.

In summary, gender differences for even these very young children appeared to be consistent with those found by older participants—favoring girls in Parent and Reading, favoring boys in Physical and, to a lesser extent, Math. Furthermore, except for Physical self-concept, there was no clear indication that the age differences observed here varied with either cross-sectional or longitudinal differences in age. For Physical self-concept, gender differences favoring boys increased significantly with the longitudinal indicator of age and increased marginally ($p = .14$) with the cross-sectional indicator of age. Whereas the age cohort effect on Physical self-concept may not warrant consideration on its own, at least the direction of the effect was consistent with the longitudinal effect.

DISCUSSION

This research addresses a variety of converging theoretical perspectives, particularly our earlier work based on the original Shavelson et al. (1976) model (e.g., Marsh, 1985, 1990; Marsh et al., 1984; Marsh & Hattie, 1996) and work by Eccles and colleagues (Eccles et al., 1993; Wigfield & Eccles, 1992; Wigfield et al., 1997) based in part on earlier research (e.g., Nicholls, 1979; Parsons & Ruble, 1977; Stipek & MacIver, 1989). Marsh (1990) proposed that self-concepts of very young children are consistently high but that with increasing life experience children learn their relative strengths and weaknesses so that with increasing levels of age, mean levels of self-concept decline, individual self-concepts become more differentiated, and self-concept becomes more highly correlated with external indicators of competence (e.g., skills, accomplishments, and self-concepts inferred by significant others). Similarly, Eccles et al. (1993) proposed that the declines in mean levels of competency self-ratings reflected an optimistic bias for very young children and increased accuracy in responses by young children as they grow older and become more realistic. A number of the empirical trends evaluated here provide new and continuing support for these theoretical predictions—support for the factor structure of multidimensional self-concept responses and the increasing distinctiveness of the different self-concept components, longitudinal and cross-sectional age comparisons, accuracy of teacher inferred self-concept ratings.

Factor Structure and Distinctiveness

From both a theoretical and a practical perspective, the most important finding of this study, perhaps, was the clearly differentiated factor structure for these very young children. The results from T2 replicated the results from T1 in that each of the a priori factors was clearly defined. The T2 results were, perhaps, stronger than the T1 results in that the factor loadings tended to be larger, the factors tended to be more distinct, the factors tended to be more reliable, and relations with T1 teacher ratings tended to be larger. Because the children were 1 year older at T2 than T1, the stronger results at T2 were not unexpected. However, particularly given problems in obtaining clearly defined factor structures for very young children—based in part on the typically poor quality of extant measurement instruments for this age group (e.g., Byrne, 1996; Crain, 1996; Wylie, 1989)—these results are very encouraging.

The distinctiveness of self-concept factors and how this varies with age is an important theoretical concern that has received inadequate attention in developmental self-concept research. The size of correlations between self-concept factors decreased across cross-sectional age cohorts and decreased over time for longitudinal comparisons. This pattern of results was evident in comparisons based on raw scale scores and comparisons based on latent factor correlations in the CFAs. However, Marsh (1989) proposed that correlations between some components of self-concept should be substantial whereas others should be small. Support for the increasing distinctiveness of SDQ factors based on correlations selected a priori to be small was particularly strong. The average size of these selected correlations did not differ from the average size of all correlations for the youngest children at T1, but the differences steadily increased over age cohorts and over time within each age cohort. In an alternative operationalization of distinctiveness, the *SD* of the eight self-concept scale scores was computed for each child at T1 and T2. These *SD*s showed a clear pattern of increasing distinctiveness (higher *SD*s) over age cohorts and over time within age cohorts. The consistent pattern of results for the various comparisons of age differences in the size of correlations and the size of *SD*s supported the conclusion that self-concepts became more distinct with age for the age range considered here. Finally, based on the MTMM studies in which multiple occasions were taken to be the multiple methods, support for the discriminant validity of the SDQ-I factors improved with age. We interpreted these various results to mean that young children were better able to distinguish between their relative strengths and weaknesses due to gaining life experience that comes with age, and that this was reflected in their self-concept responses.

Stability and Change in the Development of Self-Concept

The theoretical perspectives summarized earlier suggest that self-concepts should decline with age at least through the early adolescent period, and there seems to be reasonable support for these predictions based on responses by older children. The present investigation offered a potentially important contribution due to the MCMO design coupled with the good psychometric properties for responses by very young children. For four of eight individual scales (Physical, Peers, Math, Esteem) there were no age differences for either cross-sectional or longitudinal comparisons. The strongest and most consistent effect was for Appearance self-concept in that both cross-sectional and longitudinal comparisons indicated that self-concept declined with age. Similarly, both cross-sectional and longitudinal comparisons showed that School self-concept declined with age. Reading self-concept showed an initial increase followed by a decline that was also consistent across cross-sectional and longitudinal comparisons. Parents self-concept was particularly interesting in that it showed a significant increase in self-concept with age, although the nature of the trend differed somewhat for T1 and T2 responses (see Table 7). These results were not fully consistent with expectations that self-concept will decline during these early years based on responses by somewhat older children (Marsh, 1989). In particular, the *increase* in Parents self-concept ran counter to expectations, although Marsh et al. (1984; Marsh, 1989) found that there was no significant decline in Parent self-concept responses during early adolescence even though responses to this scale were highest of all SDQ scales. Even though these differences in Parent responses were significant, the sizes of the differences were very small (Table 1). For other scales, there is a general tendency for decreasing self-concepts with age cohort and time, but these differences were smaller than expected.

Developmental Changes in the Predictability of Self-Concept Responses

The theoretical proposals summarized earlier suggested that self-concepts of young children become more predictable with age—more closely aligned with external indicators such as objective accomplishments and the perspectives of significant others. The inclusion of teacher ratings was a new and important feature because self-other agreement on multiple dimensions of self-concept has rarely been considered for children this young or has been found to be nonsignificant (also see Wigfield et al., 1997).

Here we found that teacher ratings were significantly related to all self-concept domains except Appearance, a domain in which self-other agreement is typically lowest for all ages and is nonsignificant for preadolescent children (Marsh, 1989; Marsh & Craven, 1991). Furthermore, the significant self-other agreement between teacher and student ratings reported here was likely to underestimate the true relation because teachers were asked to complete a single summary rating for each domain rather than the multi-item ratings completed by students. These results provided important new support for the construct validity of self-concept ratings by very young children.

The longitudinal comparisons of relations between teacher ratings of students' self-concept and students' actual self-concepts also supported the proposal that young children's self-concepts were becoming more predictable as they grew older. Teacher ratings collected at T1 were significantly related to students' ratings at T1 and T2. Logically—particularly given that the ratings were collected near the end of the school year—it might be expected that the T1 teacher ratings should be more highly correlated with T1 student ratings than T2 student ratings, and the effects of T1 teacher ratings on T2 self-concepts should be mediated by T1 self-concept ratings. This follows in that T1 teacher ratings and T1 self-concepts were collected at the same time (when the teacher had been the teacher of the student for almost a year), whereas at T2 each child typically had a new teacher who had taught the child for almost a year. However, T1 teacher ratings were more highly related to T2 self-concepts than to T1 self-concepts, and the direct effects of T1 teacher ratings were significant on T2 self-concept in addition to the effect mediated through T1 self-concepts.

Gender Differences

Although not directly implicated in theoretical predictions summarized earlier, the hypothesis that gender differences evolve with age (Marsh, 1989; Marsh et al., 1991; Wigfield et al., 1997) has received surprisingly little support. Instead, researchers have found a gender stereotypic pattern of gender differences that is surprisingly consistent over age. Because of the past difficulties in measuring self-concepts with very young children, the present investigation had the potential of making an important contribution in this area. The gender differences observed here were reasonably consistent across the different analyses and reasonably consistent with extrapolations from research based on older children. There were, however, some interesting differences based on these results and extrapolations from earlier research—the higher Appearance self-concepts for very young girls (compared to typical differences favoring boys for older students), the lack of gender differences favoring girls in School self-concept, the lack of gender differences favoring boys in Esteem, and perhaps the lack of gender differences favoring girls in Peer self-concept. Nevertheless, except for Physical self-concept, the gender differences and age × gender interactions observed here were modest, and this was consistent with previous research suggesting that gender differences are surprisingly stable over age. It was also interesting to note that the pattern of gender differences in student self-concept ratings was reasonably consistent with those based on teacher inferred self-concept ratings.

Potential Strengths and Weaknesses of the Measurement Procedure

Results of the present investigation provided stronger support for the construct validity of self-concept responses than those based on other instruments designed for very young children reviewed by Byrne (1996) or Wylie (1989). Thus, it is relevant to speculate on the potential strengths and weaknesses in the instrument and administration procedure.

The individual interview-style administration was an important feature of the strategy used here that Marsh et al. (1991) showed to be more effective than group administration procedures—even when the items were read aloud to students. However, this administration procedure required much more time than the typical group administration procedure. Many of the statistical procedures used here required reasonably large sample sizes. For present purposes, the sample sizes used here were not overly large, and even larger sample sizes would have been desirable. Hence, the individual administration procedure coupled with moderately large sample sizes were important strengths of the present investigation, but they also represented a potential limitation in the added resources required to collect the data.

The test administration was conducted by a large number of different undergraduate teacher education students with some classroom experience who were given a 2 hour training program. The training included an instructional video of the instrument actually being administered and trial administrations. These results suggested that the testing procedures are easily mastered by relatively inexperienced test administrators. However, even stronger results might have been obtained if the administration had been done by a small number of more highly trained professionals.

Self-concept instruments for young children sometimes combine the use of verbal cues and pictures, but our preliminary research suggested that the pictures were counterproductive—distracting young children from the verbal content of the items. Whereas these results are only suggestive, it would be of interest to pursue these preliminary findings with more fully developed instruments that use a pictorial format.

A significant difference between this study and most other research with very young children was the length of the questionnaire—64 items. Whereas we were initially concerned with a potential fatigue effect such that the quality of responses for items near the end of the instrument deteriorated, we actually found that these items were psychometrically stronger—not weaker. Across all items from the different scales, there was a clear progression of increasing factor loadings for items presented in first, second, third, and fourth quarters of the instrument. There was support for this effect at T1 and T2, although the effect was stronger at T1 when children were younger and had not previously completed the instrument. In fact, the larger T2 factor loadings—compared to T1 factor loadings—were largely due to the stronger performance of items in the first half of the instrument at T2. These results have important implications for early childhood researchers in that the use of short instruments may be counterproductive and may account for some of the difficulties researchers have in obtaining responses from very young children that yield good psychometric properties.

The items used in this study were from a well-established instrument (see reviews of the SDQ-I by Byrne, 1996; Hattie, 1992; Wylie, 1989), but one that was designed for somewhat older children. A potential limitation of this strategy was that the wording of some of the items (e.g., those using the term "mathematics") was overly complex for some of these very young children. However, this potential limitation was apparently offset by the flexibility of the individual administration procedure in which the meaning of any item could be explained to a child who did not understand the item. Instruments specifically designed for very young children like those reviewed by Byrne (1996) and Wylie (1989) have typically not been used with older children. Hence it is not clear whether apparent problems with at least many of these instruments are inherent in the instruments instead of—or in addition to—their use with very young students. If an instrument is not effective with a population slightly older than the target population, it is unlikely to be effective with very young children.

ACKNOWLEDGMENTS

We thank Alexander Yeung for helpful comments on earlier versions of this article.

ADDRESSES AND AFFILIATIONS

Corresponding author: Herbert W. Marsh, Faculty of Education, University of Western Sydney, Macarthur, PO Box 555, Campbelltown, NSW 2560 Australia. Rhonda Craven and Raymond Debus are at the University of Sydney.

APPENDIX

SUMMARY OF THE EIGHT SELF-CONCEPT SCALES ON THE SELF-DESCRIPTION QUESTIONNAIRE

Physical Ability: Student perceptions of their skills and interest in sports, games, and physical activities.

Physical Appearance: Student perceptions of their physical attractiveness, how their appearance compares with that of others, and how others think they look.

Peer Relationships: Student perceptions of how easily they make friends, their popularity, and whether others want them as a friend.

Parent Relationships: Student perceptions of how well they get along with their parents, whether they like their parents, and the extent to which they feel parental acceptance and approval.

Reading: Student self-perceptions of their ability, enjoyment of, and interest in reading.

Math: Student self-perceptions of their ability, enjoyment of, and interest in mathematics.

School: Student self-perceptions of their ability, enjoyment of, and interest in school subjects in general.

Esteem: Student self-perceptions of themselves as effective, capable individuals who have self-confidence and self-respect and are proud and satisfied with the way they are.

REFERENCES

Baltes, P. B., & Nesselroade, J. R. (Eds.). (1979). *Longitudinal research in the study of behavior and development.* New York: Academic Press.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107,* 238–246.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Byrne, B. M. (1984). The general/academic self-concept nomological network: A review of construct validation research. *Review of Educational Research, 54,* 427–456.

Byrne, B. M. (1989). *A primer of LISREL: Basic applications and programming for confirmatory factor analytic models.* New York: Springer Verlag.

Byrne, B. (1996). *Measuring self-concept across the life span: Issues and instrumentation.* Washington, DC: American Psychological Association.

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

Chapman, J. W., & Tunmer, W. E. (1995). Development of children's reading self-concepts: An examination of emerging subcomponents and their relation with reading achievement. *Journal of Educational Psychology, 87,* 154–167.

Crain, R. M. (1996). The influence of age, race, and gender on child and adolescent multidimensional self-concept. In B. A. Bracken (Ed.), *Handbook of self-concept: Developmental, social, and clinical considerations* (pp. 395–420). New York: Wiley.

Dusek, J. B., & Flaherty, J. F. (1981). The development of self-concept during adolescent years. *Monographs of the Society for Research in Child Development, 46*(4, Serial No. 191).

Eccles, J. S., Adler, T. F., & Meece, J. L. (1984). Sex differences in achievement: A test of alternative theories. *Journal of Personality and Social Psychology, 46,* 26–43.

Eccles, J. S., & Wigfield, A. (1995). In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin, 21,* 215–225.

Eccles, J. S., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development, 64,* 830–847.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116,* 429–456.

Harter, S. (1983). Developmental perspectives on the self-system. In E. M. Hetherington (Ed.), P. H. Mussen (Series Ed.), *Handbook of child psychology: Vol. 4. Socialization, personality, and social development* (4th ed., pp. 275–385). New York: Wiley.

Harter, S. (1985). Competence as dimensions of self-evaluation: Toward a comprehensive model of self-worth. In R. L. Leahy (Ed.), *The development of self* (pp. 55–122). New York: Academic Press.

Harter, S. (1986). Processes underlying the construction, maintenance, and enhancement of self-concept in children. In S. Suls & A. Greenwald (Eds.), *Psychological perspectives of the self* (Vol. 3, pp. 136–182). Hillsdale, NJ: Erlbaum.

Harter, S., & Pike, R. (1984). The pictorial scale of perceived competence and social acceptance for young children. *Child Development, 55,* 1969–1982.

Hattie, J. (1992). *Self-concept.* Hillsdale, NJ: Erlbaum.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications.* Chicago: SPSS.

Joseph, B. W. (1979). *Pre-school and Primary Self-Concept Screening Test: Instruction manual.* Chicago: Stoelting.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences.* Stanford, CA: Stanford University Press.

Markus, H., & Wurf, E. (1987). The dynamic self-concept: A social psychological perspective. *Annual Review of Psychology, 38,* 299–337.

Marsh, H. W. (1985). Age and sex effects in multiple dimensions of preadolescent self-concept: A replication and extension. *Australian Journal of Psychology, 37,* 197–204.

Marsh, H. W. (1988). *Self Description Questionnaire: A theoretical and empirical basis for the measurement of multiple dimensions of preadolescent self-concept: A test manual and a research monograph.* San Antonio, TX: Psychological Corp.

Marsh, H. W. (1989). Age and sex effects in multiple dimensions of self-concept: Preadolescence to early-adulthood. *Journal of Educational Psychology, 81,* 417–430.

Marsh, H. W. (1990). A multidimensional, hierarchical model of self-concept: Theoretical and empirical justification. *Educational Psychology Review, 2,* 77–172.

Marsh, H. W. (1993a). Academic self-concept: Theory measurement and research. In J. Suls (Ed.), *Psychological perspectives on the self* (Vol. 4, pp. 59–98). Hillsdale, NJ: Erlbaum.

Marsh, H. W. (1993b). Stability of individual differences in multiwave panel studies: Comparison of simplex models and one-factor models. *Journal of Educational Measurement, 30,* 157–183.

Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical processes. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 315–353). Hillsdale, NJ: Erlbaum.

Marsh, H. W., Barnes, J., Cairns, L., & Tidman, M. (1984). The Self Description Questionnaire (SDQ): Age effects in the structure and level of self-concept for preadolescent children. *Journal of Educational Psychology, 76,* 940–956.

Marsh, H. W., Barnes, J., & Hocevar, D. (1985). Self-other agreement on multidimensional self-concept ratings: Factor analysis and multitrait-multimethod analysis. *Journal of Personality and Social Psychology, 49,* 1360–1377.

Marsh, H. W., & Byrne, B. M. (1993). Confirmatory factor analysis of multitrait-multimethod self-concept data: Between-group and within-group invariance constraints. *Multivariate Behavioral Research, 28,* 313–349.

Marsh, H. W., & Craven, R. G. (1991). Self-other agreement on multiple dimensions of preadolescent self-concept: The accuracy of inferences by teachers, mothers, and fathers. *Journal of Educational Psychology, 83,* 393–404.

Marsh, H. W., & Craven, R. G. (1997). Academic self-concept: Beyond the dustbowl. In G. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement and adjustment.* San Diego, CA: Academic Press.

Marsh, H. W., Craven, R. G., & Debus, R. L. (1991). Self-concepts of young children aged 5 to 8: Their measurement and multidimensional structure. *Journal of Educational Psychology, 83,* 377–392.

Marsh, H. W., & Grayson, D. (1995). Latent-variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Issues and applications* (pp. 177–198). Newbury, CA: Sage.

Marsh, H. W., & Hattie, J. (1996). Theoretical perspectives on the structure of self-concept. In B. A. Bracken (Ed.), *Handbook of self-concept.* New York: Wiley.

Marsh, H. W., & Hocevar, D. (1985). The application of confirmatory factor analysis to the study of self-concept: First and higher order factor structures and their invariance across age groups. *Psychological Bulletin, 97,* 562–582.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107,* 247–255.

Nicholls, J. (1979). Development of perceptions of own attainment and causal attributions of success and failure in reading. *Journal of Educational Psychology, 71,* 94–99.

Norusis, M. J. (1993). *SPSS for Windows.* Chicago: SPSS.

Parsons, J. E., & Ruble, D. N. (1977). The development of achievement-related expectancies. *Child Development, 48,* 1075–1079.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design and analysis: An integrated approach.* Hillsdale, NJ: Erlbaum.

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

Stipek, D. J. (1981). Children's perceptions of their own and their classmates' ability. *Journal of Educational Psychology, 73,* 404–410.

Stipek, D. J., & MacIver, D. (1989). Developmental change in children's assessment of intellectual competence. *Child Development, 60,* 521–538.

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review, 6,* 49–78.

Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review, 12,* 265–310.

Wigfield, A., Eccles, J. S., MacIver, D., Reuman, D. A., & Midgley, C. (1991). Transitions during early adolescence: Changes in children's domain specific self-perceptions and general self-esteem across the transition to junior high school. *Developmental Psychology, 27,* 552–565.

Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A. J., Freedman-Doan, C., & Blumenfeld, P. C. (1997). Changes in children's competence beliefs and subjective task values across the elementary school years: A three-year study. *Journal of Educational Psychology, 89,* 451–469.

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

Wylie, R. C. (1989). *Measures of self-concept.* Lincoln: University of Nebraska Press.