The structure, stability and measurement of young children's self-concepts: Advances in new times

Rhonda Craven
Herb Marsh

University of Sydney
University of Western Sydney

For older children, there have been considerable advances in self-concept theory, measurement and intervention design. However, these advances have not been fully applied to research with young children. In particular, psychometrically strong instruments have not been developed for young children and the factorial structure of self-concept is not well understood for this age group. A new, individual administration procedure for assessing multiple dimensions of self-concept for young children 5-8 years of age was the basis of this study. We expanded this application in a multi-cohort-multi-occasion study. Reliability, stability, factor structure, and the distinctiveness of the SDQ factors improved with age and from one year to the next, but small gender differences were reasonably stable over age. Consistent with the proposal that children's self-perceptions grow more realistic with age, T1 teacher ratings were more highly correlated with student 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 new measurement procedure.

For older children, there have been considerable advances in the quality of self-concept research due to stronger theoretical models, the development of multidimensional measurement instruments based on theoretical models, and stronger interventions (see Byrne, 1984; 1996; Harter, 1983; 1985; 1986; Marsh, 1990; 1993a; Marsh & Craven, 1997; Marsh & Hattie, 1996). These advances in theory, measurement and intervention design for older students have not, however, been fully applied to research with young children. In particular, psychometrically strong instruments have not been developed for young children and the factorial structure (or dimensionality) of self-concept is not well understood for this age group. As a result, researchers and early childhood practitioners have not been able to accurately measure and report young children's self-concepts. Researchers (Harter, 1983; Marsh and Craven, 1997; Marsh, Craven & Debus, 1991; Stipek & Maclver, 1989; Wylie, 1989) have suggested that this problem emanates from limitations of existing measurement instruments and have recommended the use of more appropriate assessment procedures. 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. There is a need to evaluate the factorial structure of self-concept responses over time for young children and to determine whether factors like those identified in studies of the responses of older children...
can be found (e.g., Harter, 1982; Marsh, 1988, 1990), but this research is dependent in part on the development of more appropriate measurement procedures. 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 Structure and Measurement of Self Concept for Children Aged 5 to 8

Harter (1985) proposed that the concept of global self-worth does not evolve before the age of about 8. In support of this claim, Harter and Pike (1984) indicated that below age 8 children either do not understand general self-worth items or do not provide reliable responses, but they did not actually present empirical support for this supposition. Subsequent research (Silon & Harter, 1985) suggested that mental age may be more important than chronological age. Factor analyses of Harter's self-concept scale for low-IQ subjects aged 9-12 (with mental ages of less than 8) revealed only two self-concept factors instead of the four factors found with normal-IQ children. These low-IQ children did not distinguish between cognitive and physical competence, and the general self-concept items did not cluster together or load on other factors.

Particularly for responses by young children, the failure to identify the intended factors may reflect problems with the particular instrument or the inability of children to accurately reflect their self-concepts with conventional paper-and-pencil tests. Here, progress in theory and practice may be stimulated by the development of better measurement procedures. In support of this suggestion, Marsh, Craven and Debus (1991) described a new, adaptive procedure for assessing multiple dimensions of self-concept for children aged 5-8. In an individual interview format, the SDQI was administered to 501 kindergarten, 1st and 2nd Year students. Approximately two weeks after the SDQI was individually administered, the SDQI was administered to Year 1 and Year 2 students using the normal group administration procedures.

The critical component of the Marsh, Craven and Debus (1991) study was the individualized interview format used to collect SDQI responses. Following instructions and sample items, each item was read aloud. The interviewer asked the child if he/she understood the item. If the child did not understand the item, the interviewer explained the sentence further, ascertained that the child understood the item, and read the item again. The child initially responded "Yes" or "No" to each item to indicate whether it was true or false as a description of the child. If the child responded "Yes," the interviewer then asked whether the child meant "yes, always" or "yes, sometimes." Similarly, if the child responded "No," the interviewer asked if the child meant "no always" or "no sometimes." The second probe was always repeated, thus providing a check on the accuracy of the child's first response. Pilot work indicated that some kindergarten students experienced difficulty with a few of the items. These items were initially presented in the original form and then in a paraphrased form. Thus, for example, children were told that mathematics meant work with numbers.
Results of the Marsh, Craven and Debus (1991) study indicated that internal consistency estimates varied systematically with age. The median reliabilities for the 8 SDQI scales were .74, .80, and .82 for kindergarten, Year 1 and Year 2 students. Reliability estimates for total academic and total non-academic self-concepts varied between .85 and .92 for all three ages. At each age level, CFA identified all 8 SDQI scales -- including the General Self scale. With increasing age, the fit of the 8-factor model improved, the size of correlations among the factors decreased, and self-concept became more differentiated.

There was an initial concern that the 64-item SDQI instrument would be too long for these very young children. Interestingly, items near the end were more effective than earlier items (in contrast to anticipated fatigue effects). 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.

Correlations between SDQI scales from individual and group administration procedures were statistically significant for all 8 SDQI scales (median correlations were .38 and .50 for Year 1 and 2 students). MTMM analyses of this data indicated support for convergent and discriminant validity, but a substantial method effect was associated with the group administration procedure that was larger for the younger students.

Due in part to this new measurement procedure, Marsh, Craven and Debus (1991) were able to provide new evidence about issues in the development of self-concept in very young children. In particular, appropriately measured multiple dimensions of self-concept are better differentiated than previously assumed. In current research, this individualized interview procedure for administering the SDQ is being evaluated in Head Start programs in several (US) states.

These results have important practical implications in that young children's self-concepts cannot be understood if the multidimensionality of self-concept is ignored in theory, instrument selection and assessment practices. Even children as young as 5 have multidimensional self-concepts and can conceptualise a general self-concept. Therefore, instrument development also leads to further refinements in theory which subsequently informs classroom practice.

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 multi-cohort-multi-occasion (MCMO) design with two waves of data collected one 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 teachers 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 realistic 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: (a) test the Shavelson, Hubner & Stanton (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; (b) evaluate the stability of young children's self-concepts over time; (c) evaluate gender and age differences for young children longitudinally as well as cross-sectionally, and (d) 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.

Methods

Sample

The sample considered in this study is a total of 396 students who at T1 were enrolled in kindergarten (n = 127), Year 1 (n = 139), and Year 2 (n = 130). Children at T1 in each of the three grade levels were predominantly 5 years of age (kindergarten), 6 years of age (Year 1), and 7 years of age (Year 2). The participants came primarily from middle class families and attended one of three schools in suburban metropolitan Sydney, Australia. 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 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) assesses three areas of academic self-concept (reading, mathematics, and school self-concept), four areas of nonacademic self-concept (physical ability, physical appearance, peer, and parent relations) and a general esteem scale. 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 8 SDQ-I scales was defined by responses to the same 8 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 (Marsh, 1986), 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 part of the study, teachers at T1 only rated each of their students on the 8 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 8 columns next to each student's name corresponding to the 8 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 nine-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 two-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. 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 and SEMs that were conducted with the SPSS version of LISREL (Joreskog & Sorbom, 1989).

Although a variety of different CFAs and SEMs are 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.

Results

Internal Consistency, Stability, and Distinctiveness

Internal consistency. As expected, coefficient a estimates of reliability increased with age, based on both cross-sectional and longitudinal comparisons (Table 1). The median a for T1 and T2 varied from .74 to .85 for the three age cohorts (Table 1) and there was a consistent pattern of a increasing as with age. This pattern was evident 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 a was lowest for the T1 kindergarten responses (.74) and largest for the T2 responses by Year 2 students (.85). This pattern is also evident for the individual scales. Interestingly, the only major exception to this pattern was for esteem; reliability estimates did not seem to be consistently related to either cross-sectional or longitudinal age differences.

Insert Table 1 About Here
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, labelled T12 in Table 1). Across the 8 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 in order to evaluate the discriminant validity of the self-concept responses. Adapting the terminology of MTMM analyses (Campbell & Fiske, 1959; Marsh & Grayson, 1994), the stability coefficients (coefficients 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 8 scales there were a total of 8 x 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 (General) for Year 2 responses, two scales (General, Peers) for year 1 responses, and four scales (General, Peers, Reading, 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 7 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 rs compared to the mean of all rs (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 rs and selected rs tended to decrease over time, the decrease was larger for the mean of rs selected a priori to be lower. For each age cohort, the difference between the mean of the selected rs and the mean of all rs 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 8 SDQ scale scores, computed separately for each of the age 3 cohort x 2 time combinations. The results (Table 1) provided a clear pattern in which responses became more differentiated (larger SDs) 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.

Insert Table 3 About Here

Factor Structure

CFA 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 particularly 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 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 7 of 8 factors in the T2 solution. For both T1 and T2 solutions the median factor loadings increased steadily for indicators one (.58 and .66 for T1 and T2 respectively), two (.66 and .70), three (.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.

Insert Table 4 About Here
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 labelled 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 labelled 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. In summarizing these two-factor models, the correlation between the two factors -- the T1/T2 stability coefficient -- is also presented. Two variations of this model were considered. In the first variation, there were no correlated uniquenesses (labelled "T1/T2 no CU" in Table 4)-- the residual variance associated with each measured variable was assumed to be uncorrelated with residual variances associated with other measured variables. In the second variation, the residual variance associated for each T1 measured variable was allowed to be correlated with its matching T2 residual variance term (labelled "T1/T2 CU" in Table 4). Because these models were nested, the difference in chi-square values can be evaluated in relation to the difference in df. Although the RNI for the less restrictive of nested models should typically be larger, the difference in RNIs provided an indicator of the substantive importance of including correlated uniquenesses. For the total sample, for example, the correlated uniqueness model provided a better fit for the analyses of Physical scale (RNIs of .96 vs. .92) and, perhaps, the Parents scale (RNIs of .97 vs. .95).

It was substantively important to evaluate correlated uniquenesses because if the correlated uniquenesses were substantial and positive, the failure to incorporate them into CFA models would result in positively biased estimates of stability coefficients. For example, the stability coefficient for Physical self-concept responses for the total sample was $r = .66$ when no correlated uniquenesses were included, but dropped to $r = .60$ when correlated uniquenesses were included. Whereas stability coefficients were smaller when correlated uniquenesses were included (Table 4), the differences were typically small. Across all 32 sets of analyses (8 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. Consistent with the evaluation of fit, the small differences in the estimated correlations suggested that the inclusion of correlated uniquenesses in this particular study was not a critical issue.

It is also important to evaluate the substantive implications of these stability coefficients. The stability coefficients in Table 4 -- even the ones that control for correlated uniquenesses -- were 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 rs of .32 and .37 with and without correlated uniquenesses) and Year 1 (median rs of .34 and .35) students, but those for Year 2 were substantially larger (median rs of .55 and .58).
Multiple Group Comparisons: Invariance Over Age Cohorts

In analyses summarized in this section, SEMs relating gender and teacher inferred ratings (of their students' self-concepts) to T1 and T2 self-concept ratings were evaluated for each SDQ 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.

Insert Tables 5 and 6 About Here

In SEM 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 all .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 were constrained to be invariant. Whereas the fit of Model 4 was statistically poorer than 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 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 6, Model 4 (with all parameters invariant except for the uniquenesses invariant across the age cohorts) 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 presented 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 (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 non-significantly 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 coefficients from gender to self-concept was significant. For Math self-concept, the path from gender to T1 self-concept was non-significant 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 are there 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 non-significant. Although teachers inferred boys to have higher Physical self-concepts than girls, the size of the gender differences were much smaller than observed in student 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 6 of 8 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 since 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 an external observers.

Insert Table 7 About Here

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 a MCMO design. The critical comparisons involved cross-sectional comparisons based on the multiple age cohorts, longitudinal comparisons based responses by the same cohort on the multiple occasions (T1 and T2), and age cohort x occasion interactions that tested the consistency of longitudinal comparisons over the different age cohorts. This MCMO design was operationalised as a 3 (age cohort) x 2 (gender) x 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 x time interaction should be non-significant. However, if the effect of age was non-linear, 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 one year and the time interval in the longitudinal comparisons was also one 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 x age cohort interaction and the gender x 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 weakness 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 over-reliance on simplistic tests of statistical significance can be counterproductive. Thus, for example, a marginally significant (longitudinal) time effect and a marginally non-significant (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 x 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 x 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 x age cohort interaction was not significant, the decline in Appearance over time did not vary as 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 (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 2, and then was stable over Year 2 and 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 comprising 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 x 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 x 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 x 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 and Summary

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 two-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 counter-productive -- 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 counter-productive 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 slightly older than the target population, it is unlikely to be effective with very young children.
References


