Depression scale scores in 8–17-year-olds: effects of age and gender

Adrian Angold,1 Alaattin Erkanli,2 Judy Silberg,3 Lindon Eaves,3 and E. Jane Costello1

1Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, USA; 2Department of Family and Community Medicine, Duke University Medical Center, USA; 3Virginia Commonwealth University, USA

Background: The excess of unipolar depression in females emerges in adolescence. However, studies of age effects on depression scale scores have produced divergent estimates of changes from childhood to adolescence. Method: We explored possible reasons for this discrepancy in two large, longitudinal samples of twins and singletons aged 8–17. Results: There were no differences between twins and singletons in their scores on the Short Mood and Feelings Questionnaire (SMFQ), a 13-item self-report depression scale. SMFQ scores for boys fell over this age-range, while those for girls fell from age 9 to age 11 and then increased from age 12 to age 17. The mean scores of girls under 12 and those 12 and over differed by only around one-fifth of a standard deviation. However, given the non-normal distribution of the scores, a cut point that selected the upper 6% of scores created the expected female:male ratio of 2:1. Conclusions: Implications for future research on adolescent depression are discussed. Keywords: Adolescent depression, scale scores, diagnosis.

Depression as a diagnosis is much more common in women and adolescent females than in men and adolescent boys

It is now generally accepted that, as the latest Diagnostic and Statistical Manual says: ‘Major Depressive Disorder (Single or Recurrent) is twice as common in adolescent and adult females as in adolescent and adult males. In prepubertal children, boys and girls are equally affected’ (American Psychiatric Association, 1994, p. 341). Numerous studies support this statement in relation to adults (Bebbington, 1996; Bebbington, Hurry, Tennant, Sturt, & Wing, 1981; Bland, Newman, & Orn, 1988a, b; Blazer, Kessler, McGonagle, & Swartz, 1994; Burke, Burke, Regier, & Rae, 1990; Canino et al., 1987; Cheng, 1989; Hwu, Yeh, & Chang, 1989; Kessler et al., 1994; Kessler, McGonagle, Swartz, Blazer, & Nelson, 1993; Lee, Han, & Choi, 1987; Weissman et al., 1993, 1996; Weissman & Klerman, 1977; Wells, Bushnell, Hornblow, Joyce, & Oakley-Browne, 1989; Wittchen, Essau, von Zerssen, Krieg, & Zaudig, 1992). More recently, epidemiologic studies of children and adolescents (including the Great Smoky Mountains Study, which is the focus of the present paper) have confirmed that the excess of females with depression diagnoses only arises after age 13, or around Tanner stage III of puberty (Angold, Costello, & Worthman, 1998, 1999; Cairney, 1998; Hankin et al., 1998; Lewinsohn, Hops, Roberts, Seeley, & Andrews, 1993; Reinherz, Giaconia, Lefkowitz, Pakiz, & Frost, 1993).

Studies of mean scores on depression questionnaires have been inconsistent in their findings with respect to gender and age

Whereas the results of studies of depressive diagnoses have been very consistent, the findings from studies using questionnaires and scale scores have been decidedly mixed. For instance, the basic psychometric study of the Children’s Depression Inventory (CDI), the most widely used questionnaire in this area (Kovacs, 1992), found that both 7–12-year-old and 13–17-year-old boys had higher mean scores than girls, but reported no increase in mean symptoms scores with age, and no interaction between age and gender. Several other investigators have reported similar findings using the CDI and other scales (Bartell & Reynolds, 1986; Finch, Saylor, & Edwards, 1985; Huntley, Phelps, & Rehm, 1987; Smucker, Craighead, Craighead, & Green, 1986), but others have found higher mean scores in girls (Doerfler, Felner, Rowlison, Raley, & Evans, 1988; Kazdin, French, & Unis, 1983; Petersen, Sargianni, & Kennedy, 1991; Reinherz et al., 1990; Wichström, 1999) or no difference between boys and girls (Faust, Baum, & Forehand, 1985; Gates, Lineberger, Crockett, & Hubbard, 1988; Haley, Fine, Marriage, Moretti, & Freeman, 1985; Helsel & Matson, 1984; Kovacs, 1983). Regardless of the direction of effects, or their statistical significance, any differences have typically been small (Compas et al., 1997). The failure to find consistent sex differences in mean symptom scores of samples that are predominantly nondepressed extends into the college...
years (e.g., Hammen & Padesky, 1977; King & Buchwald, 1982).

**Questionnaire studies of extreme scores have tended to parallel the diagnostic literature**

Roberts and Chen’s large multi-ethnic community study using the CES-D (Roberts & Chen, 1995) showed little or no difference by age in mean scores, or in the proportions scoring over the cut point of 15 generally used in studies of adults. However, compared with 11–13-year-olds, a larger proportion of adolescents aged 14 or older scored 24 or more, and 31 or more. Age-by-sex analyses were not reported. In another community study, Lewinsohn and Roberts found no sex difference in mean CES-D scores, but noted that 20% of girls and no boys met all nine DSM-III-R depression criteria, using items selected from the questionnaire (Roberts, Lewinsohn, & Seeley, 1995). In this sample of high school students the point prevalence and lifetime prevalence of interview-measured unipolar depression and MDD were significantly higher in girls than boys.

Teri (1982) found no sex difference in mean BDI scores at any age (14 through 17) among a sample of high school students. On the other hand, there were significantly more girls than boys with scores one or more standard deviations above the group mean. Compas and his colleagues (Compas et al., 1997) compared effects of gender on several questionnaire indices of depression from a non-referred adolescent general population sample with those from a large clinical sample. Gender effects on depression were non-significant or small in the non-referred sample, but much larger among referred adolescents. They concluded that most of the gender effect in adolescent depression can be attributed to the small group of children who are referred for treatment. However, their results are also susceptible to the interpretation that a relatively small group of severely depressed girls are more likely to be referred for services than depressed boys. Either way, Compas and colleagues pinpoint the question of whether this pattern of results implies that the higher levels of depression in females from adolescence onwards are attributable to the appearance of a female excess only at the extreme upper end of the distribution on continuously distributed measures of depression, or whether it is a reflection of an increase in scores across the whole distribution.

**Are twins like singletons in terms of their depression scale scores?**

Since the study we report here involves both twins and singletons, we next review the existing evidence pertaining to whether twins are different from singletons as far as psychopathology is concerned. Given that twins exist in an unusual social environment — they all have a sibling of exactly the same age and, in addition, monozygotic (MZ) twins are genetically identical — relatively little attention has been paid to the possibility that twinning itself might be a risk or protective factor for psychopathology. This is an important consideration, because the logic of the twin method for identifying genetic effects in general depends upon the assumption that the partitioning of genetic and environmental variance in twins will generalize to singletons (i.e., to the great majority of people). However, it is not difficult to come up with theoretical possibilities for the generation of differences in risk for psychopathology between twins and singletons. For instance, it has been found that raising twins is more stressful for parents (Hay & O’Brien, 1984; Thorpe, Golding, MacGillivray, & Greenwood, 1991), and that individual twins receive less parental attention (because it is shared with the co-twin) than singletons (Clark & Dickman, 1984; Conway, Lytton, & Pysh, 1980; Lytton, Conway, & Sauve, 1977; Tomasello, Mandle, & Kruger, 1986). Such differences in parenting could be the forerunners of differences in rates of psychopathology. Furthermore, twins do have higher rates of prematurity, language delay and reading retardation (Rutter & Redshaw, 1991), which could predispose them to higher rates of psychopathology, but there is little evidence that this is the case. Hay and O’Brien (1984, 1987) found slightly lower problem scores in twins than singletons on the Bristol Social Adjustment Guide and the Behavior Attitude Checklist, but twins were slightly less likely to be completely free of adjustment problems. A comparison of clinic attenders indicated that twins were more likely than singletons to receive a diagnosis of conduct disorder, although there were few twin–singleton differences in the rate of individual symptoms (Simonoff, 1992). Two studies resulting from the English National Child Development Study (Rutter & Redshaw, 1991) found perhaps minimally elevated conduct problem scores on parent questionnaire measures. A comparison of twins rated on the CBCL (Achenbach & Edelbrock, 1983) to published norms showed small elevations in mean scores of twins (Gau, Silberg, Erickson, & Hewitt, 1992). However, four comparisons of twins to population-matched singletons have shown few or no differences on multiple comparisons, and inconsistency in the directions of the effects (Gjone & Nevik, 1995; Levy, Hay, McLaughlin, Wood, & Waldman, 1996; Moilanen et al., 1999; van den Oord, Koot, Boomsma, Verhulst, & Orlebeke, 1995). Findings such as these led Rutter and Redshaw (1991) to conclude that the overall risk for socio-emotional disturbance in twins is not much different from that found in singletons. However, little attention has been paid to children’s self-reports — the studies mentioned above have mostly concentrated on parent and teacher reports. The exception is Moilanen and colleagues’ (1999) examination of Finnish children’s self-report scores on the Children’s Depression Inventory (CDI). They found no
differences between the mean scores of twins and singletons.

In summary, although several diagnostic general population studies have confirmed that the female excess of unipolar depression arises in adolescence, studies using scale scores have not. One possible explanation for this state of affairs is that there really are no differences between boys and girls, or between prepubertal and pubertal children on depression scale scores. In this case, we would have to suppose that such scores measured something quite different from depression diagnoses. A more moderate version of this position would suggest that there are real gender by age changes, but that such changes have a relatively small impact on the overall mean scores in the population, but are more easily detected at the high end of the score distribution because there small absolute differences in frequency lead to large risk ratios when the score distribution is dichotomized. In either case, studies with much larger numbers of subjects would be needed, either to provide a convincing null finding or to provide a sufficiently precise estimate of the possible real effect sizes involved. The purpose of this paper is to examine the relationship between gender, age, and depression as assessed by a self-report depression scale (the Short Mood and Feelings Questionnaire – SMFQ) with a view to resolving these apparently conflicting findings by combining data from two representative community studies involving 8,577 observations. In addition, we examine the question of whether the SMFQ self-reports of twins differ from those of singletons in any material way.

Methods

The data for this paper come from two large longitudinal samples of children and adolescents – The Great Smoky Mountains Study (GSMS) and the Virginia Twin Study of Adolescent Behavior and Development (VTSABD). Both are longitudinal studies of samples from the general population. Details of the methods of the GSMS can be found in the previous work of Costello et al. (Costello, Farmer, Angold, Burns, & Erkanli, 1997; Costello et al., 1996); details of the methods of the VTSABD can be found elsewhere (Hewitt et al., 1997). We present a summary of the methods pertinent to the analyses presented here.

The Great Smoky Mountains Study (GSMS)

A representative sample of 4,500 9-, 11-, and 13-year-olds, recruited through the Student Information Management System (SIMS) of the public school systems of eleven counties in western North Carolina, was selected using a household equal probability design. As close as possible to the child’s birthday, a screening questionnaire was administered to a parent (usually the mother), by telephone or in person. This consisted of 55 questions from the Child Behavior Checklist about the child’s behavior problems, together with some basic demographic and service use questions. All children scoring above a predetermined cutoff score of 20 (designed to include about 25% of the population) on the behavioral questions, plus a 1-in-10 random sample of those scoring below the cutoff, were recruited for the longitudinal study. In addition, all age-eligible American Indian children were recruited, regardless of their screen scores. Eighty percent of eligible families (N = 1,442 of which 349 were American Indian) agreed to participate in the interviews at least one wave.

Shortly after being screened, eligible children and one of their parents were interviewed. They were re-interviewed one, two and three years later. Interviews were conducted between 1992 and 1996. Between 80% and 94% of the sample participated in each wave. Since the SMFQ was available only from the first three waves of data collection, there were 3,857 observations from these waves for children aged 9 through 15.

Inform consent to participate in the study was obtained from parents, and assent was obtained from the children themselves.

The Virginia Twin Study of Adolescent Behavioral Development (VTSABD)

Sampling frame. During 1987 and 1989 all putative twin pairs known to the public and private schools system of the Commonwealth of Virginia were identified, yielding 3,264 twin pairs born between 1974 and 1982 with initially valid addresses. Of these, 2,791 (86%) responded to a brief request for information either by mail or telephone. The study was restricted, by design, to Caucasian families with twins who continued to meet the residence and age requirements for inclusion in the study during the period from March, 1990 to March, 1992. There were 1,894 such eligible families. Of these, 1,412 (75%) participated in a home visit and data collection at the first wave of interviews. Of the 1,302 families with twins still between the ages of 8 and 16 at the second wave of assessment, 1,022 (79%) of the eligible twin families completed a Wave II home interview.

Informed consent to participate in the study was obtained from parents, and assent was obtained from the children themselves.

Measures

The Short Mood and Feelings Questionnaire (Angold et al., 1995) is a 13-item scale derived from a 30-item depression question pool (the Mood and Feelings Questionnaire – MFQ). Each item consists of a simple statement (e.g., ‘I didn’t enjoy anything at all’), which is rated as being either ‘true’ (scores 2), ‘sometimes true’ (scores 1), or ‘not true’ (scores 0). It was designed to provide a rapidly administered questionnaire that reviewed a core set of symptoms. It correlated highly with more extensive evaluations, like the Children’s Depression Inventory (CDI) (Kovacs, 1983) or the Diagnostic Interview Schedule for Children (DISC) (Costello & Angold, 1988; Shaffer, Fisher, Piacentini, Schwab-Stone, & Wicks, 1989). In a previous study of 173 8–16-year-olds (Angold et al., 1995), the SMFQ was found to correlate r = .67 with CDI scores and r = .51 with DISC-C interview depression scores (the CDI correlated r = .48 with DISC-C depression scores in
the same study). The SMFQ discriminated clinically referred child psychiatric subjects from unselected pediatric controls, and depressed subjects from non-depressed subjects in a general population sample (Angold et al., 1995). The SMFQ is a unifactorial scale (Cronbach’s alpha = .90) (Costello, Benjamin, Angold, & Silver, 1991) with a robust single factor structure from ages 6 to 16 (Messer et al., 1995). Versions for self-completion by 6–18-year-olds and parents are available. A two-week test-retest study of 44 parent and child pairs yielded an ICC of .66 for the child questionnaire and .88 for the parent questionnaire (Costello & Angold, 1988).

In both the GSMS and VTSABD, children completed the SMFQ at home as part of a package of self-report and interview measures, read to them by the interviewer. In the VTSABD the SMFQ items reported here were extracted from the longer MFQ, of which the SMFQ is a subset. In the GSMS, only the SMFQ items were administered.

Analytic strategy

Data analysis for this study raises a number of thorny statistical issues, which we deal with under two headings concerning (i) distributional assumptions and (ii) the problems of combining estimates from two studies and dealing with the design effects of those studies.

Distributional assumptions and their effects on our analyses. Previous studies of depression scale scores in children and adolescents have relied upon statistics for normal distributions, but are such scores usually normally distributed? In the two studies reported here, they were not even approximately normally distributed. Most children had very low depression scores (as, indeed, we would expect) and there was a long right tail comprising a small group of children with high scores (see the cumulative frequency curves in Figures 1 and 5). Common transformations (such as taking logs or reciprocals) still did not produce distributions close to normality. Similarly, most factor analytic studies of depression scale scores have used principal component analysis or principal factor analysis, both of which rest upon the assumption that scores on the individual items making up the scale are normally distributed. Not surprisingly, this assumption was also not met. The commonest response on all the SMFQ items was 0. No transformation will produce a normal distribution from such data. This problem is not confined to depression scales. For instance, the modal response on most items on the Child Behavior Checklist (CBCL) is that the problem does not occur (Verhulst, Akkerhuis, & Althaus, 1985).

Our solution to this problem was to use approaches based on more reasonable statistical assumptions. In the case of total SMFQ scale scores, the data were Poisson distributed, so we used Poisson regression approaches. In times past, the fitting of such models was difficult, but now many statistical packages provide flexible generalized linear modeling routines that permit the selection of appropriate outcome variable distributional assumptions and link functions. We fit such models using the SAS GLIMMIX macro.

For the factor analyses we report below, we used Mplus’s (Muthén, 1984, 1989) approach to fitting factor models based on binomial or ordinal (as with the SMFQ) items. The items are assumed to have been derived from an underlying normal distribution with superimposed cut-points (two such cut-points being required for three-level items such as those in the SMFQ). Probits are computed for each item, and a weighted least squares approach is used to estimate the factor structure, with the asymptotic variance-covariance matrix of the probits serving as the weight matrix.

Design effects and combining the two studies

Here, we are dealing with two rather different longitudinal studies: the GSMS which is based on a two-phase sampling design with singletons as the sampling units, and the VTSABD which employed simple random sampling with twins as the sampling units. Because the GSMS was a two-phase, multi-wave study where only a subset of the screened subjects were interviewed annually, the statistical analyses for population inference have to be adjusted using design weights that are inversely proportional to the sampling fractions for the second phase (i.e., the diagnostic interview phase). Models also needed to account for longitudinal within-subject correlations. In the VTSABD study, on the other hand, while there was no need for sample-design adjustments, analyses had to be adjusted both for within twin-pair and longitudinal within-subject correlations. All these corrections needed to make suitable adjustments not only to the parameter estimates, but also to the variance-covariance matrices used for statistical inference.

To model the data from the GSMS and VTSABD simultaneously, while making all these necessary adjustments, we took a fixed-effect approach that conceptually combined the two data sets into one. To make the two data sets comparable, we defined a new weight variable as W* = the usual sampling weights for the GSMS, and W* = 1, if the subject was from the VTSABD (since such subjects are from a simple random sample). We also created a study indicator as STUDY = 0 for subjects in the GSMS, and STUDY = 1 for those in the VTSABD.

Where Y is a continuously (SMFQ score) or discretely distributed (above/below a cut-point) dependent variable, and X is the covariate vector or matrix common to both studies, such as age or sex, to assess the overall effect of X we fitted a generalized linear mixed regression model (using the SAS macro GLIMMIX) by expressing the link function as an additive function of the interaction term STUDY*X, in addition to the main effects STUDY and X. By considering both the main effects and interaction terms, we are able to address three key questions simultaneously: i) Are twins different from singletons with respect to Y (main effect of STUDY), ii) what is the overall relationship between X and Y (main effect of X), and iii) is the relationship between X and Y different in twins and singletons? With both terms being adjusted for variance heterogeneity. If we found substantial differences in the relationships between X and Y in twins and singletons,
then it would not make much sense to consider the overall relationship between X and Y (since it will be the weighted average of two different effects, and might not be very meaningful). However, if the estimates from the two studies were similar, the main effect of X would be equivalent to the weighted (pooled) estimate of the main effects from each study. To model (and account for) the within-subject correlations in GSMS, and within-twins and between twins correlations in VTSABD, we also needed to treat the identity of the subject (GSMS) or twin pair (VTSABD) as a cluster variable (referred to below as ID). All these corrections were achieved by including in GLIMMIX the statements:

repeated/type = un subjects = ID group = STUDY;
weight=w; procopt=empirical;

The statement procopt = empirical adjusted for the design effects, while group = STUDY allowed us to model different (heterogeneous) correlation structures for each study. It will be apparent that this covariance structure involved averaging across monozygotic and dizygotic twin pairs, but that was appropriate, because we were not concerned with whether there were differences between monozygotic and dizygotic twins, but with whether twins (as a class) differed from singletons.

Results
The factor structure of the SMFQ in twins and singletons

Previous work with the SMFQ used with representative community samples has indicated that it has an essentially unifactorial structure with high factor loadings on each item (Angold et al., 1995; Messer et al., 1995). Table 1 shows the item loadings on the first factor for each wave and twin group. In both studies all items had high positive loadings on the first factor.

Table 1 First factor loadings of SMFQ items in the VTSABD and the GSMS

<table>
<thead>
<tr>
<th>Item</th>
<th>VTSABD</th>
<th>GSMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>I felt miserable/unhappy</td>
<td>.68 [.67 .74 .74]</td>
<td>.66 [.56 .66]</td>
</tr>
<tr>
<td>I did not enjoy anything at all</td>
<td>.64 [.61 .75 .64]</td>
<td>.48 [.61 .72]</td>
</tr>
<tr>
<td>I was so tired I just sat around</td>
<td>.57 [.64 .72 .71]</td>
<td>.46 [.60 .67]</td>
</tr>
<tr>
<td>I was very restless</td>
<td>.57 [.64 .69 .68]</td>
<td>.40 [.57 .59]</td>
</tr>
<tr>
<td>I felt I was no good any more</td>
<td>.85 [.90 .94 .91]</td>
<td>.80 [.90 .95]</td>
</tr>
<tr>
<td>I cried a lot</td>
<td>.68 [.71 .74 .74]</td>
<td>.65 [.73 .77]</td>
</tr>
<tr>
<td>I found it hard to think/concentrate</td>
<td>.74 [.77 .76 .76]</td>
<td>.60 [.74 .74]</td>
</tr>
<tr>
<td>I hated myself</td>
<td>.89 [.93 .98 .93]</td>
<td>.83 [.90 .92]</td>
</tr>
<tr>
<td>I was a bad person</td>
<td>.77 [.80 .85 .85]</td>
<td>.64 [.74 .74]</td>
</tr>
<tr>
<td>I felt lonely</td>
<td>.72 [.83 .86 .83]</td>
<td>.74 [.80 .86]</td>
</tr>
<tr>
<td>I thought nobody really loved me</td>
<td>.84 [.86 .84 .95]</td>
<td>.84 [.88 .88]</td>
</tr>
<tr>
<td>I thought I could never be as good as others</td>
<td>.77 [.82 .87 .87]</td>
<td>.72 [.81 .79]</td>
</tr>
<tr>
<td>I did everything wrong</td>
<td>.83 [.87 .89 .92]</td>
<td>.74 [.81 .74]</td>
</tr>
</tbody>
</table>

The effects of repeated measurement on SMFQ scores and their implications for measuring age effects

Figure 1 shows the cumulative frequency distributions of SMFQ scores by wave in the VTSABD and GSMS. It will be seen that the curves are of similar shape in all cases (and distinctly non-normal). In addition, the curves for later waves within each study are shifted to the left (i.e., fewer subjects with high scores) compared with earlier waves. Poisson regression shows that this effect of wave was highly significant (parameter estimate = .86, \( p < .0001 \)). There was no significant difference in overall mean scores between the two studies (parameter estimate = 1.03, \( p = .4 \)), controlling for wave effects. With wave effects uncontrolled, there was still no significant difference in the overall means of the studies (parameter estimate = 1.02, \( p = .7 \)). The wave effect needs to be taken into account in many of the following analyses.

When we fitted age to the SMFQ data for the entire sample (i.e., ignoring potential gender effects), without controlling for wave, age was associated with significantly lower scores (parameter estimate = .96, \( p = .0004 \)). With wave controlled, there was no significant overall effect of age (parameter estimate = 1.0, \( p = .8 \)). Thus, the apparent overall effect of age appeared to be due to the fact that the children were assessed multiple times as they grew older (i.e., it was a within-subjects effect). To examine true effects of age, we must therefore control for wave, treating the within-subjects change with repeated measurement as a nuisance factor to be partialed out, and considering only between-subjects effects of age as unambiguously representing true effects of age. This pattern of falling scores on repeated administration of the SMFQ is typical of findings from many types of questionnaires and interviews, including previous work with the SMFQ (Angold et al., 1996; Lauritsen, 1998; Lucas et al., 1999; Piacentini...
et al., 1999). All of the information on age effects must be obtained from the between-subjects components of our analyses.

In order to estimate the absolute size of the effect of wave we fitted a model with age entered as a categorical variable with gender and the interaction of gender and age. The age and gender corrected means for waves 1, 2 (available in both the VTSABD and GSMS) and 3 (GSMS only) were 4.2 (95% CI = 4.0–4.4), 3.3 (3.1–3.4) and 2.8 (2.5–3.1). In other words there was a drop of about one point from wave 1 to wave 2 and of about half a point from wave 2 to wave 3.

Effects of gender and age

Figure 2 shows mean SMFQ scores by gender and age in the VTSABD (waves 1 and 2, both twins) and the GSMS (waves 1–3). The thin lines with markers show the raw means uncorrected for effects of wave. Visual inspection suggests that there was little difference between twins and singletons. There was the expected suggestion of an interaction between age and gender. It is striking that both twins and singletons of both sexes showed the same pattern of an early drop in depression scores, followed by an increase in girls and a flattening or perhaps some further fall in boys.

Putting the two studies together and using a Poisson model with suitable corrections for the twin correlations, repeated measures correlations and wave effects generated the corrected means, which are shown in Figure 2 by the heavy lines without markers.

In girls the study means (95% CI), controlling for age (treated categorically), were VTSABD – 3.78 (3.55–4.03); GSMS – 3.79 (3.42–4.20). P for study difference = .9. In boys the study means (95% CI), controlling for age (treated categorically), were VTSABD – 3.24 (3.03–3.47); GSMS – 3.19 (2.92–3.49). P for study difference = .8. We treated age categorically here to remove as much age variance as possible in order to leave as pure a study effect as possible.

Description of the age curves

Having shown that, controlling for age, the mean scores on the SMFQ from the two studies were almost identical, and having seen that the shapes of the age curves were very similar, we could combine the data from the two studies to model the effects of age.
For girls, visual inspection suggested a basically cubic form, and indeed Poisson regression revealed significant linear, quadratic and cubic terms with parameter estimates of –2.2, .17 and –.004 respectively. In boys, visual inspection suggested a linear or quadratic function of age. Poisson regression indicated only a significant linear effect with a parameter estimate of –.02.

**Overall model of effects of age and sex**

Having satisfied ourselves as to the shape of the gender-specific age curves, we could now consider its implications for previous findings concerning effects of gender and age. A common approach has been to dichotomize age at some point and to fit that and sex and their interaction. We therefore split the children into those under age 12 and those aged 12 and older. This approach produced small, but highly significant effects (Figure 3) (Poisson regression: Sex OR = 1.32, \(p = 8 \times 10^{-3}\); Age group OR = 1.20, \(p = .3\); Interaction OR = .75, \(p = 6 \times 10^{-5}\)). Planned contrasts indicated that there was no significant difference between the mean scores of younger boys and girls \(p = .8\), but that older girls had significantly higher scores than younger girls \(p = 4 \times 10^{-4}\), while older boys had marginally significantly lower mean scores than younger boys \(p = .03\). As a result, older girls had significantly higher mean scores than older boys \(p = 1 \times 10^{-7}\). However, the change in both males and females was about \(\frac{1}{5}\) an SMFQ point, resulting in a difference of about one point between males and females aged 12 or more. Since the standard deviation of a Poisson distribution is equal to its mean, the difference between the age groups within sex was about one-fifth of a SD, with the resulting difference between sexes at 12 or more being about \(\frac{1}{4}\) SD.

**Extreme scores on the SMFQ**

A cut-point of 11 or more on the SMFQ selected the top 6% SMFQ scores. We compared the probabilities of being in this group by age and sex. As can be seen in Figure 4, in those aged 12 and above, the result was a 2.2:1 female: male ratio, just as expected from the diagnostic literature.

Logistic regression indicated the presence of a significant interaction of sex and age group: Sex OR = 2.3, \(p = .0002\); Age group OR = 1.4, \(p = .4\); Interaction OR = .52, \(p = .007\). Planned contrasts showed that there was no significant difference between younger boys and girls \(p = .4\) or between younger and older boys \(p = .3\), but that older girls were significantly more likely to be high scorers than younger girls \(p = .007\), with the result that older girls had significantly higher rates of ‘depression’ than older boys \(p = 1 \times 10^{-5}\).

**A general shift in the SMFQ distribution or the addition of a few high-scoring individuals?**

It has been suggested that increased rates of depressive symptoms in adolescent girls are best accounted for by the addition of a relatively small number of high-scoring individuals to the top end of the scale score distribution, while the rest of the distribution remains relatively unaffected by adolescence. In order to examine this proposition, we plotted the cumulative frequencies for SMFQ scores in younger (11 or less) and older (12+) boys and girls. The results are shown in Figure 5. It is clear that for both boys and girls age affected the whole distribution, rather than only the upper end.

**Conclusions**

As we showed in the introduction to this paper, previous work on the effects of age and gender on self-report depression scale scores across the
period from late childhood through adolescence has resulted in a confusing series of results, with effects in all possible directions reported from different studies. On the other hand, the diagnostic literature has been quite consistent in indicating that depression becomes more common in girls during adolescence, and that a 2–3:1 female:male ratio then persists at least throughout the female reproductive life span. Our aim was to use data from two large, general population samples to conduct a more powerful test of this question than has previously been possible. Since one of the available data sets was derived from a sample of twins, we also had the opportunity to determine whether there were differences in the scale-based self-ratings of twins and singletons.

There were no significant differences between twins and singletons

We found no differences between twins and singletons on the SMFQ in relation to the factor structure of the instrument itself, its susceptibility to occasion of measurement effects, mean population SMFQ scores, or the effects of gender and age on the scale scores. These results provide strong confirmation of Moilanen and colleagues’ (1999) finding that there were no differences between the mean scores of twins and singletons, and extend it by showing that the lack of differences extends beyond overall mean scores to encompass the patterns of correlation among scale items, and developmental changes in scores. The implication here is that the results of twin studies of self-report depression scores in children and teenagers can confidently be extrapolated to the population of singletons. In other words, when genetic or environmental effects are observed in twin studies, we can assume that similar genetic and environmental effects are at work in singletons. Were depression scale self-reports of twins to behave markedly differently from those of singletons, then genetic analyses of twins would apply only to twins (and be of relatively little value in understanding the etiology of depression in most people). However, that is clearly not the case in relation to depression, and, as we discussed in the introduction to this paper, there is no evidence for substantial effects of twinning in relation to other scale score measures of psychopathology.
There were small but highly significant interaction effects of gender and age on depression scores

We have presented the largest study of self-report depression scores to date, and it shows that there were consistent, but small (around one-fifth of a standard deviation) effects of age and gender on depression scores. Most importantly, scores increased in adolescent girls compared with younger girls and in comparison with boys, while they fell slightly in boys. These results suggest that the confusion that has characterized the previous literature has been due, in part, to sample sizes too small to detect relatively small interaction effects with confidence. We also used statistics appropriate to the observed distribution of the SMFQ data (and most depression scale data in the general population), which was distinctly non-normal. Little attention has been paid to whether normal distributional assumptions were met in other studies that have examined depression scores in this age range, and ANOVA approaches are not very robust to situations in which the distributions of scores are very severely skewed (as here, and we suspect in many general population studies using other depression questionnaires). Lack of attention to the distributional properties of the data may also, therefore, have contributed to the heterogeneity of findings in this literature.

Several previous studies have reported that more girls than boys have highly deviant depression scale scores in adolescence, even in the absence of significant effects on mean scores (Compas et al., 1997; Roberts & Chen, 1995; Roberts, Lewinsohn, & Seeley, 1995; Teri, 1982), and this had led some (e.g., Compas et al., 1997) to conclude that higher mean scores are attributable to a relatively small group of individuals who have extreme scores on continuously distributed measures of depression. Our data indicate that this is not the case. We found that the entire distribution of scores was shifted to the right in older girls. Given the distributions of the cumulative frequency curves for boys and girls, this resulted in girls aged 12 and older being twice as likely as boys aged 12 and older to have scores in the top 6% of all SMFQ scores. In other words, such a cutoff on the MFQ resulted in a gender by age interaction significant effects with confidence. We included in those groups. Again, this question deserves more detailed study.

Falling depression scores in late childhood

Another notable feature of the age curves we have described is the fall in depression scores with age in boys. A similar effect has been described with the SMFQ in the Pittsburgh Youth Study (Angold et al., 1996). That study began at age 6, and suggested that the most rapid decrease occurred before age 10. However, the age-specific mean SMFQ scores for the ages represented in both studies are remarkably similar, suggesting that this is quite a robust effect that deserves further study, since we do not know its cause. Clearly it cannot be a pubertal effect, since it appears to begin many years before puberty, but it could result from elevations in adrenal androgens (adrenarche), which begin around age 6 or before. Of course, it might also have nothing whatever to do with hormones, but result from psychosocial or even emergent genetic processes. The main point is that, until now, the Pittsburgh Youth Study data have stood alone in presenting this unexpected finding. These new data suggest that we should pay more attention to the possibility that depression scale scores fall in boys in mid to late childhood.

We were surprised to see that mean depression scores also fell in girls prior to the age of 11. This effect does not appear to have been described before, but it could also help to explain why previous comparisons between female children’s and adolescents’ scale scores have shown such variable results. If the age-depression score function is actually U-shaped, then the results of comparing ‘older’ and ‘younger’ children would be expected to vary depending upon exactly which ages were included in those groups. Again, this question deserves more detailed study.
Correspondence to
Adrian Angold, Center for Developmental Epidemiology, DUMC Box 3454, Durham, NC 27710, USA; Tel: 919-687-4686 x. 222; Fax: 919-687-4737; Email: Aangold@psych.mc.duke.edu

References


Manuscript accepted 19 March 2002