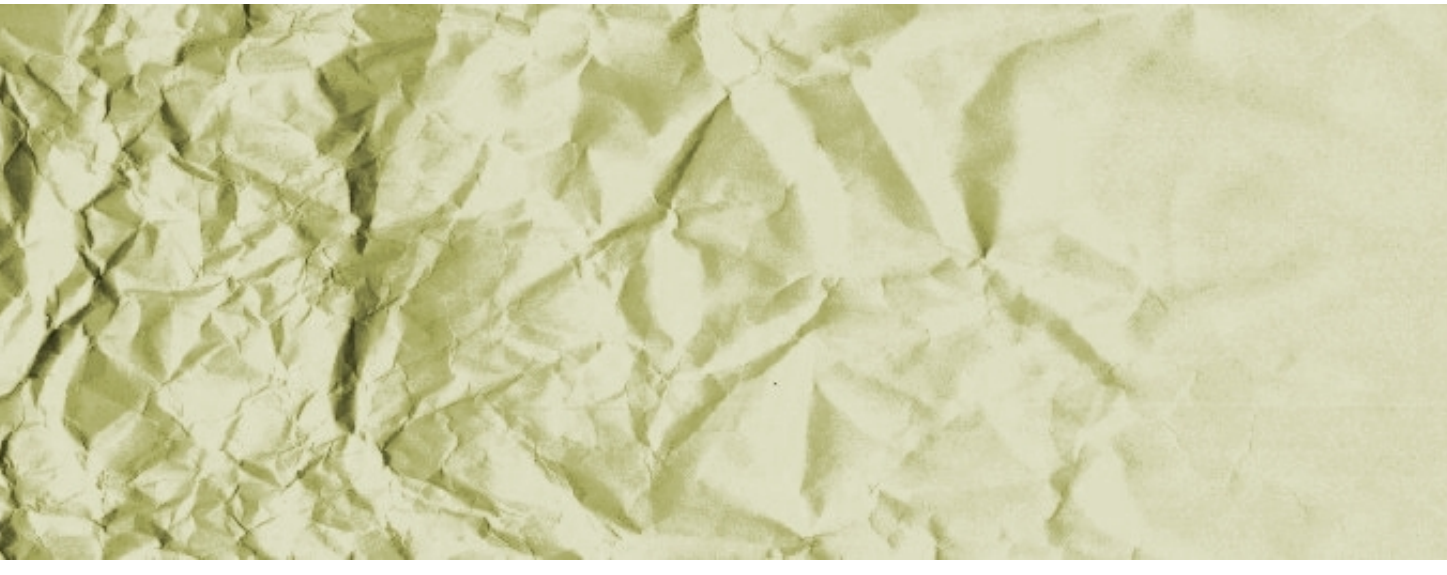
The Measurement of Adolescent Depression

Leigh Roeger



FLINDERS UNIVERSITY INSTITUTE OF INTERNATIONAL EDUCATION
RESEARCH COLLECTION
NUMBER 14

STUDIES IN COMPARATIVE AND INTERNATIONAL EDUCATION
Number 14

The Measurement of Adolescent Depression

LEIGH ROEGER

FLINDERS UNIVERSITY
INSTITUTE OF INTERNATIONAL EDUCATION

Produced by the Flinders University Institute of International Education
Sturt Road, Bedford Park, Adelaide SA 5000

Edited by John P. Keeves and I Gusti Ngurah Darmawan
Designed by Katherine L. Dix
Published by Shannon Research Press, South Australia
ISBN: 1-920736-10-7

Preface

The Center for Epidemiological Studies Depression Scale (CES-D: Radloff, 1977) is one of the most widely used self-report measures of depressive symptomatology. Studies employing the CES-D have generally found that girls on average report higher CES-D scores than boys. These findings appear to indicate that girls experience higher rates of depressive symptomatology but to date researchers have not ruled out the possibility of gender bias (or Differential Item Functioning: DIF) in the CES-D itself. In addition, it is widely believed that differences between schools may be important in terms of shaping student mental health, including depression, but to date there has been very little research addressing this issue.

Using longitudinal data collected from around 2500 students across 26 South Australian high schools over three years (Years 8 to 10: Ages 13 to 15) the present study applies Item Response Theory (IRT: TestGraf), Structural Equation Modelling (SEM: M*plus*) and Hierarchical Linear Modelling (HLM) statistical techniques to test for CES-D gender and school effects. The results from the IRT and SEM analyses indicate that for equivalent levels of depressive symptomatology the following item scores are higher for girls compared with boys: *Bothered* (1), *Appetite* (2), *Blues* (3), *Good* (4), *Sleep* (11), *Cry* (17) and *Sad* (18). Conversely for equivalent levels of depressive symptomatology the following items scores are higher for boys: *Effort* (7), *Happy* (12) and *Unfriendly (15).* Using HLM techniques statistically significant differences in mean CES-D scores are found between schools and these differences increase between Years 8 to 10 consistent with a school effect.

The sizes of the gender and school effects found in the present study are best described as small. Using a latent mean analysis, it is estimated that CES-D gender DIF adds around one half of a point to girls scores. The magnitude of this bias is not sufficient to account for the observed gender differences in the dataset. Although school differences in mean levels of depressive symptomatology are statistically significant, most (98%) variation in CES-D depression scores is found to be at the student level. The implications of these findings for the measurement of adolescent depressive symptomatology with the CES-D and for the types of mental health preventative programs to be provided in schools are discussed.

Acknowledgments

Professor John Keeves, Professor Graham Martin, Dr Steve Allison, Ms Vikki Dadds and my wife (Marilyn) and daughters (Shauna and Caroline) provided all the necessary ingredients for this project.

Contents

Figures

Tables

1

Introduction

> If data are no good, then theorizing based on those data cannot be any good either. For this reason, an important part of every science is the invention, construction, and validation of data-gathering tools. (Funder, 1993, p. 121)
>
> During adolescence, the social environment of the school plays an important role in shaping current and future health. (Sheehan, Marshall, Cahill, Rowling & Holdsworth, 1999, p. 47)

Adolescent depressive symptomatology is a major area of current psychological inquiry and the Center for Epidemiological Studies Depression Scale (CES-D: Radloff, 1977) is a cornerstone to a large part of this research. Many studies have found that adolescent girls, on average, have higher total CES-D scores than adolescent boys. These results have been interpreted as indicating that girls experience higher levels of depressive symptomatology than boys. Several theories have been developed to explain these gender differences and not an inconsiderable amount of research effort has been directed to garnering empirical support for them.

The first key question addressed by the present study is whether CES-D scores obtained from boys and girls across early adolescence can be meaningfully compared. It is important to be clear that the question here does not so much concern whether adolescent boys or girls differ in their mean level of depressive symptomatology - the construct that the CES-D is designed to measure. Rather the issue is whether the CES-D measures depressive symptomatology on the same measuring scale across gender. If it does, then the CES-D would be said to exhibit measurement invariance for this comparison. Without measurement invariance any observed CES-D difference between boys and girls might simply reflect differences in the measurement operation.

The question of CES-D gender measurement invariance is unabashedly a methodological one at its core. It is appreciated that not everyone finds measurement issues to be of tremendous interest but few people openly quarrel with their importance. Questions about invariance in psychological measurement or measurement bias have long been considered (Horn & McArdle, 1992) and have generated much research and debate (Reisse, Widaman & Pugh, 1993). Questions about measurement invariance are important questions because as Funder (1993)

points outs (see introductory quote) bad data leads to bad theorising and for this reason the validation of data gathering instruments is a necessary part of any scientific endeavour.

To her credit the principal developer of the CES-D recognised the fundamental importance of measurement invariance for the CES-D across population subgroups. Radloff stated that: "To compare results from one subgroup to another, the scale must be shown to measure the same thing in both groups" (1977, p. 386). In keeping with this sentiment Radloff investigated whether the CES-D had adequate reliability and validity and a similar factor structure across various population sub-groups defined by age, gender, race (American Whites and Blacks) and level of education. The results from the original CES-D analyses indicated that with some minor exceptions the psychometric properties of the CES-D were similar across the population sub-groups which were examined.

On the basis of her analyses Radloff (1977) concluded that the CES-D was suitable for comparing levels of depressive symptomatology across the groups that had been examined. By today's standards the tests applied by Radloff, while commonplace and acceptable during the 1970s, would not now be considered sufficient to establish that the CES-D was in fact measuring depressive symptomatology equivalently across these groups. Subsequent to the original CES-D reliability and validity study a great many other researchers have also investigated the issue of CES-D measurement invariance across different age ranges, cultures and gender.

The CES-D measurement invariance studies carried out to date have generated a bewildering array of conflicting findings and have generally provided mixed support for the notion of CES-D measurement invariance. Later it is argued in this thesis that the evidence either for or against CES-D gender measurement invariance is weak. This lack of evidence means that despite being used in hundreds of substantive studies, measurement invariance for the CES-D has not yet been demonstrated. It is possible therefore that the observed gender mean CES-D differences and the different gender pattern of correlations found between CES-D scores and external variables are entirely artifactual and without substantive meaning.

This study addresses the question of CES-D measurement invariance using statistical techniques drawn from two very different measurement traditions. The first approach, more familiar to the psychological community, is based around factor analysis and classical measurement theory. A definition of the general problem from this perspective is provided by Horn and McArdle (1992, p. 117) who state that the question of invariance in measurement "… is one of whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute". Reflecting their shared factor analytic background the general framing of the problem by Horn and McArdle is very similar to that adopted by Radloff (see earlier quote) who also emphasised that scales that possess measurement invariance measure the same thing across groups.

The second approach to measurement invariance is based on item response theory (IRT: Lord, 1980). IRT models are routinely used in large scale educational testing programs to examine whether test items function differently in different groups. An impetus for the use of IRT models for this purpose grew out of the United States civil rights era of the 1960s (Scheuneman & Bleistein, 1999). This era was characterised by a unprecedented concern for equal opportunity and attacks on systems, including educational, that were viewed as discriminatory against minority groups and women (Cole, 1993). In this climate the onus was placed on testing authorities to show that group differences in performance arose purely as the result of earlier social disadvantage and were not an artefact of the testing process itself.

In response to social concerns for test fairness and equality, testing authorities in America such as the Educational Testing Service and the American College Testing Program developed a variety of IRT models for detecting biased test items. Underlying most of these methods is the notion that test items are unbiased when they have the same item response function in every group (Lord,1980). This means that people of the same ability or skill have the same probability of getting an item correct regardless of their group membership. In contemporary IRT studies the expression 'differential item functioning' (DIF) has come to replace the term 'item bias'. The term 'DIF' is preferred because it more accurately reflects the broad and neutral question of whether items have the same probability of being correct for comparable groups. The term 'bias' on the other hand suggests unfairness, but this requires a judgement beyond item statistics (Camilli, 1993).

Very few researchers have reviewed and contrasted both CFA and IRT approaches to measurement invariance. A notable exception is Reisse et al. (1993) who argued that the central principle consistently evident in the vast literature on measurement invariance is that psychological measurements are on the same scale or comparable when "… the empirical relations between the trait indicators (e.g. test items) and the trait of interest are invariant across groups" (p. 552). This is a very useful definition and one which will be adopted in the present study. Of course the definition begs the question of how the empirical relations between the test items and trait are to be expressed or calculated. The answer to this question in turn depends on whether factor analytic or IRT methods are used but on the face of it the definition can encompass and provide a framework for both these techniques

During the 1970s and 1980s factor analysis meant exploratory factor analysis and various techniques were proposed for examining the invariance of factor structures across groups (see Reynolds & Harding, 1983). Since the advent of Structural Equation Modelling (SEM) interest in these techniques has declined (Reisse et al., 1993) and Confirmatory Factor Analysis (CFA: Jöreskog, 1969) is now the main factor analytic approach for investigating measurement invariance. Similarly older IRT analyses of measurement invariance relied heavily on the logistic function and were applicable only to test items that were dichotomous (e.g. correct / incorrect). Because not all scale items are modelled correctly with the logistic function and many psychological scales use an ordinal response format, newer methods and techniques have been developed. Both CFA and a modern IRT model are used in the present study to investigate the question of CES-D measurement invariance.

The second key question addressed in the present study is whether schools exert effects on student levels of depressive symptomatology independently of, or in addition to, individual level characteristics. The incidence and prevalence of adolescent depression may be increasing during this century (Lewinsohn, Rohde, Seeley & Fischer, 1993), and since the genetic make-up of the population has not altered substantially (Birmaher et al., 1996) it is possible that the social environment is influencing rates of depression in young people. Schools have a unique role in the lives of children and adolescents and if the increasing rates of depression are socially influenced then schools may be one of the main social agents involved.

It is widely assumed that differences between schools may be important in terms of shaping student mental health – for example, see the introductory quote. In reality there is little empirical research to gauge the importance of the social context for mental health in either young people or adults (Taylor, Repetti & Seeman, 1997) and there is no published scholarly writing specifically addressing the possibility of school effects on student levels of depression. This second question concerning possible school effects on student levels of depressive symptomatology therefore is

quite novel and to date has generated very little empirical research or critical analysis. In this sense the research does not enjoy (or is burdened by) a large literature to guide the analyses. This research will enter uncharted waters and will be more speculative in nature.

General aims of the study

Very broadly the general aims of this study are:

(a) to examine whether the CES-D measures depressive symptomatology equivalently across adolescent boys and girls.

(b) to examine whether schools exert effects on student levels of depressive symptomatology independently of individual level characteristics.

In the course of this study quite a number of subsidiary questions are also addressed. Most of these questions centre around the psychometric properties of the CES-D scale when used with adolescent samples. A detailed account of these questions is provided in Chapter 3.

Significance of the study

The possible lack of measurement invariance in the CES-D is clearly of great concern given that it is a major tool in current epidemiological research. The key contribution of this study is that it examines the measurement properties of the CES-D using modern statistical techniques that have not yet been applied to the CES-D. The application of these modern techniques represents a considerable improvement over previous analyses of CES-D measurement invariance. The present study, which will use data from a large scale community survey of high school students, therefore can provide a much better understanding of whether CES-D scores are in fact comparable across adolescent boys and girls.

The results from this study may provide evidence supporting the measurement properties of the CES-D for gender comparisons in samples of adolescents. If this is the case then the psychological research community can confidently continue to employ the CES-D, without modification, in substantive gender based analyses of adolescent depressive symptomatology. On the other hand if the results indicate that CES-D scores are not comparable across boys and girls, the magnitude of the discrepancy will have been estimated and the items that contribute to this lack of measurement invariance identified.

School effects on student levels of depressive symptomatology is an important issue to the applied psychological community. Schools are places where whole populations of young people can be accessed easily and present ideal opportunities for preventative mental health programs. Because of this, schools are assuming greater prominence for the delivery of child and adolescent mental health services. If school level characteristics are important then the so-called ‘whole of school’ programs directed at changing these may be able to produce improvements to student levels of depressive symptomatology. This study is the first to test specifically for school effects on student levels of depression and the results can be useful for guiding the further development of mental health programs to schools.

2

Literature Review

Depression, gender and adolescence

Clinical depression is by far the most common psychiatric disorder, annually affecting more than 100 million people worldwide (Gotlib, Lewinsohn, Seeley, Rohde & Redner, 1993). Estimates from large scale epidemiological studies of the prevalence of clinical depression vary between countries but life time prevalence rates of between around 5 to 11 per cent have been reported for Westernised countries such as the United States (Robins & Regier, 1991), Canada (Bland, Orn & Newman, 1988) and New Zealand (Wells, Bushnell, Hornblow, Joyce & Oakley-Browne, 1989).

The relatively large numbers of people suffering from clinical depression, and the disabling nature of the disorder, explain why depression has become a major public health issue. In addition, cohort trends appear to indicate that individuals born after the Second World War have a higher risk of developing clinical depression than those born earlier (Lewinsohn et al., 1993; Ryan et al., 1992). The World Health Organisation's global burden of disease research has reported that in terms of overall burden of diseases in the world unipolar major depression was ranked fourth and was predicted to rise to second by the year 2020 (Murray & Lopez, 1996).

Consistent with the overseas literature, similar prevalence estimates for clinical depression and the associated burden have been reported in Australia. The Australian National Survey of Mental Health (Andrews, Hall, Teesson & Henderson, 1999) found that in the 12 months prior to interview 5.8 per cent of the adult population had experienced one or more depressive disorders, with depressive disorder more frequent in women (7.4%) than in men (4.2%). Respondents to this survey without any mental or physical disorders, on average, reported that they had been unable to carry out their usual activities for one day – presumably due to fleeting and minor conditions such as colds or flu. People with depression on the other hand had on average 2.7 days where they were unable to carry out their usual activities.

A striking and now predictable finding from large scale epidemiological studies is the greater prevalence of clinical depression in adult females compared with adult males

(see Wolk & Weissman, 1995 for review). This difference has been found in many countries and across diverse cultures. Among adults, women are diagnosed with depression about twice as frequently as men (Weisman & Klerman, 1985) although this sex ratio might be smaller among the elderly (Bebbington et al., 1998; Jorm, 1987). Initially researchers investigated the possibility that the observed sex difference in rates of clinical depression might be artifactual.

It was hypothesised that women might be more willing to acknowledge affective symptoms as these are more feminine (King & Buchwald, 1982) or that men might be more likely to forget depressive symptoms (Kessler, McGonagle, Swartz, Blazer & Nelson, 1993). While some support was found for these hypotheses it is now widely accepted that at best these explanations could only account for a very small part of the gender ratio discrepancy. Given this, it is nearly unanimously believed that adult gender difference in rates of clinical depression reflect real differences. The major research task then is to develop and test theories which purport to explain the factors responsible for this difference.

A broad range of theories have been developed to explain gender differences in adult rates of clinical depression. These include genetic and biological explanations, social approaches which emphasise women's role in society and finally psychological or cognitive theories about gender differences in personality. Genetic and biological (focussing on female reproductive endocrinology) are appealing because potentially they could explain the universality of the higher rates of female depression across diverse cultures. This evidence is complex and controversial but most reviewers (see Bebbington, 1998; Wolk & Weissman, 1995) have concluded that while it is possible that there are genetic effects in the transmission of depression these effects do not cause the sex difference. Similarly while there is strong evidence that biological factors are involved in the development of depression as yet no sex related biological mechanism has been convincingly identified.

With the failure of genetic and biological explanations much attention has focussed on the role of social and cognitive factors. Women's traditional, and many would argue disadvantaged, role in society is thought to contribute to their relatively high rates of clinical depression. Consistent with this explanation is the finding that the sex difference in clinical depression is most marked in the reproductive years when male and female roles maximally diverge (Wilhelm & Parker, 1990). Interestingly among university students the gender difference in rates of depression is considerably smaller than in the general population (Gladstone & Koenig, 1994). This seems to provide support for the notion that when the roles and status of men and women are comparable then rates of depression will also be similar (Nolen-Hoeksema, 1990) but it may simply reflect that a healthier group of women attend university (Wolk & Weissman, 1995).

Simple versions of the role-strain hypothesis in explaining gender differences in rates of depression generally have not received support (see Weich, Sloggett & Lewis, 2001). Bebbington (1998), however, has argued persuasively that depression is linked to the things that people do and that finer grained study of role-based behaviour may yet provide a better understanding of sex differences in depression. A number of studies have shown a link between stressful life events and negative health outcomes including depression (Kessler, Price & Wortman, 1985; Lloyd, 1980). It is possible that women experience more negative events or are more prone (perhaps because of a lack of social support) to depression in response to negative life events than men. While some studies have found that women experience more stressful events others have not and the evidence regarding access to social support is conflicting (Turner & Marino, 1994).

Psychological explanations for the gender difference in depression have focussed on the response styles of men and women to negative events which cause an initial lowering of mood. Nolen-Hoeksema (1990) has proposed that when faced with depressed mood females are more likely than males to engage in self-focussed ruminative responses and these serve to maintain and amplify the depressed mood. Males on the other hand are more likely to engage in distracting responses which shorten and dampen depressed moods. There is some empirical evidence to support this (see Nolen-Hoeksema, 1991) but it is also possible that the tendency of women to ruminate about negative events is a symptom rather than a cause of their higher rates of depression (Wolk & Weissman, 1995).

While sex differences in rates of depressive disorder have not yet been convincingly explained (Bebbington, 1998) many commentators view adolescence as a crucial developmental stage for understanding both the nature and course of depression and the sex difference in rates of depression (Gjerde & Block, 1996; Petersen et al., 1993). This is because during adolescence prevalence rates of clinical depression increase markedly and the female preponderance of clinical depression is first observed (Angold & Costello, 2001; Wade, Cairney & Pevalin, 2002). The increased focus on depression in young people is in contrast with an earlier dominant consensus during the 1950s and 1960s that children and adolescents lacked the cognitive and affective mechanisms necessary to experience depression (Holsen, Kraft & Vitterso, 2000).

Today, depressive disorders in children and adolescents are viewed as isomorphic to depression in adults (Kutcher & Marton, 1989). Major problems associated with the early onset of clinical depression include the increased risk of suicide (Kovacs, Goldston & Gatsonis, 1993; Ryan et al., 1987) and the costs to social, cognitive and interpersonal development (Altmann & Gotlib, 1988; Coyne, Downey & Boergers, 1992). It is also possible that the first depressive episode may sensitise children and adolescents to future episodes. This is thought to occur because of long-lasting changes in biological processes and an enduring increased responsivity to stressors (Post, 1992; Segal, Williams, Teasdale & Gemar, 1996). For this reason young adults who first experienced clinical depression during childhood would be expected to have a greater number of episodes by any age than their peers with later onset depression (Kovacs, 1997).

Large-scale population studies conducted overseas have reported very wide prevalence rates of between 0.4 per cent and 2.5 per cent in children and between 0.4 per cent and 8.3 per cent in adolescents for clinical depression (see Bird, 1996; Birmaher et al., 1996; Hammen & Rudolph, 1996 for reviews). These estimates are for clinical depression as defined by a categorical diagnostic system such as the *Diagnostic and Statistic Manual of Mental Disorders,* fourth edition (DSM-IV; American Psychiatric Association, 1994) or the *International Classification of Diseases and Health – Related Problems,* tenth revision (ICD-10; World Health Organisation, 1992). Because of important methodological differences in the manner these large scale population studies were carried out, it is unclear whether the wide variation in prevalence rates are of substantive interest (Bird, 1996).

The Child and Adolescent Component of the National Survey of Mental Health and Wellbeing (Sawyer et al., 2000; Sawyer et al., 2001) was recently completed and for the first time provided prevalence estimates of mental health disorders among Australian children (6 to 12 years of age) and adolescents (13 to 17 years of age). Parents were interviewed using the Diagnostic Interview Schedule for Children (DISC: Shaffer, Fisher, Lucas, Dulcan & Schwab-Stone, 2000) which is based on

DSM-IV criteria. Adolescents (but not children) completed the *Child Behavior Checklist* (CBCL: Achenbach, 1991).

Sawyer et al. (2000) identified that the overall prevalence (based on the parent diagnostic interview) of depressive disorders was 3.7 per cent and that children (Boys: 3.7%; Girls: 2.1%) had lower prevalence rates than adolescents (Boys: 4.8%; Girls: 4.9%). In a later report (Sawyer et al., 2001), using a more recent scoring algorithm, the overall prevalence estimate of 3.7 per cent was revised down to 3.0 per cent. A breakdown showing the impact of this revision on the estimates for children and adolescents separately was not shown.

It is widely believed that before adolescence rates of clinical depression are approximately the same for boys and girls but during mid-adolescence (around 15 years of age) girls will experience higher rates than boys (Garrison et al., 1997; Hankin et al., 1998; Lewinsohn, Hops, Roberts Seeley & Andrews, 1993; Nolen-Hoeksema, 1990). The research evidence for this belief is much less robust than is commonly assumed. A review of some of the major epidemiological studies that have examined gender depression differences is provided by Compas (1997) who concluded the emergence of the differences might be limited to a small subgroup of adolescent girls who represent an extreme of the distribution of depressive symptoms among the adolescent population.

Notably the Australian national survey (Sawyer et al., 2000) failed to find significant gender differences in rates of clinical depression among older adolescents. This might have been because the disorder estimates provided in the Australian survey were derived from parent's answers to the DISC. It is generally appreciated that parents provide less sensitive reports of internalising problems in their children compared with externalising symptoms that are more concrete and observable (Herjanic & Reich, 1997; Sawyer, Clark & Baghurst, 1993). This is a key reason large scale adolescent epidemiological studies of mental health usually interview young people directly. It is possible therefore that this methodological artefact in the Australian survey may have reduced the size of the true gender differences in rates of adolescent depression in the sample.

Despite the somewhat inconsistent research results a considerable amount of theorising has been directed to explaining gender differences in depression during adolescence. In many respects these theories and explanations have paralleled those advanced to explain gender difference in adult rates of depression. From a biological perspective given that the emergence of gender differences in depression is believed to coincide with the period of greatest pubertal change researchers have focussed attention on the role of reproductive hormones. The evidence implicating reproductive hormones, however, is mixed and in a key study of psychiatric patients Angold and Rutter (1992) showed that pubertal status did not predict depression scores.

Other researchers, for example Susman et al. (1987), have found that estrogen levels are associated with depressed affect in pubertal girls but it does appear that social factors including negative life events and their interaction with pubertal status (but not hormonal status) account for more of the variance in negative affect than biological factors alone (Brooks-Gunn & Warren, 1989). In this respect more than 40 cross-sectional studies have established a link between stressful events in adolescence and depression (Compas, Ey & Grant, 1993). In addition, longitudinal studies have found that even when initial level of depression is controlled, recent stressful events are associated with increases in depression (Allgood-Merten, Lewinsohn & Hops, 1990; Lewinsohn, Joiner & Rohde, 2001).

Since there is an association between stressful life events and depression it is possible that adolescent girls experience more stressful life events than adolescent boys and this explains the female preponderance of depression. The evidence that adolescent girls experience more stressful life events than adolescent boys, however, is mixed (Petersen, Sarigiani & Kennedy, 1991). For example, Compas, Davis and Forsythe (1985) found that among young adolescents (12-14 years of age) girls did report more negative daily events than boys but this difference was not evident among older adolescents. Interestingly, several studies have found that parental divorce is more likely for girls than for boys during early adolescence (Block, Block & Gjerde, 1986).

It does seem clear that stressful life events have more negative effects on women than men (Kessler & McLeod, 1984), and although this pattern might not be as strong among children and adolescents most studies find that it is girls who show the most negative reactions to life events (Compas, 1987). In particular, adolescent girls appear to experience more interpersonal stressful events and are more likely to respond to interpersonal stress with depressive symptoms (Petersen, Kennedy & Sarigiani, 1991). Researchers have also examined the role of gender-related developmental challenges during adolescence and a number of possible contributing stressors have been identified.

Girls are more likely than boys to go through puberty before or during the transition to secondary school and the synchronicity of these two changes has been suggested as a particular stressor to girl's adjustment (Seiffge-Krenke & Stemmler, 2002; Simmons, Burgeson, Carlton-Ford & Blyth, 1987). Changes to physical appearance associated with puberty may lead to poorer satisfaction with body image in girls compared with boys (Petersen, Sarigiani, & Kennedy, 1991). This may be particularly important given that physical appearance is a key factor in global self-esteem among adolescents (Cairns, McWhirter, Duffy & Barry, 1990), and poor self-esteem is closely associated with depression (Beck, 1976).

Under the gender intensification hypothesis (Hill & Lynch, 1983) femininity correlates positively with depressed mood. Early adolescence marks the beginning of gender identity formation, and the pressure for girls to behave in less masculine and more feminine ways therefore may explain the emergence of gender differences in depression around this time. Cross-sectional studies, however, have found that the correlation between femininity and depressed mood is low (WichstrØm, 1999) and two longitudinal studies (Allgood-Merten et al., 1990; Petersen, Sarigiani & Kennedy, 1991) failed to establish an effect for changing levels of femininity on depressed mood.

It is clear that further studies are required in order to understand better possible gender differences in adolescent depression. These studies are required because adolescent depression is a major public health issue in its own right and also because an increased knowledge of the antecedents of depression in early life may contribute to an improved understanding of depression in later life (Allgood-Merten et al., 1990). This research depends on valid and reliable measures of depression in adolescents samples.

To date, and with some confidence it can be predicted that in the future, the majority of adolescent depression research studies will be based on data collected from self-report depression scales. Self-report depression scales are the cornerstone of adolescent depression research and it is to the measurement of depression and the gender differences with these instruments to which attention is now given.

Depression scales and gender differences in adolescent samples

Paper and pencil self-report questionnaires are the primary research methods for the assessment of depression in children and adolescents. Although more than 30 self-administered scales for the measurement of depression have been developed (Moran & Lambert, 1983) only a handful have achieved widespread use (Gotlib & Cane, 1989). The most prominent of these include the Beck Depression Inventory (BDI: Beck, Ward, Mendelson, Mock & Erbaugh, 1961), the Children's Depression Inventory (CDI: Kovacs, 1992), and the CES-D.

The BDI was designed for adults and appears to be particularly suited for use with clinical populations (Beck, Steer & Garbin, 1988). The CDI was specifically developed for children and is considered to be a downward revision and extension of the BDI (Reynolds, 1992). The CES-D was developed for use in community samples of adults but has been widely used with adolescents. These three scales would be the most frequently used and well-validated self-report measures of depression. In general, self-report depression scales share several common features.

Self-report depression scales are fast and relatively inexpensive to implement in large groups and are an economical method for establishing overall mean levels of depressive symptomatology in a population. Typically these scales are not designed to provide a discrete diagnosis of depression but rather they aim to measure depression as a single dimension of psychopathology which cuts across a wide variety of diagnostic categories (Gotlib & Cane, 1989). The emotions or symptoms listed reflect the central features of depressive disorders but will nearly always include depressed mood (the presence of unhappiness or sadness - the affective component of depression) along with other symptoms of depression. Scores from these scales therefore are not pure indexes of depressed mood (Compas et al., 1993) and are generally referred to as indicating levels of depressive symptomatology.

In children and adolescents high levels of depressive symptomatology predict later clinical depression (Kovacs, Feinberg, Crouse-Novak, Paulauskas & Finkelstein, 1984; Kovacs, Feinberg, Crouse-Novak, Paulauskas, Pollack & Finkelstein, 1984; Prescott et al., 1998) and difficulties in psychological adjustment and interpersonal functioning in adulthood (Kandel & Davies, 1986). High levels of depressive symptomatology is the best predictor of referral status for child and adolescent psychiatric treatment (Achenbach, 1991; Petersen et al., 1993) and are associated with using school counselling services (Mattison, Handford, Kales, Goodman & McLaughlin, 1990). High levels of depressive symptomatology in children impair academic performance (Nolen-Hoeksema, Girgus & Seligman, 1992) and increase the risk for dropping out of school (Kovacs, 1989), suicidal ideation (Garland & Zigler, 1993; Kandel, Raveis & Davies, 1991) and suicidal behaviour (Cole, 1989; Lewinsohn, Rohde & Seeley, 1994; Reinherz et al., 1993).

In a large community sample of adolescents Gotlib, Lewinsohn and Seeley (1995) showed that adolescents who had scored at a high level on the CES-D but did not meet the criteria for clinical depression (so called 'false positives') were far from a normal group. In fact, the false positives did not differ significantly from the true positives (high CES-D scores and a diagnosis of clinical depression) across a wide range of measures of psychosocial dysfunction. This implied that these individuals were experiencing significant levels of distress and impairment. These findings are complementary to research with adult samples which has shown that even in the absence of clinical depression, symptoms of depression can cause impairment and

that the majority of disability in a population is attributable to depression among persons who do not meet the criteria for clinical depression (Horwath, Johnson, Klerman & Weissman, 1994; Wells, Burnam, Rogers, Hays & Camp, 1992).

It is widely assumed that during adolescence overall levels of depressive symptomatology increase markedly from childhood and that female adolescents have significantly higher levels of depressive symptomatology than male adolescents. Consistent with this view three major and well executed longitudinal studies (Ge, Lorenz, Conger, Elder & Simons, 1994; Petersen, Sarigiani & Kennedy, 1991; WichstrØm, 1999) found that gender differences in overall levels of depressive symptomatology emerged around the ages of 13 to 15 years and occurred primarily as a result of adolescent girls reporting higher levels of depressive symptomatology than adolescent boys.

On the other hand, reviews of what could be loosely termed the 'depressed affect in adolescent populations' literature have provided mixed support for gender effects. Petersen, Compas and Brooks-Gunn (1992) examined 30 studies and found that two of the 13 studies to test for gender effects found no significant differences. In a later review, Leadbeater, Blatt and Quinlan (1995) found that six studies from 21 to use the BDI, CDI or CES-D showed no sex differences and one study found higher levels for boys rather than girls. In a compilation of CDI studies, Kovacs (1992) noted that the research literature related to CDI gender differences had been inconsistent. In support of this assessment, references were provided to 15 studies that were evenly divided between those which had found boys to score higher than girls, girls to score higher than boys, and finally studies that showed no difference between boys and girls.

In a recent large scale meta-analysis, Twenge and Nolen-Hoeksema (2000) examined the pattern of gender differences in CDI scores. Using 310 samples of children and adolescents (ages 8 to 16 years; N = 61,424) these authors confirmed that prior to age 13 years there was not a statistically significant gender CDI difference, but beginning in the adolescent years (from age 13), because of the higher scores from girls, there was a statistically significant gender difference. The size of this gender difference however was not large. The effect sizes reported by Twenge and Nolen-Hoeksema (2000) varied by age (13: 0.08; 14: 0.22; 15: 0.22; 16 0.18) but for adolescents (13-16 years of age) overall, the effect size was 0.16. The authors noted that this relatively small effect size was in contrast to studies which had found larger gender differences in selected samples of high scoring adolescents. Twenge and Nolen-Hoeksema (2000) recommended that future research should examine the magnitude of gender depression differences at different levels of severity.

One possible explanation for the somewhat inconsistent gender results is that the effect size for gender might be quite small. In a large school based community study of adolescents designed to provide normative data for the CDI, Chartier and Lassen (1994) found that although females showed statistically higher CDI scores than males the magnitude of this effect was trivial. In addition, no gender differences were found in the proportions of adolescents scoring above a cut-point score of 19 indicating that the higher scores for females cluster more in the moderate rather than the severe range. These authors suggested that because CDI studies typically employ large samples even trivial effects can produce a statistically significant difference and this may lead to inaccurate interpretations.

A further possible reason for the inconsistent gender effects for depression scale scores is that inappropriate statistical techniques have been applied. Nearly all prior analyses of depression scales have used statistical techniques that rest on the

assumption that scores on individual items are normally distributed. This issue is discussed later in this thesis but for now it can be noted that distributional assumptions have not been met in most previous analyses. According to Angold, Erkanli, Silberg, Eaves and Costello (2002, p. 1060) this deficiency may account for the inconsistent gender findings in the adolescent depression scale literature:

> Little attention has been paid to whether normal distributional assumptions were met in other studies that have examined depression scores in this age range [8 to 17 year olds], and ANOVA approaches are not very robust to situations in which the distributions of scores are very severely skewed. Lack of attention to the distributional properties of the data may also, therefore, have contributed to the heterogeneity of findings in the literature.

At the time of writing no systematic analysis of the CES-D literature with respect to gender differences in adolescent samples had been carried out. In view of this, a literature review was conducted of studies that had used the CES-D in adolescent samples. Initially studies were located using PSYLIT and MEDLINE searches for the period from 1977 to the year 2000 inclusive. When these source papers were obtained they were carefully checked for references to other CES-D studies in adolescent samples. Just over 60 studies were located in this manner and these are summarised in Appendix A. These studies would constitute a very large proportion of all published material relating to the CES-D in adolescent samples. The overwhelming majority of the studies were carried out in the United States with high school students.

The provision of a precise estimate of mean gender CES-D differences found in adolescent samples is not simple. First, in quite a number of papers the actual gender means were not reported. Second, a number of authors have published several CES-D papers but in many cases these were no more than the re-analysis of different aspects of previously collected and published data. Clearly data collected from one group of students should only count once when calculating an estimate for mean CES-D gender differences. Third, some CES-D studies were longitudinal and it was not obvious which time point should be taken to examine gender mean differences. Taking these complexities into account 18 studies from the total of 63 possible candidates were selected to examine CES-D mean gender differences in adolescent samples. The selected 18 studies are summarised in Table 1.

The selected 18 studies do not include repeated entries for the same sample of students and where longitudinal data are involved only data collected at the first time point are shown. The first time point was chosen because there is a possibility that repeated administrations of the CES-D may reduce the size of observed gender differences (Aseltine, Gore & Colten, 1998). It is appreciated that in selecting these studies some arbitrary decisions were made and that other researchers might have chosen a slightly different set of studies. The full list of studies is included in Appendix A and shows that a different selection of studies would do little to alter the general picture of CES-D mean gender differences reported in Table 1.

As far as can be ascertained only one study (Allison, Roeger, Martin & Keeves. 2001) has published CES-D data for a community sample of Australian adolescents. In Allison et al. (2001) the present author and colleagues found that girls on average reported higher mean CES-D scores than boys. In addition, the proportion of girls reporting suicidal ideation was higher than for boys. CES-D scores were strongly associated with the risk of suicidal ideation with increasing levels of depressive symptomatology correlated with a greater risk of suicidal ideation.

Table 1 CES-D gender means and effect size in adolescent samples

Study	Year	Sample	Boys Mean (SD)	Girls Mean (SD)	Effect Size
Tolor & Murphy	1985	U.S. (Connecticut) N = 285, Age: 13-17.	14.66 (9.11)	18.48 (10.81)	0.35
Doerfler, Felner, Rowlison, Raley & Evans	1988	U.S. (rural Southern) N = 1207, Grade: 8.	15.43 (9.55)	18.41 (11.94)	0.24
Gjerde, Block & Block (Modified CES-D)	1988	U.S. (California) N = 106, Age: 18.	19.77 (10.75)	22.50 (11.10)	0.25
Garrison, Schluchter, Schoenbach & Kaplan (Whites only)	1989	U.S. (North Carolina) N = 677, Age: 12-15.	15.21 (9.24)	16.54 (9.40)	0.14
Garrison, Jackson, Marsteller, McKeown & Addy (Whites only)	1990	U.S. (Southeast) N = 550, Age: 12-13.	13.98 (8.52)	15.80 (9.58)	0.19
Manson, Ackerson, Dick, Baron & Fleming (American Indians)	1990	U.S. (Southeastern) N = 188, Age: 15-17.	16.7 (8.0)	21.7 (10.0)	0.50
Garrison, Jackson, Addy, McKeown & Waller (Diagnostic sample)	1991c	U.S. (South Carolina) N = 226, Age: 12-14.	18.90 (11.20)	26.84 (13.24)	0.60
Gjerde & Block (Modified CES-D)	1991	U.S. (California) N = 106, Age: 16.	19.77 (10.75)	22.50 (11.10)	0.25
Roberts, Lewinsohn & Seeley	1991	U.S. (Oregon) N = 1710, Age: 16.	15.70 (10.5)	18.12 (10.5)	0.23
Avison & McAlpine	1992	Canada N = 306, Age: 17.	15.45 (9.82)	18.98 (11.86)	0.30
Berganza & Agular (Modified CES-D)	1992	Guatemala N = 339, Age: 15.	15.69 (7.43)	20.78 (8.91)	0.57
Clarke, Hawkins, Murphy & Sheeber (Estimated values)	1993	U.S. (Oregon) N = 513, Age: 15.	14.74 (10.6)	19.36 (13.1)	0.35
Gore, Aseltine & Colten	1993	U.S. (Boston) N = 1208, Grade: 9-11.	11.2 (7.5)	14.6 (9.1)	0.37
Sheeber, Hops, Alpert, Davis & Andrews.	1997	U.S. (Oregon) N = 421, Age: 16.	15.40 (9.80)	17.92 (11.17)	0.23
Windle & Windle	1997	U.S. (New York) N = 975, Age: 15.	13.65 (9.47)	16.11 (10.55)	0.23
Gjerde & Westenberg (Modified CES-D)	1998	U.S. (California) N = 106, Age: 18.	19.77 (10.75)	22.50 (11.10)	0.24
Marcotte	1996	Canada (Quebec) N = 349, Age: 11-18.	13.15 (8.69)	18.18 (10.98)	0.43
Allison, Roeger, Martin & Keeves	2001	Australia (SA) N = 2489 Age: 13	11.4 (9.0)	13.9 (11.6)	0.22

This association was found for both boys and girls but it was also shown that the risk of suicidal ideation at moderate levels of depressive symptomatology was much greater for girls than boys.

It can be readily seen from Table 1 in all studies adolescent girls consistently obtained higher mean CES-D total scores than boys. In fact no study could be identified in which adolescent boys were reported to have higher mean CES-D scores than adolescent girls. A gender effect size was calculated by taking the difference between the boy group mean and the girl group mean and dividing this difference by the girl group standard deviation. This effect size can be interpreted as the degree, in standard deviation units, that the average girl was different from the average boy. In calculating an effect size there are statistical advantages to using the pooled standard deviation as the divisor (McGaw & Glass, 1980) but in many of the reviewed CES-D studies the pooled standard deviation was not reported and hence the choice of the girl group standard deviation.

Cohen (1977) suggests that an effect size of 0.20 represents a small difference between the groups, 0.50 represents a moderate effect and 0.80 a large effect. Using these criteria in the studies reviewed the effect sizes ranged between 0.19 and 0.60 but for the most part they were in the order of between 0.20 and 0.40. Effect sizes between 0.20 and 0.40 would be categorised as small. In a few studies with unusual samples the gender effect was much larger. These samples included a diagnostic group of high scoring adolescents and a small group of adolescents from Guatemala using a modified version of the CES-D.

In summary, this review of the published literature indicates that adolescent girls, on average, consistently report higher mean total CES-D scores than adolescent boys but the size of this gender effect is small. This finding is in contrast with studies using other questionnaires such as the CDI or the BDI that have tended to produce more conflicting evidence about gender differences (Angold, & Costello, 2001).

In an interesting test of mean gender CES-D total score differences Greenberger, Chen, Tally and Dong (2000) speculated that while being female was a significant risk factor for depression in the United States of America, decades of centralised political and ideological effort had been directed towards eliminating gender inequalities in China. It was hypothesised that among Chinese adolescents gender might not be such a risk factor for elevated levels of depressive symptomatology. Using data derived from samples of Chinese and United States adolescents Chinese girls reported significantly higher CES-D scores than their male counterparts, but the size of this effect was smaller than that observed in the United States sample. Unfortunately, gender means are not reported in the paper and for this reason the study is not included in Table 1 but it is summarised in Appendix A.

While it is clear that the average adolescent girl reports a higher CES-D total score than the average adolescent boy, few researchers have examined in detail the distribution of CES-D scores by gender. In the majority of studies, where it is reported, a higher proportion of females have been identified as ‘high scoring cases’ compared with males. This suggests that the higher female CES-D average score arises at least in part because female scores cluster more in the severe range than do male scores. There are exceptions to this, however. For example Wells, Klerman and Deykin (1987) found no differences in the proportion of males and females scoring above 16. Much depends of course on what cut-point is employed to categorise high scoring cases with the further complication that in some studies different cut-points are used for males and females.

The CES-D gender studies reviewed in this section are based on comparisons of raw (or observed) mean total CES-D scores between males and females. In the studies that examined whether these mean gender differences were statistically significant simple procedures (such as *t* tests or analysis of variance) were performed. These simple statistical procedures have traditionally been widely used in psychological research for group mean comparisons. More recently group comparisons between latent means through structural equation modelling (SEM) has been recommended as a better and more flexible approach (Cole, Maxwell, Arvey & Salas, 1993).

Traditional analyses are limited because observed means may be contaminated with measurement error and the adequacy (reliability and validity) of the measure cannot be evaluated (Li, Harmer & Acock, 1996). In addition, and most importantly in the context of present study, latent mean analysis allows item bias to be detected and controlled for in the comparison of latent means (Byrne, Shavelson & Muthén, 1989; Hoyle & Smith, 1994).

Until relatively recently performing a latent mean analysis through SEM was very difficult. The older discussion papers of the methodology for structured means analyses are very technical in nature (Aiken, Stein & Bentler, 1994) and early versions of SEM software (which were less than user friendly to begin with) required complex additional programming (Byrne, 1998). With the publication of several explanatory substantive papers (see Schaie, Maitland, Willis & Intrieri, 1998) and improvements to SEM software, latent mean analyses are now able to be performed by researchers relatively easily. To date there has not yet been a latent mean analysis of CES-D gender differences and this is in fact one of the subsidiary aims of the present study.

In the next few sections of this literature review a detailed examination of the CES-D and its psychometric properties are presented in order to understand better CES-D gender differences and how they might arise. This review outlines the development of the scale and summarises the quite considerable number of basic reliability, validity and exploratory factor analyses that have been performed on the CES-D. Attention is particularly focused on the use of the CES-D in adolescent samples. Following this, and reflecting the fact that this study is not the first to investigate CES-D gender differences, material relating to these differences at the item or sub-scale level is reviewed.

The development of the CES-D scale

The CES-D scale was developed by researchers at the National Institute of Mental Health in the United States for use in studies of the epidemiology of depressive symptomatology. The CES-D developers intended that the scale would enable the targeting of treatment programs by accurately identifying groups at high risk of depression and that it would prove a valuable tool in studying the relationship between depressive symptomatology and other variables. Since its inception the CES-D has become one of the most frequently used self-report measures of depression (Gotlib & Cane, 1989). A MEDLINE and PSYLIT search in refereed journals published between the years 1990 and 1999 found over 540 different studies using the key-words ‘CES-D’ and ‘CES_D’.

The CES-D has been widely used to compare levels of depressive symptomatology across groups (males and females, young and elderly, etc.) and also to examine whether levels of depressive symptomatology have different correlates across groups,

for example, the relationship between family functioning and depressive symptomatology might vary between males and females. It has been used as an outcome measure in prevention programs (see Clarke et al., 1992; Clarke, Hawkins, Murphy & Sheeber, 1993; Lewinsohn, Clarke, Hops & Andrews, 1990; Peden et al., 2000).

A version of the CES-D for children has been developed (the CES-DC: Weissman, Orvaschel & Padian, 1980) although this is not widely used (but see Blatt, Hart, Quinlan, Leadbeater & Auerbach, 1993, for one example). In addition, a short form (10 items) of the scale has been developed specifically for screening for clinical depression (CESD-10: Andresen, Malmgren, Carter & Patrick, 1994; Boey , 1999).

The CES-D scale comprises 20 items selected from previously validated depression scales such as the Beck Depression Inventory (BDI: Beck et al., 1961), the Zung Depression Scale (Zung, 1965) and the Minnesota Multiphasic Personality Inventory Depression Scale (MMPI: Hathaway & McKinley, 1940) to represent the major components of depressive symptomatology. Sixteen items were chosen to cover depressed mood, feelings of guilt and worthlessness, feeling of helplessness and hopelessness, psychomotor retardation, loss of appetite and finally sleep disturbance. In addition, four positively worded items were included to break a tendency toward a response set and to measure positive affect.

The CES-D items are listed below together with the abbreviations that are used in the present study shown in parentheses. The positively worded items are identified by an asterisk. In the text, CES-D item abbreviations are shown with italics and factor names are capitalised.

1. I was bothered by things that don't usually bother me. (*Bothered*)
2. I did not feel like eating; my appetite was poor. (*Appetite*)
3. I felt that I could not shake off the blues even with help from my family or friends. (*Blues*)
4. I felt that I was just as good as other people. (*Good*) *
5. I had trouble keeping my mind on what I was doing. (*Mind*)
6. I felt depressed. (*Depress*)
7. I felt that everything I did was an effort. (*Effort*)
8. I felt hopeful about the future. (*Hopeful*) *
9. I thought my life had been a failure. (*Failure*)
10. I felt fearful. (*Fearful*)
11. My sleep was restless. (*Sleep*)
12. I was happy. (*Happy*) *
13. I talked less than usual. (*Talk*)
14. I felt lonely. (*Lonely*)
15. People were unfriendly. (*Unfriendly*)
16. I enjoyed life. (*Enjoy*) *
17. I had crying spells. (*Cry*)
18. I felt sad. (*Sad*)
19. I felt that people dislike me. (*Dislike*)
20. I could not get going. (*Getgoing*)

The CES-D asks respondents to indicate the frequency with which he or she experienced each item during the past week by checking one of four alternatives: rarely or none of the time (less than 1 day), some or a little of the time (1-2 days), occasionally or a moderate amount of the time (3-4 days), or most or all of the time (5-7 days). In scoring the CES-D the responses (scored from 0 to 3, with positive items reversed) of all 20 CES-D items are summed to produce a total CES-D score with a potential range between 0 and 60. Several points about the CES-D response format and scoring can be noted.

The CES-D response format emphasises the frequency that symptoms are experienced in the previous week in contrast to the BDI or the CDI which asks respondents to endorse graded statements reflecting different degrees of severity for each of the items. CES-D item means therefore are a weighted average of both the frequency of a symptom and the duration of the symptom among those who experience it. Consequently identical item means can be obtained by many subjects experiencing the item briefly or by a few subjects experiencing the symptom for a longer period of time (Wells et al., 1987).

The summary total CES-D score reflects both the number and strength of the endorsement of items and similar scores can be obtained from the strong endorsement of a few items or the weaker endorsement of more items (Fechner-Bates, Coyne & Schwenk, 1994; Wells et al., 1987). Because all items are all given the same weight the CES-D scoring system assumes that each item is equally informative of depressive symptomatology. It is also assumed that each item is uniformly effective across all levels of depressive severity and that the intervals among options (e.g. between Option 0: rarely or none of the time and Option 1: some or a little of the time) are psychologically identical (Santor, Zuroff, Ramsay, Cervantes & Palacios, 1995) This means that in scoring the CES-D, differences between options are taken to reflect an interval scale and the ordinal nature of the response format is disregarded.

Commensurate with the fact that the CES-D is one of the most frequently used self-report depression scales over the past 20 years quite a considerable amount of attention has been directed to its psychometric properties. In the following sections CES-D reliability, validity and factor studies are reviewed. This background provides the foundation to the literature which has examined CES-D gender differences at the item or factor level. Not surprisingly given the importance of understanding gender depression differences and the key role of the CES-D, several investigators have studied whether the CES-D exhibits measurement invariance across gender. These studies are critically reviewed later and it is argued that on the available evidence it is unclear whether the CES-D provides an unbiased measure of gender differences when used in adolescent samples.

CES-D internal properties

In order to test the measurement properties of the CES-D, in the original reliability and validity study, Radloff examined the internal consistency, distribution of scores, test-retest correlations and item factor loadings in three culturally diverse adult samples. In all samples the coefficient alpha was above 0.80 indicating that items appeared to be measuring a single underlying dimension. The distribution of scores obtained in the community samples were positively skewed (low scores observed more commonly than high scores - indices ranging from 1.50 to 1.69) and groups with higher means also tended to have higher variances. Test-retest reliabilities were

found to be in the order of 0.50 across various retesting intervals of two, four, six and eight week intervals.

Researchers have generally viewed the basic psychometric properties of the CES-D favourably (but see Reynolds, 1992, for an exception). A considerable number of studies have found that the internal reliability of the CES-D as measured by Cronbach's coefficient alpha exceeds recommended minimum standards. A follow-up study of the CES-D by its developer in a sample of adolescents and young adults also obtained high internal consistency estimates for the scale as a whole as well as for the four sub-scales (Radloff, 1991). In two well known studies using large scale community samples of adolescents Garrison et al. (1991a) reported alphas for boys of 0.84 and for girls 0.89 while Roberts, Andrews, Lewinsohn and Hops (1990a) reported coefficients for boys of 0.88 and for girls of 0.91.

The highly skewed distribution of CES-D scores produced in community samples resembles a reverse J curve. This indicates a preponderance of low scores and the endorsement of a large number of Option 0 item responses (Chapleski, Lamphere, Kaczynski, Lichtenberg & Dwyer, 1997; Knight, Williams, McGee & Olaman, 1997). Radloff (1977) warned that because of the skewed distribution and the fact that groups with higher means tended to also have higher variances, that standard parametric statistical tests with CES-D data would not be exact.

References to these statistical caveats in the CES-D literature are extremely rare and by and large researchers appear to have accepted the skewed distribution of CES-D scores on the grounds that the prevalence of depressive disorders (which would be reflected in elevated total CES-D scores) is usually less than 10 per cent (Devins & Orme, 1985). Alternatively as noted by Hertzog, Van Alstine, Usala, Hultsch and Dixon (1990) the CES-D might not be sensitive to lower levels of depressive symptomatology.

Test-retest reliability concerns the extent to which a test yields equivalent measurement across time assuming that the underlying construct has not changed in the testing interval. Nunnally (1978) suggests that these values should exceed 0.80. Test-retest studies of the CES-D have produced results consistent with those shown in the original Radloff (1977) study and lower than the recommended standard. In adolescent (and most adult) samples these reliabilities have been found to be around 0.50 – 0.60 across retesting intervals of one month (Andrews, Lewinsohn, Hops & Roberts, 1993; Garrison et al., 1990; Lasko et al., 1996; Roberts et al., 1990a; Roberts, Lewinsohn & Seeley, 1991), three months (Garrison et al., 1989), six months (Tolor & Murphy, 1985), and one to two years (Garrison et al., 1990).

One explanation for the low CES-D test-retest values is that the instrument focuses on state depression rather than chronic depression (i.e. the CES-D asks about symptoms experienced only over the previous week) (Gjerde, 1995). On this view of the CES-D, the scale is measuring a labile state and the observed instability in scores should not be taken to indicate that the measure is unreliable or invalid (Hertzog & Nesselroade, 1987). It is difficult to determine from traditional test-retest analyses whether the marginal temporal stability of the CES-D is due to real change or to the unreliability of the CES-D. Even so, reliabilities in the range of 0.50 suggest that responses of individuals to the 20 items tend to be quite volatile (Roberts et al., 1990a). This appears particularly the case for adolescents where depressive symptoms may be more transient and fluctuating than compared to adults (Garrison et al., 1990).

The CES-D was originally designed for use in the general adult population but it has also been widely used with adolescents. In order to investigate whether the CES-D is

reliable when used with adolescents, Wells et al. (1987) compared the symptom profiles of a sample of college students (aged between 16 to 19 years) with profiles obtained from adult community samples. The means for each item were ranked for both groups and found to be significantly correlated. For most items the differences in ranks for adults and adolescents were less than five suggesting that the CES-D was performing similarly in both populations. The response patterns of high and low scoring subjects were also compared. The mean value of all items were higher for the high scoring subjects implying that differences between low and high scoring subjects was more a matter of the amount or duration of symptoms rather than a difference in the specific pattern of symptoms.

Overall the results from Wells et al. (1987) support the use of the CES-D in adolescent samples. Using a similar methodology, and also based on a review of the literature, Roberts et al. (1990a), in a widely quoted study, reached this same conclusion, namely, that the CES-D appears as equally reliable in adolescent and adult samples. Roberts et al. also examined these basic internal consistency statistics across gender and found no consistent or dramatic effects. In addition, data on the mobility of symptoms (the change in symptoms over the one month follow-up) did not reveal any meaningful gender differences.

CES-D validity studies

CES-D validity studies address the issue of what the CES-D measures – and how well it measures it. In the literature CES-D scores are variously referred to as indicating levels of 'depression', 'depressive symptomatology', 'depressed mood', 'depressive symptoms', 'depressive tendencies', 'depression severity', 'dysphoria', 'psychological distress', 'emotional distress', 'subjective well being' and 'generalised psychopathology'. Justification for the use of these terms is rarely provided, and to the consternation of influential commentators (see Kendall, Hollon, Beck, Hammen & Ingram, 1987; Kendall, Cantwell & Kazdin, 1989) these terms are often used as if they were interchangeable.

For present purposes it is not necessary to provide detailed argument regarding the best choice of terms and it is appreciated that there are differing opinions and an inconsistent use of terminology for self-report depression scales (Haaga & Solomon, 1993). Conscious of not contributing to the so-called 'jangle problem' in psychology in which identical constructs are given different names (Krueger & Finger, 2001) on the basis of convention, CES-D scores in the present study are referred to as reflecting levels of 'depressive symptomatology'. A brief review of CES-D studies that address what the CES-D measures is presented below, organised around bolded headings identifying the various types of validity.

Content validity. The extent to which items correspond to the content of the theoretical concept the scale is designed to measure is referred to as 'content validity'. Since the CES-D items were selected to represent the major components of depressive symptomatology, the scale is based on symptoms of depression as seen in clinical cases (Radloff, 1977). Despite this seemingly straightforward approach it can easily be demonstrated that CES-D items are not limited exclusively to the construct of clinical depression.

Gotlib and Cane (1989) argue that only 10 CES-D items measure clinical depression, four items measure symptoms common to both depression and anxiety, one item measures anxiety and five items are unrelated to either depression or anxiety (see also

Roberts, Rhoades & Vernon, 1990b). The exact manner in which Gotlib and Cane (1989) calculated which CES-D items were allocated to which disorder is not provided but it is not difficult to deduce and is probably as follows:

CES-D items measuring depression: *Blues* (3), *Depress* (6), *Happy* (12), *Enjoy* (16), *Cry* (17), *Sad* (18), *Appetite* (2), *Good* (4), *Failure* (9) and *Hopeful* (8);

CES-D items assessing symptoms common to both depression and anxiety: *Mind* (5), *Sleep* (11), *Effort* (7) and *Getgoing* (20);

CES-D items measuring anxiety: *Fearful* (10);

CES-D items unrelated to either depression or anxiety: *Bothered* (1), *Talk* (13), *Lonely* (14), *Unfriendly* (15) and *Dislike* (19).

Convergent validity. Considerable evidence supports the convergent validity (the degree to which scores from a scale are in agreement with those provided by other measures of the same or a related construct) of the CES-D. Radloff (1977) reported some initial data which indicated strong correlations between CES-D scores and a number of other scales designed to measure symptoms of depression (e.g. the *Lubin Depressive Adjective Checklist*; Lubin & Himelstein, 1976).

Subsequent researchers have also reported strong correlations between the CES-D and the CDI in children and adolescents (Doerfler, Felner, Rowlison, Raley & Evans, 1988), the BDI (Masten, Caldwell-Colbert, Alcala & Mijares, 1986; Roberts et al., 1991; Soler et al., 1997), the Hamilton Rating Scale (Fava, 1983) and the Geriatric Depression Scale (Chan, 1996). Furthermore CES-D scores have been shown to be correlated with measures of self-competence, self-concept, locus of control, loneliness and self-efficacy (Doerfler et al., 1988; Gore & Aseltine, 1995).

Discriminant validity. The evidence that the CES-D measures only depressive symptomatology and not other constructs is not encouraging (Gotlib, 1984; Gotlib & Cane, 1989; Roberts & Vernon; 1983; Vernon & Roberts, 1981). In a key study Gotlib (1984) administered seven questionnaires (comprising 17 different pathology scales) to a sample of university students. All the scales were found to be highly inter-correlated and a factor analysis suggested that the scales measured a unitary factor most appropriately termed ‘general psychological distress’.

Further evidence was provided by Weissman, Sholomskas, Pottenger, Prusoff and Locke (1977) who found that CES-D scores were significantly correlated with the Symptom Check List-90 which is a general screening measure that assesses nine independent psychiatric domains. Orme, Reis and Herz (1986) administered the CES-D along with instruments measuring self-esteem and state-trait anxiety and found that many CES-D items had substantial correlations with these three related but putatively distinct constructs. These authors also concluded that the CES-D does not measure solely depression – and that caution should be exercised in using the CES-D in the evaluation of interventions designed to alleviate depression because of the threat that this may pose to the construct validity of the evaluation.

The failure of the CES-D to distinguish between clinical depression and generalised anxiety was also noted by Breslau (1985) who found that the strength of association between CES-D scores and depression and anxiety were similar and that the two disorders tended to have an additive effect on the CES-D. Vernon and Roberts (1981) reported that the correlation between CES-D scores and the Demoralisation Scale (a broad measure of psychological distress that measures anxiety, depression, self-esteem and hopelessness) was as high as possible given the reliabilities of the two instruments.

In a thoughtful discussion of the lack of CES-D discriminant validity, Roberts and Vernon (1983) suggested that while the symptom items included in the CES-D were typical of depression they were also the most prevalent manifestations of psychological distress that would be common to many types of psychiatric disorder. They argued that most self-report depression instruments (including the CES-D) measured the same thing and this phenomenon was best described as non-specific psychological distress. These difficulties highlight the fact that establishing the validity of self-report depression scales has been challenging because external criteria for validation have been difficult to identify (Compas, 1997).

Criterion validity. Studies examining the criterion validity of the CES-D have sought to test whether the CES-D can discriminate between groups formed on the basis of other independent criteria. Using a concurrent approach to criterion validity the CES-D has been shown to discriminate very well between psychiatric patient and community population samples, with, as expected, the psychiatric patient group obtaining significantly higher CES-D scores (Radloff, 1977; Weissman et al., 1977). Using a predictive approach to criterion validity quite a large number of studies have examined whether the CES-D could be used as a screening instrument for clinical depression. These studies are reviewed in a later section of this chapter.

Construct validity. Approaches to construct validity seek to test theoretically or empirically derived hypotheses about how a measure should behave. Convergent and discriminant validity are two aspects of construct validity and as covered earlier for the CES-D have provided mixed evidence. On a more encouraging note there is a very large literature showing that CES-D scores are associated with variables already known to be associated with the construct of depressive symptomatology. This literature is much too large to review here but for example CES-D scores have been shown to be related to eating disorders (Killen et al., 1994), suicidal behaviours (Garrison et al., 1991a, 1991c; Reifman & Windle, 1995), poor family functioning (Sheeber, Hops, Alpert, Davis & Andrews., 1997), stressful events (Costello, 1982), and drug use (Dick, Manson & Beals, 1993; Swanson, Linskey, Quintero-Salinas, Pumariega & Holzer, 1992).

Variables such as eating disorders, suicidal behaviour and family functioning have also been found to be related to scores from other self-report depression measures such as the CDI and the BDI. For example CDI scores have been found to be correlated with eating disorders (Strauss, Smith, Frame & Forehand, 1985), poorer family functioning (Nishide & Natsuno, 1997) and suicidal behaviours (Feldman & Wilson, 1997; Marciano & Kazdin, 1994). Therefore the pattern of results supports the construct validity of the CES-D in as much as CES-D scores appear related to a variety of variables in the same manner as other measures of depressive symptomatology (Devins & Orme, 1985).

CES-D exploratory factor analyses

Factor analysis has been used extensively to examine the hypothetical constructs underlying CES-D scores. In Radloff's (1977) original reliability and validity study, factor analysis revealed four factors which were interpreted as follows:

Depressed Affect: *Blues* (3)*, Depress* (6), *Lonely* (14), *Cry* (17) and *Sad* (18);

Somatic: *Bothered* (1), *Appetite* (2), *Effort* (7), *Sleep* (11) and *Getgoing* (20);

Positive Affect: *Good* (4), *Hopeful* (8), *Happy* (12) and *Enjoy* (16);

Interpersonal: *Unfriendly* (15) and *Dislike* (19).

The originally reported CES-D factor structure includes only 16 items. This is because only items with loadings of greater than 0.40 were included. Radloff suggested that if a less stringent criteria were applied (loadings of 0.35) then the remaining four items would be allocated as follows: Item 9 (*Failure*) and Item 10 (*Fearful*) to the Depressed Affect factor, and Item 5 (*Mind*) and Item 13 (*Talk*) to the Somatic factor. In subsequent work (Radloff, 1991; Radloff & Terri, 1986) Item 5 (*Mind*) was added to the Somatic factor but the remaining three items were not included. Not surprisingly there is some confusion in the literature about whether the original CES-D four factor solution comprises 16, 17 or 20 items.

In the present study the CES-D four factor model is taken to comprise the full 20 items loading to factors as shown above. With respect to factor labels, the original Radloff labels, that are the most widely used, are adopted. The four positive mood items: *Good* (4), *Hopeful* (8), *Happy* (12) and *Enjoy* (16) that comprise the Positive Affect factor are reversed scored before adding them to the total CES-D score. As such these items no longer represent the presence of positive mood but rather they indicate the absence of positive affect (Devins et al., 1988; Papassotiropoulos, Heun & Maier, 1999). Arguably this factor would be better labelled 'Absence of Positive Affect' but to maintain consistency with the established literature the original label (Positive Affect) is used.

A considerable number of exploratory factor analyses of the CES-D have been carried out using a variety of different approaches. Sometimes not all 20 items have been used and a wide variety of statistical criteria and extraction methods have been employed. Predictably therefore the items comprising the CES-D factors have not been consistent across studies (Callahan & Wolinsky, 1994). While many commentators give the impression that the CES-D four factor solution has been well replicated, a closer examination of the studies reveals much less agreement (Helmes & Nielson, 1998). Indeed, in the literature support can be found for two (Cheung & Bagley, 1998), three (Beals, Manson, Keane & Dick, 1991; Dick et al., 1993; Manson et al., 1990; Ying, 1988), four (Devins et al., 1988; Zich, Attkisson & Greenfield, 1990), or five (Thorson & Powell, 1993) factor solutions to the CES-D.

Helmes and Nielson (1998, p. 741) contend that the discrepant exploratory factor analyses of CES-D are predictable because the instrument was never designed to assess specifically different facets (Depressed Affect, Positive Affect, Somatic, Interpersonal difficulties) of depression:

The case of the CES-D illustrates the types of instability and confusion over subscale structure that are likely to arise when a measure lacking such *a priori* structure is subjected to an exploratory factor analysis.

It should be noted that this confusion arose subsequent to Radloff (1977) who cautioned strongly against placing undue emphasis on the separate factors and recommended that for epidemiological research a simple total score (of all 20 items) should be used. Researchers in the main have followed this advice and so in this context the key question concerns the assumption that scale scores arise from a unidimensional process. If it does then scores can be taken to represent an index of a single construct (Gibbons, Clark & Kupfer, 1993).

The CES-D and unidimensionality

In the literature the internal consistency of the CES-D as measured by Cronbach's coefficient alpha has consistently been found to be high. This evidence is consistent with the view that CES-D items are all related to the same underlying construct of depressive symptomatology. It is also the case that in community samples (where the scale is recommended for use) CES-D scores are highly skewed indicating a preponderance of low scores and a large number of zero item responses. Because of this, measures of internal consistency are inflated since nearly all people who score a modal zero on a particular item also score zeros on most other items (Coyne, 1994). This means that even a set of unrelated items may obtain a high measure of internal consistency, or coefficient alpha, if a large majority of subjects score a zero for each item. Quite clearly more stringent tests of the CES-D are required to establish the unidimensionality of the CES-D.

Prior to the widespread availability of SEM software packages only one study had systematically examined whether the CES-D was measuring a single dimension. This study by Clark, Aneshensel, Frerichs and Morgan (1981) used traditional factor analysis to extract four factors which were found to be highly correlated. A further second-order factor was able to be extracted from the correlation matrix of the four factors. This showed that the CES-D appeared to be measuring a single dimension. The 20 items were then projected onto this second-order factor by multiplying the loading on the first-order factor by the factor loading on the second-order factor. Generally the magnitude of these factor loadings was acceptable for the majority of items around 0.50 but the range was quite large from 0.24 (*Good* (4) to 0.73 (*Depress* (6)).

During the 1990s SEM software became widely available and allowed researchers to test more adequately the unidimensionality of the CES-D through what are termed 'second-order CFAs'. A second-order CFA examines the extent to which information in the correlations between the first-order factor can be accounted for by one superordinate factor. If the results indicate the presence of a well defined general factor accounting for most of total common variance then this provides evidence that the scale is unidimensional and supports the practice of calculating a total scale score. A CES-D second-order SEM path diagram is presented later in this report.

In previous second-order CFAs, researchers have initially estimated a CES-D four factor model (with factor correlations unconstrained) of the CES-D and then examined the loss of fit to this model when the factor correlations are forced to load to a single second-order factor. If the loss of fit is small then this shows that most of the information in the first order factor correlations is accounted for by the second-order factor loadings.

In the first second-order CFA of the CES-D Hertzog et al. (1990) used two samples (total N = 725) of predominantly Caucasian North Americans. The second-order CFA indicated that the four first-order dimensions could be modelled using a single second-order factor with a loss of fit (compared with the first-order model with unconstrained factor correlations) being not significant at the 0.01 level. Standardised second-order factor loadings relating the first-order factors to the general Depression factor were in the order 0.80 – 0.99 except for the Interpersonal first-order factor which was around 0.60. According to Hertzog et al. (1990) the results from their analyses justified the use of the total CES-D score.

In a replication study using a large (N = 2705) Australian community sample of aged persons, McCallum, Mackinnon, Simons and Simons (1995) also confirmed that

essentially all the information in the correlations between the first-order factors was accounted for by the second-order factor. In their analysis McCallum et al. initially used virtually the same statistical techniques as Hertzog et al. (1990) but in order to test better whether the four CES-D factors arise from one superordinate factor, what is known as a Schmid and Leiman (1957) second-order parameterisation was also performed.

The results from McCallum et al. (1995) showed the presence of a well defined general factor which accounted for over three-quarters (76%) of the total common variance while the specific factors (Depressive Affect: 5%; Positive Affect: 6%; Somatic: 5%; Interpersonal: 7%) each accounted for very much less of the common variance. These variance estimates totalled to 100 per cent and no estimate of the variation attributable to error (or more precisely not accounted for by the factors) was provided. These authors argued that their results supported the unidimensionality of the CES-D and they recommended its use as a single depression measure in English speaking countries.

The analysis performed by McCallum et al. (1995) represents a considerable improvement over previous higher order factor analyses of the CES-D. However, only 16 items were included in the analyses. Items: *Good* (4), *Hopeful* (8), *Sleep* (11) and *Cry* (17) were dropped to maintain consistency with a Japanese comparison sample and only three CES-D response categories were used. In addition, arguably a better modelling approach was available. This better modelling approach is a variation of the Schmid and Leiman (1957) second-order factor model and is known as a nested factor model. The advantage of this model is that it allows more fully for the decomposition of the variance-covariance information at the latent level (Little, 1997; Loehlin, 1998). Specifically a nested factor model, pioneered by Gustafsson (see Gustafsson, 1992; Gustafsson & Balke, 1993) enables the per cent of variation attributable to error to be calculated. It is proposed to apply a nested model in the present study.

Sheehan, Fifield, Reisine and Tennen (1995) performed a second-order CFA in a sample of 813 adults with rheumatoid arthritis. Obviously by mistake (identified by Helmes & Nielson, 1998) Sheehan et al. (1995) loaded Item 9 (*Failure*) and Item 10 (*Fearful*) items to the Interpersonal factor instead of loading these items to the Depressed Affect factor. Never-the-less their results showed that there was little to distinguish between the four factor and second-order four factor models. A further error occurred when the researchers interpreted their results as providing convincing evidence that there were four dimensions of depressive symptoms underlying the CES-D scale and they recommended that future researchers who wish to use a single score should base it on the second-order model rather than the single factor model as is now the current practice. This interpretation is unusual given that the presence of a general second-order factor supports the practice of calculating a total scale score. Whether factor scores should be used instead of observed scores is a different question.

Further studies confirming the existence of a CES-D second-order factor are reported by Davidson, Feldman and Crawford (1994) in a small (N = 303) sample of frail and disabled elderly people and by Knight et al. (1997) in a sample of New Zealand women in middle life. Two studies have reported results appearing not to support the existence of a second-order CES-D factor. Chapleski et al. (1997) found in a small (N = 277) sample of American Indian elders, that a second-order model of the CES-D (full 20 items) led to a significant deterioration of fit compared with the four factor model.

Wong (2000) in a sample of homeless adults mistakenly believed that because a second-order factor model did not yield a significantly **better** fit than the four factor model this meant that her results did not support the presence of a single second-order depression factor. This is incorrect because if a second-order CES-D model fits no **worse** than a four factor model then the second-order model is to be preferred on the grounds of parsimony (Rindskopf & Rose, 1988).

On balance the available evidence from the SEM studies supports the notion of CES-D unidimensionality when used in adult non-minority samples. Importantly, there has not been a second-order CFA of the CES-D carried out with an adolescent sample. In addition, the SEM analyses reported in this section were performed using ML estimation techniques which among other things assume that all variables arise from a multivariate normal distribution (MVN). This is unlikely to be the case because CES-D items are ordinal or polychotomous in nature and significantly skewed. This methodological shortcoming is discussed later in more detail. For now it is sufficient to note that a subsidiary aim of the present study is to examine whether the CES-D exhibits unidimensionality in a adolescent sample using SEM statistical techniques specifically developed for ordinal data.

Screening for clinical depression with the CES-D

The most reliable and valid method to identify individuals who are depressed is through a semi-structured diagnostic interview by a trained clinician. The main disadvantage of this approach is that it is costly and for large scale community programs impractical. It is impractical not only because of the skilled personnel required but also because the prevalence of diagnosable depression is so low: 1 to 3 per cent (Coyne, 1994). Self-report depression inventories on the other hand are fast and relatively inexpensive to implement in large groups and offer the potential to be an economical method for identifying cases in the general population (Dierker et al., 2001). Screening is particularly appealing in community samples where the disorder is likely to be milder, associated with less impairment and more likely to go undetected and untreated when compared with clinical populations.

Given the potential benefits of screening and the popularity of the CES-D, quite a number of studies have investigated effectiveness of the CES-D as a screen for clinical depression. As a general rule it has not been expected of the CES-D that it establishes a definitive diagnosis. Rather the CES-D has been used to partition a population into two groups: the first consisting of individuals presumed not to be cases (negative on screening) and the second consisting of individuals scoring above a particular cut-point who are presumed to be cases (positive on screening) (Roberts & Vernon, 1983). Studies examining the use of the CES-D as a screen for clinical depression have produced similar results across adolescent, adult and older age groups (see Lewinsohn, Seeley, Roberts & Allen, 1997; Radloff & Teri, 1986). For the purposes of this review material relating to the screening effectiveness of the CES-D in adolescent samples is the most pertinent and is reviewed.

Several measures to evaluate the effectiveness of a screening instrument have been developed and these are briefly described in the following paragraphs. The sensitivity (or true positive rate) of an instrument refers to the proportion of individuals with a particular disorder who also score above the cut-point and hence are correctly identified as ‘cases’ by the instrument. Specificity (true negative rate) refers to the proportion of individuals without the disorder who score below the cut-point. Receiver operating characteristic (ROC) curves are used to examine the

sensitivity and specificity of an instrument at various cut-points with a view to selecting the most optimal cut-point. The positive predictive value (PPV) of an instrument represents the probability that an individual has the disorder when the instrument indicates that the disorder is present.

Garrison et al. (1991b) conducted a large scale (around 2500 students) screening of middle school students (majority between 12 to 14 years of age). A sample of these students (N = 332) was also administered the Present Episode Version of the Schedule for Affective Disorders and Schizophrenia in School Age Children (K-SADS: Chambers et al., 1985). CES-D scores were significantly higher among students diagnosed with a major depressive disorder as compared with those judged not to be experiencing a disorder. In addition, a significant trend of increasing scores was observed from dysthymia without major depression, to major depression without dysthymia, to double depression (both major depression and dysthymia).

Garrison et al. (1991b) calculated that the optimal cut-point for males was a score of 12 and above and for females a score of 22 and above. The CES-D performed considerably better with females than males but even in females at the optimal cut-point for major depression the sensitivity of the CES-D was 83 per cent, the specificity was 77 per cent and the PPV 25 per cent. The CES-D was also found to perform comparably as a screen for probable cases of any psychiatric disorder as it did for major depression. It was concluded that the low PPVs of the CES-D were of considerable concern because if the CES-D was used in a two stage screening procedure around 75 per cent of identified cases would not have the disorder. In Garrison's et al. (1991b) opinion the high costs of providing further more intensive intervention to those not requiring it and the potential iatrogenic effects caused by erroneously identifying individuals as clinically depressed meant that the CES-D was not a useful depression screening measure.

In a sample of high school students (average age 16.6 years), Roberts et al. (1991) compared the screening efficiency of both the CES-D and the Beck Depression Inventory using a modified version of the K-SADS. Optimal cut-points for the CES-D were identified as 24 for females and 22 for males and using these cut-points for major depression (current) the sensitivity of the CES-D was 84 per cent, specificity 75 per cent and the PPV 8 per cent. Similar results were obtained for the BDI but this scale identified slightly fewer false positives and produced closer point prevalence estimates (e.g. prevalence from diagnostic interview, 3%, BDI estimate, 4%, and CES-D estimate, 12%). The researchers concluded that neither the BDI or CES-D scale represented an efficient or cost effective method for identifying cases of depression. Curiously the optimal cut-points for the BDI were 11 for females and 15 for males but for the CES-D this gender ratio was reversed (i.e. 24 for females and 22 for males). Roberts and colleagues were unable to offer any explanation for this anomaly.

The studies by Roberts et al. (1991) and Garrison et al. (1991b) both showed that a high proportion of adolescents with elevated CES-D scores do not meet the criteria for clinical depression. This basic finding is consistent with the results from studies in other populations and has prompted a detailed assessment of why self-report depression scales such as the CES-D are not more successful when used as screening instruments for clinical depression. One obvious explanation concerns the lack of content validity for the CES-D identified by Gotlib and Cane (1989) and referred to earlier. Coyne (1994), Fechner-Bates et al. (1994) and Zimmerman and Coryell (1994) argue that the mismatch between CES-D items and clinical depression symptoms serves to reduce the screening efficacy of the CES-D in a number of ways.

First, because a number of the criteria or symptoms for clinical depression are not assessed by the CES-D (e.g. suicidal ideation) this has the effect of reducing the CES-D's sensitivity (the ability of a scale to detect individuals with depression). Second, because some CES-D items are not related to the diagnostic criteria for depression (e.g. *Bothered* (1)) and some items are redundant in the sense that they tap the same symptom (e.g. *Depress* (6) & *Sad* (*18*)) this reduces the CES-D's specificity (the ability of a scale to detect individuals without depression). Third, it can be noted that the scoring procedure for the CES-D counts symptoms of short duration and mild intensity. This means that it is possible for respondents to obtain high total CES-D scores without having a single symptom that would count toward a diagnosis of a major depressive disorder.

Several strategies have been tried to improve the screening effectiveness of self-report depression scales. These strategies have included raising cut-points (Andresen et al., 1994; Boey, 1999), reducing the number of items to match more closely DSM criteria (Santor & Coyne, 1997), administering the tests on two different occasions (Deardorff & Funabiki, 1985), or designing a new scale specifically tied to DSM inclusion criteria (e.g. the *Inventory to Diagnose Depression*: Zimmerman & Coryell, 1987). Unfortunately these attempts have proved largely unsuccessful and according to Kendall and Flannery-Schroeder (1995) paper-and-pencil self-report tests simply do not provide the same information seeking capacities as do structured professional interviews which can factor the interpersonal process into a diagnosis. For this reason Kendall and Flannery-Schroeder (1995, p. 893) conclude that "… there exists no reason to believe that a new self-report measure would be in any way devoid of the problems of the presently used self-report measures".

In this section a number of studies have been reviewed that have examined the effectiveness of the CES-D for identifying cases of clinical depression. Undoubtably these studies have provided useful knowledge about whether self-report depression scales can be used as epidemiological screening instruments. But it is also true that the CES-D was never intended by its developer to be used as a clinical screening tool. There is a subtle yet important point here and it is that defining a case of clinical depression and measuring the severity of a depressive syndrome are different although related objectives (Gibbons et al., 1993; Santor et al., 1995).

Arguably the CES-D should be judged less on the basis of its screening characteristics and more on its ability to differentiate (or discriminate) between individuals with different levels of depressive symptomatology. In contrast to the quite substantial literature which has examined the ability of the CES-D to act as a screen for clinical depression very few studies have investigated the issue of CES-D discriminability. The research which has examined the ability of the CES-D to discriminate has been performed using IRT models and these studies are reviewed in a later section of this chapter.

CES-D gender differences at the item or sub-scale level

Earlier it was established that the CES-D literature clearly shows that in adolescent samples, girls, on average, report higher CES-D total scores than boys. Theoretically these gender mean CES-D total score differences could be reflected in differences at the sub-scale, item or response option level. Girls on average might have slightly higher mean scores on all 20 items or alternatively might have markedly higher

scores for just one or two items. If only a few items are involved these may all relate to the same factor or alternatively the items may be drawn from a number of different factors. At the response option level some options may be differentially endorsed by boys and girls. For example even at high levels of depressive symptomatology boys might be less likely than girls to endorse response option four (most or all of the time) for any particular symptom.

Several researchers have examined CES-D gender differences at the sub-scale, item and response option level. In a very thorough early study of the effects of gender on CES-D items Clark et al. (1981) using a sample of 1000 adults from Los Angeles compared the means, response option patterns and item-scale correlations of males and females. Females obtained higher mean scores on all items except Item 7 (*Effort*) and were more likely to report the presence (score 1, 2 or 3) of all symptoms except for Item 7 (*Effort*) and Item 1 (*Bothered*). The largest female to male ratio for the presence of symptoms was found for the Depressed Affect items and in particular *Cry* (17), *Sad* (18) and *Failure* (9). Nearly all (18 out of 20) item-scale correlations were higher for females than males.

Clark et al. (1981) speculated that it was possible that CES-D items operated differently for men and women and that at a given level of depression females appeared to be reacting differently to some items. It was concluded that mean gender differences in total scores might constitute real differences in levels of symptomatology or equally reflect measurement bias. The finding of Clark et al. (1981) relating to the high female Depressed Affect item scores is consistent with analyses of other self-report depression scales which also show that high female total scores are largely a function of excess scores with respect to Depressed Affect (Craig & Van Natta, 1979; Newmann, 1984).

It will be recalled that the CES-D comprises items which are designed to measure components of depressive symptomatology. Although there is a clear emphasis on depressed affect, items covering feelings of guilt and worthlessness, feelings of helplessness and hopelessness, and somatic items such as psychomotor retardation, loss of appetite, positive affect and finally sleep disturbance are also included. The evidence from Clark et al. (1981), Craig and Van Natta (1979) and Newmann (1984) raises the possibility that the high levels of observed female depressive symptomatology might be due simply to the presence of transient symptoms of depressed affect rather than to a depressive syndrome.

Theoretically the selection of the types of symptoms measured might be one methodological factor which influences the magnitude of the gender difference observed in scores from self-report depression scales. It might be hypothesised, for example, that if the CES-D comprised more items measuring depressed affect then the gender difference in total scores could be larger. Surprisingly there is very little literature which has examined the balance of the different types of symptoms that should be included in self-report depression scales.

It can be noted that CES-D scale comprises covering Depressed Affect (seven items), Positive Affect (four items), Somatic (seven items) but only two items covering interpersonal difficulties. The importance of the types of items to be included in a depression scale was realised by Clark et al. (1981, p. 179) who argued that if the correct balance was not achieved this could potentially bias any subgroup comparisons:

If depression is a complex state which has many different dimensions for different people, then a scale that includes a wide range of types of items with different response rates for different types of subjects is necessary….If one wants to use a

scale to compare subgroups of a sample then the need for balance is critical as possible biases will affect the results obtained.

In one of the few studies to examine this issue Compas et al. (1997) compared the extent of gender differences in self-reports of depressed mood, mixed symptoms of anxiety-depression and an analogue of clinical depression in a large sample of referred and nonreferred adolescents. These researchers found that among non-referred adolescents gender differences overall were quite small in magnitude but that adolescent girls did report more depressed mood and higher scores on the anxiety – depression syndrome. Compas et al. (1997) noted that their findings were consistent with the research of Silverstein, Caceres, Perdue and Cimarolli (1995) who also found that the largest gender difference occurred in symptoms of affective distress (depressed and anxious mood) with smaller differences in the other symptoms of depression (e.g. sleep difficulties) and no difference with respect to a more pure index of symptoms of major depression.

Using a slightly different approach to examining possible CES-D gender bias Roberts et al. (1990a) in a large community sample of young adolescents assessed the importance of individual items to boys and girls by ranking the means of each item separately for boys and girls. Roberts et al. found that both ranks were similar (rank correlation = 0.82, $p < 0.001$) but not identical (data not shown). The greatest disparity in ranks occurred for Item 2 (*Appetite*) (Girls: 12; Boys: 18) and Item 15 (*Unfriendly*) (Boys: 12; Girls: 19) For all other scale items the difference in ranks was five or less.

At the factor or sub-scale level, Manson et al. (1990) in a small sample of American Indian adolescents found that girls endorsed symptoms pertaining to the Depressed Affect, Somatic and Interpersonal factors far more frequently than boys. No differences were observed between boys and girls with regard to items comprising the Positive Affect factor. Consistent with these findings in a Japanese sample of adolescents Iwata, Saito and Roberts (1994) found that symptom presence (a score of 1, 2 or 3) on negative items (items comprising the Depressed Affect, Somatic and Interpersonal factors) was more common among girls than boys but that for the positive items (items comprising the Positive Affect factor) symptom frequencies were comparable across gender.

CES-D gender impact or bias?

The CES-D gender differences in total scores, sub-scale scores or at the item level reported in the studies reviewed in the previous sections provide useful but limited preliminary information about possible gender bias in the CES-D. The information is limited because analyses which rely solely on manifest (or observed) variables are not diagnostic of bias or the lack of bias (Meredith & Millsap, 1992). As Mackinnon et al. (1995) explain the use of item endorsement frequencies to investigate possible item bias confounds the two factors that affect item endorsement: differences between groups on the trait that the items measure and differences in the manner in which the item functions in different groups.

While it seems to make intuitive sense that the term ‘impact’ can be taken to be the same as bias this is not true. In the educational testing and statistical literature, Simpson’s paradox (Simpson, 1951) is often used to illustrate why appropriate judgements about differences in item function can only be made after groups have been matched (or conditioned) on the construct that the item purports to measure.

Table 2 presents a hypothetical example of Simpson's paradox adapted from Dorans and Holland (1993). In this example imaginary means for Item 17 (*Cry*) are presented for males and females.

Table 2 Hypothetical Simpson's paradox for the CES-D Cry item

	Males		Females	
	N	Mean	N	Mean
CES-D range				
Low	1000	0.2	400	0.1
Moderate	1000	0.6	1000	0.5
Severe	400	1.0	1000	0.9
Total	2400	0.5	2400	0.6

From Table 2 it can be seen that the sample is taken to comprise 2400 males and 2400 females and the overall female mean score for Item 17 (*Cry*) item is set higher than the overall male mean score (0.6 versus 0.5). When this mean CES-D item score is decomposed into three groups (low, moderate and severe levels of depressive symptomatology), however, a quite different picture emerges. At each of the three different group levels of depressive symptomatology the mean item score is actually 0.1 less for females than it is for males (e.g. 0.1, 0.5 and 0.9 for females compared with 0.2, 0.6 and 1.0 for males).

Paradoxically when item means are compared at each level of depressive symptomatology this item appears to produce higher scores for males and not vice versa as indicated by the overall mean item score. The perplexing result arises because of the unequal distributions of depressive symptomatology in the two groups. Although the overall item means in this hypothetical example suggest that Item 17 (*Cry*) is producing higher scores for females in actual fact it is more likely than not the reverse is true.

If bias cannot be proved simply by the presence of an item score difference what then is required? Bias, or DIF as it is now generally referred to, is evident when differences in item functioning are observed after groups have been matched with respect to the ability or trait that the item measures. In other words for DIF to be properly tested individuals must be equated along a continuum so that the responses to items can be examined for individuals who are equally depressed.

In this manner, as Reisse et al. (1993) proposed, the empirical relations between the test items and the trait of interest can then be compared across groups. Currently there are two main statistical approaches for examining the relation between test items and their underlying traits, namely IRT and CFA. In the sections that follow the literature which has investigated measurement invariance in the CES-D using these techniques is reviewed.

Item response theory and psychological scales

IRT models are routinely used, and have been for many years, in large scale educational testing programs to examine whether test items function differently in different groups. A detailed account IRT models for detecting DIF in educational tests is provided by Holland and Wainer (1993). During the 1990s the widespread availability of easy to use desktop computer software enabled these methods to be used more widely by researchers in psychology and psychiatry. The conceptual and practical advantages of IRT techniques for the evaluation of psychiatric and psychological rating scales have been convincingly demonstrated in a number of key studies.

In the psychiatric literature the advantages of IRT models are demonstrated in studies of the *Beck Depression Inventory* (BDI) (Bedi, Maraun & Chrisjohn, 2001; Gibbons, Clark, VonAmmon-Cavanaugh & Davis, 1985; Clark, Gibbons, Haviland & Hendryx, 1993), the *Hamilton Depression Rating Scale* (Gibbons et al., 1993) and of the *Toronto Alexithymic Scale* (Hendryx, Haviland, Gibbons & Clark, 1992). Quality of life instruments such as the *General Health Questionnaire* and the *SF-36 Health Survey* have been evaluated using IRT models (see Andrich & Van Schoubroeck, 1989; Raczek et al., 1998) as well as the *Mini-Mental State Examination* (Marshall, Mungas, Weldon, Reed & Haan, 1997). In Australia the NHMRC Psychiatric Epidemiology Research Centre has used IRT models to examine the possibility of age related DIF in the *Eysenck Personality Questionnaire* (Duncan-Jones, Grayson & Moran, 1986; Mackinnon et al., 1995).

It is generally acknowledged that the psychological community has been slow to apply IRT models to measurement tasks and that the real contribution of this method to the discipline is yet to be realised (Embretson, 1996; Zickar, 1998). Although IRT techniques have not yet been applied widely in psychological studies they have been used on a limited basis to investigate substantive research questions with the BDI and CES-D and the MMPI (Waller, Thompson & Wenk, 2000). Zickar and colleagues have used IRT models to detect faking on personality instruments in the context of recruitment to military services (Robie, Zickar & Schmit, 2001; Zickar & Drasgow, 1996; Zickar & Robie, 1999). The present author and colleagues (Allison et al., 2001) used non-parametric IRT techniques to analyse gender differences in the relationship between depressive symptomatology and suicidal ideation in young adolescents.

Santor and colleagues have published IRT analyses of the BDI and CES-D in a series of papers (viz. Santor & Coyne, 1997; Santor, Ramsay & Zuroff, 1994; Santor et al., 1995). Several lines of inquiry have been followed in these studies but the most salient of these concern the possibility of gender bias in the BDI and scale discriminability in the BDI and CES-D. The main findings from these studies are briefly summarised below. The analytical approach (based on nonparametric kernel-smoothing techniques) and associated software (TestGraf) which was developed by one of the members of this research team is used in the present study and is described later.

Santor et al. (1994) investigated possible gender BDI item bias using a sample of American college students (average 20 years of age) and a sample of adult outpatients with clinical depression. In the outpatient sample three items (Item 6: sense of punishment, Item 10: *Crying* and Item 14: *Distortion of body image*) showed a small degree of DIF. For Items 10 and 14, at all levels of depression, women tended to respond more strongly to options reflecting more severe depression. For Item 6 this pattern was reversed, with men, for the same level of depression, indicating that they

felt more punished than did women. In the college sample again Item 14 exhibited the largest amount of DIF but the other two items showed very little DIF. When items showing DIF were removed, women in their sample remained significantly more depressed than men indicating that the BDI was relatively unbiased when used in their samples.

The following year Santor et al. (1995) published a paper investigating the ability of the CES-D and BDI to discriminate among individuals at different levels of depressive severity. Scale discriminability was defined as how effectively the BDI and CES-D discriminated among individuals along the continuum of depressive severity. The scales would be shown to be effective if they discriminated equally across the full range of depressive severity. That is, differences between two individuals in the low range of depressive severity should be as easily detected as the difference between two individuals in the high range of scores. The results from the analyses showed that in the sample of college students the CES-D was more discriminating than the BDI and in the sample of outpatients the CES-D was not less effective than the BDI in detecting individual differences in depressive severity.

In a later review paper of the application of IRT models to the measurement of depression, Santor and Ramsay (1998, p. 357) presented a figure showing test information functions (TIF) for the CES-D and BDI produced using their college student sample referred to earlier. TIFs derive from the idea that scales differ with respect to how much information they can provide about the individuals they are used to assess. TIFs take into account both how well tests discriminate between individuals and how precisely they measure the amount of the trait (in this case depressive symptomatology) an individual has. The better a test is able to discriminate between individuals and estimate those differences precisely the more information is provided by that test. Interestingly the CES-D provided more information than the BDI at moderate to severe levels of depressive symptomatology but both scales provided less information at low levels of depressive symptomatology.

In what appears is the only other published IRT analysis of the CES-D, Gelin and Zumo (2003) recently investigated the possibility of CES-D gender DIF using a mixture of ordinal logistic regression (to identify items showing DIF) and IRT techniques (for post-hoc graphical displays) in a sample Canadian adults. Their results indicated that when the CES-D is scored in the traditional manner (using the ordinal format) Item 17 (*Cry*) was found to show higher scores for women. Using a presence scoring method (the respondent reports the presence of a symptom at least some of the time) Item 17 (*Cry*) was again identified as showing gender DIF. Finally using a persistence scoring method (the respondent reports the presence of a symptom a moderate amount or most of the time) Item 7 (*Effort*) and Item 8 (*Hopeful*) showed gender DIF with these items increasing scores for males.

The existing IRT analyses of the CES-D have been carried out in samples of adults and IRT models have not yet been used to examine the psychometric properties (in particular scale discriminability or information function) of the CES-D in a sample of adolescents. IRT models are applied in the present study as one of the two key techniques for investigating whether CES-D scores are comparable across adolescent boys and girls. CFA is also used as a complementary approach and in the next few sections of this chapter the quite extensive CES-D factor analytic measurement invariance literature is reviewed.

Structural equation modelling and the CES-D

Multiple group confirmatory factor analysis (MG-CFA) is widely considered to represent one of the most powerful and versatile approaches to testing measurement invariance in psychological scales (Cole, 1987; Dolan, 2000; Steenkamp & Baumgartner, 1998). It is also a relatively new approach that is quite challenging to implement in practice. MG-CFA models for testing measurement invariance originated in the early 1970s (see Jöreskog, 1971) but it was only with availability of computer SEM programs such as LISREL (Jöreskog & Sörbom, 1996) and EQS (Bentler & Wu, 1995) and the publication of several excellent introductory texts (see Bollen, 1989; Byrne, 1998) that this method has become more widely accessible to measurement researchers.

The basic idea behind a MG-CFA is the estimation of the same measurement model in two or more groups and then the testing of the equality of estimates of particular parameters in the different groups. Some types of parameters may show equality while other types of parameters may not. For this reason in a SEM framework measurement invariance is said to exist at several different levels. Unfortunately in the literature there has been a lack of agreed upon terminology to refer to these different levels (Steenkamp & Baumgartner, 1998). Together with the inherent methodological complexities involved in performing a MG-CFA analysis and the lack of guidelines for appropriate research practice (Floyd & Widaman, 1995) there is considerable ambiguity in the literature about the extent to which measures must be equivalent for comparisons to be meaningful.

A description of MG-CFA and an explanation of the different levels of measurement invariance is outlined in Chapter 8. In subsequent sections of this chapter the CES-D substantive literature which has used factor analysis methods to investigate measurement invariance across age, illness and cultural groups is presented. This material has been reviewed to examine the conceptual framework and statistical techniques used to investigate CES-D measurement invariance. Following this, studies that have investigated possible gender bias in CES-D are reviewed in some detail. A critique of these studies is provided and an argument developed that further research is required to establish the validity of the CES-D for meaningful gender comparisons.

Although an explanation of the use of factor analysis and MG-CFA to investigate measurement invariance is presented later it is necessary prior to reviewing the CES-D substantive measurement invariance literature to clarify briefly and distinguish between two levels of measurement invariance. The first level concerns whether item responses (manifest variables) load on the same constructs (configural invariance) while the second level addresses whether the factor loadings to the constructs are equivalent (metric invariance). Both these levels of measurement invariance are necessary (but not sufficient) conditions for meaningful comparisons to be made across groups. To date CES-D researchers have primarily focused on these two levels of measurement invariance.

At the first level researchers have tested whether the factor structure for the CES-D is the same across groups. This level of measurement invariance is known as 'configural invariance'. For example, it has been hypothesised that people from non-Western cultures minimise the difference between depressive and bodily complaints. This hypothesis can be tested by examining whether one factor accounts for the correlations between items loading to the Depressed Affect and Somatic factors. If it does then a three factor CES-D model with the Depressed Affect and Somatic factors collapsed would be preferred in these populations to the traditional CES-D four factor

model. This finding would also imply that people from non-Western and Western cultures construe depression differently and that CES-D scores should not be compared across these groups.

The second level of measurement invariance is known as 'metric invariance'. This level of invariance examines whether groups are responding to scale items in a similar fashion. Metric invariance is tested by constraining factor loadings to be the same across groups. Metric invariance provides a stronger test of invariance than configural invariance and has received the most consideration in the general psychometric literature (see Schaie et al., 1998 for a summary). The importance of this level of measurement invariance is that if differences between factor loadings across groups are found then this indicates an asymmetric relationship between true scores and observed scores. Given a lack of invariant factor loadings quantitative comparisons of observed scores are meaningless because a one-unit change in true scores does not translate into the same amount of change in observed scores for both groups (Schaie & Hertzog, 1985).

CES-D measurement invariance across age and illness groups

Epidemiological studies using structured clinical interviews generally show that the prevalence of depression decreases with age. On the other hand studies using self-report depression scales show increasing levels of depressive symptomatology in later life (see Newmann, 1989 for a review). This discrepancy has directed attention to whether self-report measures might artificially raise scores in the elderly. One plausible way this could happen is because the elderly are more likely to suffer various types of physical malaise than younger people. Hence they might also be more likely to give affirmative responses to the somatic items (loss of energy, sleep and appetite disturbances) in self-report depression scales (Blazer, 1982). If CES-D somatic items for the elderly reflect both underlying levels of depressive symptomatology and also the aging process itself then it is quite possible that total CES-D scores would be artificially inflated.

In one of the first MG-CFA studies of the CES-D, Liang, Tran, Krause and Markides (1989) tested the factorial invariance of the CES-D across three samples (reflecting three age generations: total N = 1125) of Mexican Americans. Liang et al. (1989) proposed a three factor model of the CES-D with the assignment and selection of items based on face validity and reliability. Their three factor model excluded the Interpersonal factor items: *Unfriendly* (15) and *Dislike* (19) because conceptually these items seemed to confound a lack of social support with depressive symptoms. A further number of items were also deleted: *Bothered* (1), *Effort* (7), *Talk* (13), *Fearful* (10), *Failure* (9) and *Good* (4), but note that even in this reduced model several Somatic factor items: *Appetite* (2), *Sleep* (11), *Getgoing* (20) and *Mind* (5) were included.

The Liang et al. (1989) 12 item, three factor CES-D model had an adequate fit to their data. The correlation between the latent factors of Depressed Affect and Somatic was high (> 0.90) in all three samples but Positive Affect correlated only moderately with the other two factors. Liang et al. (1989, p. 117) reported that: "When simultaneous factor analysis was applied, structural variations across three generations of Mexican Americans were found. Whereas the first-order factor loadings are invariant, measurement error variances are not". Because of the lack of invariance in measurement error variances Liang et al. concluded that the meaning of inter-

generational differences was ambiguous – due either to mean differences, structural differences or both.

Arguably, Liang et al. (1989) were mistaken in their view that they had shown that the CES-D should not be used for across age group comparisons because invariance of error variances is not considered necessary for the comparison of latent means (Horn & McArdle, 1992; Meredith, 1993; Steenkamp & Baumgartner, 1998). This ambiguity was to cause some confusion which became evident in the next study to examine this issue. Based on the full (20 item) CES-D four factor model Hertzog et al. (1990) found that factor loadings across age ranges were invariant but factor variances and covariances were not equivalent. Measurement error variances were not tested. The lack of invariance for factor variances and covariances was thought possibly to be due to the skewed distribution of the CES-D scores and it was recommended that future studies use special techniques for analysing non-normal ordinal data.

Hertzog et al. (1990) concluded from their results that the CES-D scale had age-invariant measurement properties and could therefore be legitimately used for quantitative comparisons across age groups. They noted (p. 64) that their results were "… inconsistent with recent findings of Liang et al. (1989), who found generational differences in CES-D item factor loadings in a Mexican-American population". In reality this was not the case; both research groups had found equivalent factor loadings across age groups and the difference was simply that Liang et al. had tested for the equivalence of error variances and found that these were not invariant. Thus two studies have provided evidence of configural and metric invariance (similar factor pattern and factor loadings) for the CES-D across age generation groups.

Earlier in this review it was noted that three major longitudinal studies (viz. Ge et al., 1994; Petersen, Sarigiani & Kennedy, 1991; WichstrØm, 1999) found that gender differences in overall levels of depressive symptomatology began to emerge around the ages of 13 to 15 years primarily as a result of adolescent girls reporting higher levels of depressive symptomatology than adolescent boys. The findings from these major longitudinal studies raise the possibility that the meaning of depressive symptomatology might be changing for young people during early adolescence. For example, it is possible that overall true levels of depressive symptomatology for girls remain equivalent between the ages of 13 and 15 years but some symptoms or items are just more likely to be endorsed.

This possibility was examined by Roberts et al. (1990a) using SEM techniques in a sample of adolescents between Grade levels 9 and 12 and aged mainly from 15 to 17 years. Their results indicated that CES-D item factor loadings were invariant across age year groups. It was concluded from the SEM analyses, and supported by some additional descriptive statistics that the CES-D was extremely stable across these late adolescent age groups. One of the subsidiary aims of the present study is to replicate this analysis in a sample of younger adolescents, where a lack of measurement invariance might be more evident and to test more fully different levels of measurement invariance.

The CES-D has also been applied in medical populations to examine the impact of illness, for example, low blood pressure (see Paterniti, Verdier-Taillefer, Geneste, Bisserbe & Alpeovitch, 2000), or cancer (see Beeber, Shea & McCorkle, 1998; Hann, Winter & Jacobsen, 1999) on levels of depressive symptomatology. Conversely, the CES-D has been used to assess whether depressive symptomatology itself is a risk factor for illnesses such as cancer (Zonderman, 1995). An obvious potential problem

for examining the relationship between depressive symptomatology and medical illness is symptom overlap – that is, overall levels of depressive symptomatology may be overestimated due to the mis-identification of somatic illness symptoms for symptoms of depression.

In substance, this concern is the same as raised for the use of the CES-D across age groups, with total CES-D scores reflecting not only levels of depressive symptomatology but also complaints associated with the aging process. Only one study appears to have specifically investigated whether the CES-D measures depressive symptomatology equivalently across healthy and ill groups. Devins et al. (1988) examined the reliability, factor structure and potential for item bias of the CES-D from data collected from five adult samples with varying levels of health and illness (from healthy undergraduates to end-stage renal disease patients).

Devins et al. (1988) found that the CES-D exhibited good internal consistency across the five groups and the factor composition was very similar. A five group, 20 item repeated measures analysis of variance indicated a main effect for both groups and items. Using the criterion of a difference of greater than one scale point as evidence of a substantively meaningful difference, 13 differences across the CES-D items involving the Positive Affect factor were identified. Despite this finding it was concluded that overall the CES-D was not compromised seriously by item bias when used in non-psychiatric medical patient populations.

CES-D measurement invariance across cultural groups

The CES-D was developed for use with Anglo-Americans and the possibility that the CES-D may not be a valid measuring instrument in other ethnic groups has generated two distinct lines of inquiry. The first line of inquiry concerns the notion that people from some non-Western cultures tend to somatise depressive symptoms (the Somatisation Hypothesis: Kleinman & Good, 1985) and that as a result CES-D affective and somatic depressive symptoms may not be clearly differentiated. The second issue concerns whether people from Asian cultures tend to suppress positive affect expression. This means that people from Asian cultures may give less positive responses to the CES-D positive affect items (such as 'I was happy') than people from Western cultures even when they share similar levels of depressive symptomatology.

The evidence that people from non-Western and Western cultures construe depression differently because non-Western people conflate depression's affective and somatic components is unclear. If it is the case, then factor analyses of the CES-D should show that one factor accounts for the correlations between items loading to the Depressed Affect and Somatic factors in non-Western cultures. Therefore a three factor CES-D model, with the Depressed Affect and Somatic factors joined, would be preferred in these populations. The issue is important because it concerns a possible lack of configural invariance, or factor pattern, for the CES-D and it has been investigated using Hispanic, American Indian and Asian samples.

Guarnaccia, Angel and Worobey (1989) examined the factor structure of the CES-D in a large adult sample of the three major Hispanic groups in the United States (Mexican–Americans, Puerto Ricans & Cubans). On the basis of a highly significant χ^2 statistic, Guarnaccia and colleagues judged that the CES-D four factor model provided a poor fit to their data. They then proceeded to perform six (three cultural

groups by gender) exploratory factor analyses (EFA's). EFA was used because "... since we have no other *a priori* basis on which to hypothesize factor structures for any of the Hispanic sub-groups we proceed in an exploratory fashion" (Guarnaccia et al., 1989, p. 87). They found that each of the six analyses yielded a similar three factor structure – with items from the Depressed Affect and Somatic factors loading to a single factor. On the basis of the EFA results it was concluded that the meaning of CES-D items differed between Hispanics and non-Hispanics.

At around the same time, factor analysis was used by Golding and Aneshensel (1989) to assess whether the structure of the CES-D was the same in three adult samples of non-Hispanic United States Whites, United States born Mexican Americans and Mexican born Mexican Americans. EFA in the three samples yielded a model very similar to the Radloff four factor solution. An initial CFA of this model showed a poor fit and to improve fit a number of error covariances were allowed to correlate. In addition, Item 11 (*Sleep*) was allowed to load on both the Depressed Affect and Somatic factors. MG-CFA showed that factor loadings were not equal across the three groups but the correlations of the factor loadings between each pair of groups were high (above 0.80) indicating, in the researchers' view, that while the factor loadings were not identical they were, with the exception of Item 11 (*Sleep*), substantively similar.

Golding and Aneshensel (1989, p. 167) surmised that their results had provided "… high but imperfect conceptual equivalence of the CES-D among Mexican-Americans and non-Hispanic Whites". Conceptual equivalence was defined as the extent to which responses measure the same basic construct across groups. The two studies (Guarnaccia et al., 1989; Golding & Aneshensel, 1989) investigated the same problem, at roughly the same time, with similar samples and statistical techniques, yet they arrived at opposite conclusions.

One clue to the discrepant results is the very different analytical strategies employed. Both research teams used CFA to test the fit of the original CES-D four factor model and found that this model did not fit their data well. Guarnaccia et al. (1989) proceeded to a EFA whereas Golding and Aneshensel (1989) improved the four factor model fit by allowing a number of measurement errors to correlate and allowing Item 11 (*Sleep*) to load to both the Depressed Affect and Somatic factors. Several points can be made about these two studies of whether the CES-D exhibits cross cultural (Hispanic) configural invariance.

The first point to be made is that it is hardly credible that Guarnaccia et al. (1989) lacked any other theoretical model other than the four factor model to test using CFA. At least one other model should have been obvious, namely, to use a three factor model combining items from the Depressed Affect and Somatic factors and to test this against the four factor model in a CFA framework. In addition, rejecting the four factor model solely on the basis of a significant χ^2 statistic of fit seems unwise because given their large sample size (around 5000), it was nearly inevitable that this statistic would indicate a poor fitting model. Other alternative indices of fit should have been considered (and also some judicious modifications) before resorting to EFA techniques.

Golding and Aneshensel (1989) proceeded in a more logical fashion but provide few details about the number (or specific items) of the error covariances which were allowed to vary. In addition, they do not appear to have recognised the importance of the selection of the marker item (i.e. the item which serves to define the scale of latent variable) across groups for each factor. In their analysis the first item for each factor across groups was arbitrarily fixed at unity and served as the marker item. These

items are unlikely to be the most metrically invariant across groups and this approach may have caused other items to appear (incorrectly) not to be invariant (see Cheung & Rensvold, 1999 for a discussion of this issue). More positively the researchers tackled the difficult issue of assessing the substantive importance of the finding that factor loadings varied across groups.

In CES-D CFA studies using samples of American Indians, a similar picture of discrepant findings is evident. Beals et al. (1991) compared the fits of a one factor, three factor (Depressed Affect and Somatic factors combined) and four factor CES-D models in a sample of 605 American Indian college students. They found that the one factor model generally did not fit the data well and that although the fits of the three and four factor models were very similar, the Depressed Affect and Somatic factors were so highly correlated as to be practically indistinguishable. On this basis the authors argued that the three factor solution was the most appropriate for their data. This finding was later replicated by the same research team in a small (N = 188) sample of American Indian adolescents (Dick, Beals, Keane & Manson, 1994) although in this study Item 7 (*Effort*) was found to load consistently on the Positive Affect factor (rather than the Somatic factor).

Chapleski et al. (1997) in a small sample of American Indian elders compared a one factor model, a two factor model (Depressed Affect, Somatic and Interpersonal forming one factor and Positive Affect the other), a three factor model (Depressed Affect and Somatic forming one factor) and the four factor CES-D model. In addition, the 12 item three factor model proposed by Liang et al. (1989) was tested. With respect to models comprising the full 20 CES-D items there was little to distinguish (in terms of fit) between the three and four factor models. However, because the four factor model was consistent with the *a priori* theoretical conception of the CES-D this model was preferred. The 12 item three factor model was found to provide the best fit of all the models tested although no discussion of whether this was due to the reduced number of items in this model was provided.

In summary, the evidence from the studies that have examined the Somatisation hypothesis in samples of American Indians is mixed and it is difficult to draw any firm conclusions from the findings. During the 1980s, EFAs of the CES-D in Asian samples also produced conflicting evidence about whether these groups differentiated between the Depressed Affect and Somatic factors (Ying, 1988; but cf. Kuo, 1984). This issue was examined by Noh, Avison and Kaspar (1992) who carried out a quite detailed assessment of the utility of a translated version of the CES-D in a sample of Koreans living in Canada. Both a so-called 'replicatory factor analysis' (no further details provided) and a EFA yielded (number of eigenvalues greater than one) four factors closely resembling the original four factor solution. The researchers concluded that their findings did not support the argument that Asians tended to somatise depression.

Following Noh et al. (1992), Cheung and Bagley (1998), in a small sample of Hong Kong Chinese married couples, used CFA to test a two factor (Positive Affect, Depressed Affect and Somatic factors combined), three factor (Depressed Affect and Somatic factor combined) and the original four factor model of the CES-D. The proposed two factor model of Cheung and Bagley (1998) was based both on the idea that Asians may somatise depressive symptoms and also the idea (derived from an EFA in sample of Black North Americans by Callahan & Wolinsky, 1994) that the Positive Affect items are simply antonyms for the Depressed Affect items and therefore these items should be combined. On the basis of the parsimonious normed goodness-of-fit index (PNFI) the authors concluded that the two factor model provided the best fit to their data.

Closer inspection of the results from Cheung and Bagley (1998) reveals that while the PNFI measure of fit did indeed favour the two factor model all other indices of fit reported in the study (χ^2, GFI, CFI, NFI, RMSEA) favoured the three factor model. The principal difficulty with the three factor model was that the correlation between the combined Depressed Affect / Somatic factor and the Positive Affect factor was very high (around 0.90) and arguably two factors lacked discriminant validity. In any event the results did support the notion that Asians tended to blur the distinction between the affective and somatic symptoms of depression. Consistent with the results from several other studies (Liang et al., 1989; Dick et al., 1994) the factor loading for Item 7 (*Effort*) did not reach significance indicating poor convergent validity.

In the most recent (at the time of writing) cross cultural CFA, Greenberger et al. (2000) examined the equivalence of a modified (five point rating scale rather than the traditional four point) CES-D across Chinese and United States adolescents (age 16-17; total N = 703) as part of a study of the correlates of depressive symptoms among adolescents. Configural invariance across the two groups was tested using the four factor model and with the addition of an error term correlation between *Failure* (9) and *Hopeful* (8) this was found to provide a good fit to the data. Factor loadings on the other hand were not equivalent across the two groups, although the model still showed a reasonably good fit to the data. Covariances among factors did not differ between the two groups but when factor variances and error terms were constrained to be the same across both groups model fit deteriorated substantively. Greenberger et al. (2000, p. 212) concluded:

In summary tests of several models for which we examined goodness-of-fit statistics and chi-square values indicated that the U.S. and Chinese samples showed the same factor pattern and the same interrelations among factors but differed in specific factor loadings and the variances of individual items. Overall, the structure of the CES-D is sufficiently similar to warrant its use with these two culturally different samples.

While Greenberger and colleagues are to be commended for investigating the possibility of a lack of measurement invariance prior to carrying out analyses using analysis of variance and regression techniques with observed variables, few psychometricians would agree with their conclusion that the CES-D exhibited measurement invariance across these two culturally diverse groups of adolescents. Rather, the researchers should have performed their substantive analyses with latent variables under the assumption of partial measurement invariance.

The CES-D contains four positively worded items: *Good* (4), *Hopeful* (8), *Happy* (12) and *Enjoy* (16) which were included to minimise the influence of a negative response set and also to measure positive affect. When calculating a total score these items are reversed scored. This scoring method assumes that both the positive and negative keyed items measure the same (although different aspects) construct of depressive symptomatology. It is also implicitly assumed that these positive items measure depressive symptomatology equivalently across sub-groups. This assumption has been tested in relation to whether people from Asian backgrounds may tend to give less positive responses to the Positive Affect CES-D items.

The possibility that people from Asian cultures might respond differently to the positively worded CES-D items compared with people from Western cultures first appears to have been examined by Noh et al. (1992). By comparing mean item values these researchers observed that Koreans appeared more likely to make dysphoric responses to the Positive Affect items: *Good* (4), *Hopeful* (8), *Happy* (12) and *Enjoy* (16) compared with North Americans. When positive items were included

in total CES-D scores Koreans had substantially higher mean scores than North Americans but when total scores were recalculated without these items the mean difference between these two groups was significantly reduced. Noh et al. recommended that these items should be omitted in Koreans samples.

The basic finding of Noh et al. (1992) was replicated by Iwata et al. (1994) who compared Japanese adolescents (12 to 15 years of age) with North American adolescents using data drawn from Garrison et al. (1989) and Schoenbach (1982). They found that while the average CES-D score was higher in the Japanese sample compared with the North American sample (means 18.6 and 16.6 respectively) overall responses to negatively worded items were similar between the two groups. These authors concluded that the tendency to make more dysphoric responses to the Positive Affect items had a negative effect on the psychometric properties of the CES-D and artificially raised total CES-D scores in the Japanese sample.

Iwata et al. (1994) concluded that the CES-D should be revised by changing the positively worded items to be negatively worded (e.g. Item 12 'I was happy' to 'I was unhappy'). Further evidence that Asian people may be reluctant to agree with the Positive Affect CES-D items was reported by Cheung and Bagley (1998) who found in a small sample of Hong Kong Chinese married couples that the slightly higher mean in their sample compared to North Americans was due to dysphoric responses to the CES-D Positive Affect items.

In summary, several studies have shown that Asians have on average higher CES-D total scores than non-Asians. It appears that this increase is related to the four positively worded CES-D items on which Asians give less positive responses compared with what might be expected given their responses to the negatively worded CES-D items. On the current evidence the suspicion of possible cultural bias is raised, but until more sophisticated analyses are performed it is only one possible explanation for the discrepancy. An alternative explanation is simply that Asians on average experience less positive affect than non-Asians and therefore quite correctly are shown by the CES-D to have higher levels of depressive symptomatology. Thus it remains to be shown that for the same level of depressive symptomatology, Asians on average report lower levels of positive affect than non-Asians.

CES-D measurement invariance across gender

Five studies have examined possible gender bias in the CES-D using MG-CFA. In the most widely known of these, Roberts et al. (1990a) with a large community sample (N = 2160) of young adolescents performed a LISREL CFA to test the fit of the original CES-D four factor model. The goodness-of-fit index was 0.97 and the mean square residual was 0.029 providing evidence that this model fitted their data well. Using this model a MG-CFA was performed and item factor loadings were found not to be invariant across boys and girls.

Further analyses by Roberts et al. (1990a) indicated that the factor loadings for two items were different between boys and girls. Item 17 (*Cry*) had a higher loading to the Depressive Affect factor for girls and Item 2 (*Appetite*) had a lower loading to the Somatic factor for girls. When the loadings for these two items were estimated separately and all other factor loadings were constrained to be equal across genders the fit of the model improved significantly and did not differ from a fully unconstrained model. Roberts et al. interpreted these findings as indicating that there were few dramatic differences between boys and girls on the CES-D.

Beals et al. (1991) in a sample of 605 American Indian college students (mean age = 25 years) performed a LISREL MG-CFA to test for gender differences in the factor loadings of CES-D items. Using a three factor model of the CES-D (Depressed Affect and Somatic factors combined) they found that the loss of fit between a model in which the factor loadings were constrained to be equal compared with a model in which the loadings were allowed to vary between males and females was statistically significant. Two items were identified as being responsible for the loss of fit, Item 17 (*Cry*) (higher loading for females) and Item 12 (*Happy*) (higher loading for males).

Beals et al. (1991, p. 627) concluded that because Item 17 (C*ry*) had now been found twice (their study and Roberts et al. (1990a) to be different across gender this item was "… differentially valid across gender". A later study (Dick et al., 1994), carried out by the same research team of Beals et al. (1991) in a small (N = 188) sample of American Indian adolescents, however, found that when a three factor model was tested across gender, item factor loadings were not statistically different. Unfortunately, very few other details about this analysis are provided.

In an important study from a methodological perspective, Stommel et al. (1993) with a sample of adult cancer patients and their care-givers (average 63 years of age) tested the CES-D for gender bias in a CFA framework. The software package used for the analyses is unclear but probably it was LISREL or possibly EQS. Stommel et al. noted a very skewed response pattern for the Interpersonal factor items: *Unfriendly* (15) and *Dislike* (19) and on this basis removed them from further analysis. This left a three factor CES-D model with 18 items. Through a series Lagrange Multiplier tests, three items: *Failure* (9), *Talk* (13) and *Cry* (17) were found to have significantly different male and female factor loadings and in addition, the inter-factor correlation between Depressed Affect and Somatic was found to be significantly different.

Stommel et al. (1993) considered that the result for Item 9 (*Failure*) was due to that item's poor psychometric properties but the results relating to Item 13 (*Talk*) and Item 17 (*Cry*) were considered to be robust. In order to assist the interpretation of these results they then carried out a rather unusual type of CFA. In this analysis the two discrepant items: *Talk* (13) and *Cry* (17) and the remaining items were simultaneously regressed on a dummy variable for gender. Based on the direction of the regression coefficient (positive or negative) this analysis indicated that men who otherwise had the same level of depressive symptomatology as women were less likely to report crying spells but more likely to indicate that they talked less.

Stommel et al. (1993) recalculated the mean difference between men and women's total CES-D score and found that without the two interpersonal items and the three biased items the mean difference between men and women was reduced from a difference of 1.6 points to 1.3. The revised 15 item scale was highly correlated (0.98) with the full 20 item scale and it was recommended that being shorter and unbiased this scale should be preferred. Stommel et al. noted that his findings were only partially consistent with those of Roberts et al. (1990a) (with both researchers finding gender differences on Item 17 (C*ry*)) and speculated that the discrepancy between the two studies might be due to their different samples (older adults versus young adolescents).

The results of Stommel et al. (1993) have commonsense appeal but two aspects of their analysis warrants questioning. First, the analysis had certain similarities to what are termed MIMIC analyses (see Christensen et al., 1999; Gallo, Anthony & Muthén, 1994). These analyses are in effect tests for scalar invariance (see later for discussion) and are usually only performed on items which have been demonstrated to exhibit metric equivalence, that is invariant factor loadings. According to Bollen

(1989, p. 366) "Applications to date require that at a minimum the invariance of form and the invariance of factor loadings should hold before testing restrictions on means and intercepts. I follow this convention". Contrary to this convention Stommel et al. tested items which had specifically been shown not to exhibit metric equivalence.

Second, MIMIC analyses to date have tested items individually as opposed to testing all items simultaneously in the same analysis. This means that the effect of gender on individual CES-D items (other than the two discrepant items: *Talk* (13) and *Cry* (17)) were not tested for scalar invariance. Because of these deficiencies, as well as the fact that the sample was older adult cancer patients, the results of Stommel et al. (1993) provide little guidance to whether the CES-D would exhibit gender measurement invariance in a young adolescent sample.

The fifth study to examine possible gender bias in the CES-D using MG-CFA, was carried out by Breithaupt and Zumbo (2002) in a large longitudinal population based sample of 6621 elderly participants. Somewhat unusually a binary scoring system was employed with items scored either for a zero (symptom not present) or a one (for symptom present). A single latent factor was specified as the measurement model and the regression path for Item 6 (*Depress*) was set to be equal (that is act as the marker item) across comparison groups. An asymptotic covariance matrix based on tetrachoric item correlations (appropriate for binary data) available in LISREL was used for the analysis.

Three main tests of measurement invariance were performed by Breithaupt and Zumbo (2002). In the first test of gender invariance, configural invariance was examined and the results indicated that item responses (manifest variables) loaded on to the same construct. This finding suggested that males and females in the sample were employing the same conceptual frame of reference to the construct (depressive symptomatology) hypothesised to underlie CES-D items. Second, factor loadings were constrained to be equal across groups and a significant reduction in closeness of fit was observed. This finding also suggested that one or more item factor loadings were not equivalent across groups. Finally error variances were constrained to be equal across groups and no significant deterioration in the fit indices was detected suggesting that the items were equally reliable for males and females.

The study by Breithaupt and Zumbo (2002) has several commendable features. It is one of the few SEM analyses of the CES-D to have used an estimation technique appropriate for categorical data and importantly justifications were clearly presented for the specification of the measurement model and the selection of the marker item. On the negative side, noting that Zumbo has argued (Gelin & Zumbo, 2003) that DIF results might change depending on how items are scored, the CES-D binary scoring system employed in this study makes it unclear whether the results would apply to the CES-D when scored conventionally. A further criticism is that the researchers on finding a lack of CES-D gender measurement invariance failed to go on to test which particular items caused this invariance.

In summary, researchers using CFA to examine gender CES-D measurement invariance have (with the exceptions of Stommel et al., 1993; Breithaupt & Zumbo, 2002) confined their analyses to examining whether CES-D models with item factor loadings constrained to be equal across gender fit the data as well as models with the factor loading unconstrained. Several items: *Cry* (17), *Appetite* (2), *Failure* (9), *Talk* (13) and *Happy* (12) have been shown to have significantly different factor loading across gender but only Item 17 (*Cry*) has been found not to be invariant on more than one occasion.

Perhaps because of the rather conflicting results produced from the existing gender CES-D studies most researchers continue to sum all 20 CES-D items for both males and females and to assume that these scores are comparable. Only one research team (viz. Aseltine et al., 1998) appears to be acting on the evidence of potential gender CES-D bias by excluding Item 17 (*Cry*) when performing CES-D gender comparisons. Whether this is appropriate is difficult to judge because the existing evidence either for or against gender CES-D measurement invariance is weak. The evidence is weak first because of methodological flaws in the existing studies and second because studies to date have only partially tested for measurement invariance.

Methodological weaknesses in CES-D SEM analyses

Previous SEM analyses of the CES-D, as far as it can be ascertained, have been performed with the LISREL software program. This is to be expected given that LISREL was the dominant SEM software package during the 1990s when most of these analyses were published. The LISREL software program includes a number of different estimation techniques but in the majority of analyses the estimation technique termed 'Maximum Likelihood (ML)' was used. The ML estimation technique assumes, among other things, that the data used to test models arise from a multivariate normal distribution (MVN).

In an important paper setting out the principles and practice for reporting SEM analyses, McDonald and Ho (2002) note that much social and behavioural science data may fail to satisfy the assumption of MVN. These authors recommend that researchers apply the normality tests available in SEM software and if a problem exists this should be reported. In the CES-D gender SEM analyses reviewed, only one reported whether normality tests had been performed. This study, Beals et al. (1991) provided few details of these tests, but simply stated that only moderate levels of multivariate kurtosis were found in their data and thus a ML estimation would perform adequately.

The level of normality testing applied in SEM analyses of the CES-D could not be considered adequate. This is particularly troublesome because it is unlikely that the assumption of MVN would be met in a typically very skewed CES-D data set. The list of potential practical problems inherent in a ML factor analysis of ordinal data where the assumption of MVN is not met is daunting. These problems include biased estimates of factor loadings (Dolan, 1994, Olsson, 1979), inflated chi-square test statistics (Amemiya & Anderson, 1990), deflated standard errors (Hoijtink, Rooks & Wilmink, 1999; Satorra & Bentler, 1994) and the discovery of response specific difficulty factors (Benstein & Teng, 1989; Schaie & Hertzog, 1985).

How serious is this problem? Well, as Rigdon (1996, ¶ 2) wryly notes: "Any significant deviation from normality will be taken by critics as a weakness, while being dismissed by authors as inconsequential". Unfortunately there is no widely accepted general rule of thumb to indicate how far from normal data can be before these problems become evident. In one of the few simulation studies of factor analysis with non-normal categorical variables Muthén and Kaplan (1985) found that if variables have skewnesses and kurtoses between –1.0 and 1.0 then little distortion will occur from using the ML technique. CES-D values exceeding these limits have been reported regularly since Radloff's (1977) original CES-D study.

Clearly the general issue of violating the assumption of MVN in SEM is not inconsequential and much effort has been directed to developing strategies in response to it. To date work has focussed on the problems of the inflated chi-square test of model fit and the deflated standard errors used to test the significance of parameter estimates. The consequence of an inflated chi-square statistic of model fit is that it will lead researchers to reject models that may not be false. The consequence of deflated standard errors is that this will lead researchers to accept that some parameter estimates are statistically significantly different from zero when in fact this is not the case (i.e. a Type 1 error).

One approach to these problems is to perform a normalising transformation on ordinal items prior to ML estimation but this creates difficulties in the interpretation of the parameter estimates (Rasmussen & Dunlap, 1991). Another approach, introduced by Satorra and Bentler (2001), is to adjust the obtained chi-square statistic and standard errors to take into account the non-normality of the sample data. These procedures are implemented in the EQS program (Bentler, 1995) and in effect they adjust downward the obtained model fit chi-square and adjust upwards the standard errors of parameter estimates by the amount of non-normality in the sample data. A second approach implemented in the AMOS (Arbuckle, 1997) software package is not to adjust the chi-square statistic but to adjust the critical value of the chi-square test by a process termed 'bootstrapping' (see Bollen & Stine, 1992).

These solutions are supported by simulation studies that show with a reasonable number of response categories (> 6) and in the absence of excessive skewness, factor loadings and other parameter estimates will remain valid. But this overlooks the fact that many psychological scales employ fewer response categories (often four or five options) and, particularly for self-report depression scales, are excessively skewed with strong floor effects. In this case, the parameter estimates themselves are likely to be biased and not remedied by adjusted chi-square statistics and standard errors. As Muthén (2001b, ¶ 2) explains:

For variables with strong floor or ceiling effects (say more than 50% pile up at the bottom or top of the scale), a model with linear relationships between variables is theoretically not realistic (for instance, residuals cannot be both positive and negative at the scale end points as a linear model assumes), which means that analyzing the usual continuous-variable sample statistics is not appropriate. So, the problem is not solved by using chi-square and standard error formulas that are robust against deviations from non-normality - the parameter estimates themselves are suspect. Treating the variables as ordered categorical (ordinal) is a way to avoid this problem.

Given that CES-D items are ordinal, treating them as such in a SEM analysis appears to be rather self-evident. There are however good reasons why so few substantive CES-D researchers have done this. In LISREL, an analysis of ordinal variables is achieved by using the weighted least squares (WLS) estimator with polychoric correlations and the asymptotic covariance matrix (see Jöreskog, 2001a,b,c; 2002). The WLS estimator is an asymptotic distribution free estimator that does not require MVN. It is assumed, however, that there is an underlying continuous variable for each ordinal variable which is reflected in the ordinal variables being bivariately normally distributed.

The LISREL WLS approach represents a clear improvement to ML estimation for categorical variable SEM analyses, but there are two potential problems. The first is that the WLS estimator requires a very large sample size, between 2500 and 5000 observations (Hu, Bentler & Kano, 1992; West, Finch & Curran, 1995). Few CES-D studies to date would have met this requirement. The second possible difficulty is the

assumption that there is an underlying continuous variable for each ordinal variable which is reflected in the ordinal variables being bivariately normally distributed. It is not known whether this assumption would be justified for CES-D items and conceivably it would not be.

A subsidiary aim of the present study is to test the normality assumptions required for SEM analyses of the CES-D and to compare ML and WLS estimation techniques using a large sample of adolescents. These analyses will provide an indication of the extent of the possible problems in previous SEM analyses of the CES-D which have relied solely on the ML technique. The analyses will be carried out using LISREL and a specialised SEM software program called M*plus* which has particular strengths (detailed in Chapter 7) for the analysis of categorical data.

Conceptual flaws in CES-D SEM analyses

The key conceptual flaw for existing CFAs of bias in the CES-D is that to date only two levels of measurement invariance have been tested. At the first level (configural invariance) investigators have examined whether the factor structure is similar across groups. At the second level, metric invariance or weak factorial invariance as it is sometimes termed, has been tested by examining whether factor loadings are equivalent between groups. Both configural and metric invariance are necessary conditions for full measurement invariance but they are not in themselves sufficient.

Another set of hypotheses concerns the intercepts for the measurement equations. It could be that one group tends to respond systematically higher or lower to the questions even if both groups have the same slope [ie factor loadings] relating the measures to the latent variables. This would be revealed in a difference in the intercepts for the same indicator for males versus females. (Bollen, 1989, p. 367)

This third level of measurement invariance is known as ‘scalar invariance’ or in the terminology of Meredith (1993) ‘strong factorial invariance’ or ‘additive bias’. Unless this bias is removed, comparisons across groups based on such additively biased items are meaningless. As Muthén (2000, ¶ 1) explains:

> One needs invariant measurement intercepts in addition to invariant slopes to be able to compare observed means - if this invariance is not present, two people with the same observed value have different factor values (and are therefore different).

The importance of scalar invariance has not always been fully appreciated and substantive examples in the CES-D literature can be found of researchers erroneously concluding that because their CFA results showed equivalent factor loadings between groups this proved that CES-D scores could be validly compared across groups (see Chan, 2000 for a discussion of this issue).

One of the best accounts of the distinction between metric and scalar invariance is provided by Reise, Smith and Furr (2001, p. 85) in the context of a discussion of measurement invariance issues and a neuroticism scale. Given the current ambiguity in the literature it is worth presenting in detail:

Tests of weak factorial invariance evaluate the degree to which the strength of the indicator (facets) to latent trait relationships (i.e. factor loadings) are equivalent across gender groups. Finding weak factorial invariance demonstrates that the indicators of neuroticism function equivalently for men and women. This in turn implies qualitative similarity in trait manifestation and that the same latent variable is

being measured in both groups. Weak factorial invariance does *not* mean that the scale is free of differential scale functioning (Raju, van der Linden, & Fleer, 1995). A measurement instrument displays differential scale functioning when members from different groups who have the same position on a latent variable do not have the same expected raw score on the scale. Under a CSA [multiple-group covariance structure analysis] framework, differential scale functioning can only be evaluated by tests of strong factorial invariance as described next.

Strong factorial invariance is tested by adding intercept terms into the factor model and then evaluating whether, for each of the facets, the regression of the facet on the latent trait is equivalent across groups (Meredith, 1993; Millsap, 1998). Strong factorial invariance holds when the regression parameters (intercept and loading) shown in Equation 1 are equal across groups.

(1) $$FacetRawScore = Intercept + Loading\ (Factor)$$

Finding strong factorial invariance implies: (a) mean differences between groups on the latent variable are identifiable and can be interpreted as reflecting valid differences on the latent trait underlying facet responses, and (b) mean raw score differences on the facets are completely explainable by mean differences on the latent variable. These properties have further implications for evaluating differential prediction as elaborated in Millsap (1998). If strong factorial invariance is not found, this implies a gender by item content interaction and the interpretation of gender differences on the neuroticism dimension becomes more complex.

Further theoretical accounts of the distinction between metric and scalar invariance are available (Cole et al., 1993; Horn & McArdle, 1992; Millsap, 1998; Pentz & Chou, 1994; Raju, Laffitte & Byrne, 2002; Steenkamp & Baumgartner, 1998). Substantive examples are rare but Cheung and Rensvold (2000) tested both factor loadings and intercepts in a cross-cultural analysis of work orientation. As it turned out on one construct (job content) factor loadings were equivalent but intercepts were not and as a result this construct could not be validly compared across cultures. A second example is provided by Lubke, Dolan and Kelderman (2001) in an interesting analysis of black – white group differences on cognitive tests. To date there has not been one MG-CFA to examine whether the CES-D meets the requirement for scalar invariance.

A second conceptual shortcoming in the existing CES-D gender literature is that researchers have not evaluated what impact potentially biased CES-D items have at the total score level. For example, as reviewed earlier, three studies have identified that the factor loading for Item 17 (*Cry*) is not invariant across gender. This is important, but most research and clinical use of psychological scales is conducted using total scores. The question therefore that has most relevance is whether items that comprise the scale yield biased test results after they have been summed together to produce a total score (Marshall et al., 1997).

This issue is not as simple as it first might appear. The existence of biased items in a test does not prove that total scores from the test are biased. This is because it is possible that the bias may cancel itself out. For example a sub-group of items might be biased against males and a different sub-group of items biased against females. When these items are combined the effect of this bias at the scale level is eliminated. In the IRT literature the terms 'DIF amplification' and 'DIF cancellation' are sometimes used to describe an item's contribution to overall differential test functioning (DTF: Nandakumar, 1993; Shealy & Stout, 1993).

In a similar vein, but from a factor analytic perspective, Labouvie and Ruetsch (1995) argued that because most group comparisons are made at the total scale level, measurement invariance is not required for items individually but only for the set of items as a whole (but see Meredith, 1995; Nesselroade, 1995). On this argument it is possible that the approach of Aseltine et al. (1998) to leave out Item 17 (*Cry*) could itself actually **cause** CES-D total test scores to become biased.

In summary, there has been insufficient attention paid to the measurement properties of the CES-D when used to make gender comparisons. The existing evidence is conflicting and it is not clear whether the group gender differences in depressive symptomatology based on CES-D scores are in fact meaningful. Previous analyses of gender bias in the CES-D have not addressed the issue of scalar invariance and the ordinal nature of the CES-D response format has been largely ignored. In order to be fair it should be noted that software for testing scalar invariance (particularly for ordinal data) is only now becoming widely available and it is only relatively recently that a framework to guide appropriate research practice for testing measurement invariance using CFA has been developed.

The deficiencies identified in the existing literature are not peculiar to the CES-D. For example, the MMPI (Hathaway & McKinley, 1940) is one of the most frequently used psychological tests in the world. Virtually since its inception there has been considerable debate about whether black – white differences on the MMPI scales are valid. These MMPI differences have very significant legal, ethical and practical consequences (Gottfredson, 1994) and so not surprisingly the literature regarding MMPI race differences is both voluminous and polemic. What is surprising is that modern techniques (either IRT or CFA) for examining measurement invariance have only recently (Waller et al., 2000) been applied to the MMPI. In essence the situation for the CES-D is exactly the same.

School effects on adolescent depression

The assumption that the school social environment affects student mental health is reflected in material produced from an Australian Commonwealth funded mental health initiative titled ‘MindMatters’ which was quoted at the beginning of this study. For ease of reference it is repeated here: “During adolescence, the social environment of the school plays an important role in shaping current and future health” (Sheehan et al., 1999, p. 47). These claims are not recent Australian inventions but are backed up by emphatic statements from authoritative bodies such as the World Health Organisation for example: “Schools powerfully affect the psychosocial development of children” (Wolff, 1993, p. 1).

Unfortunately there is very little evidence to support these claims in so far as they relate to internalising disorders such as depression or anxiety. Only two studies (Larson, Raffaelli, Richards, Ham & Jewell, 1990; Sawyer, Sarris, Baghurst, Cornish & Kalucy, 1990) have examined school differences with internalising problems. Both showed higher rates of depression and emotional disorder in children attending schools of lower average socioeconomic status. These two studies, however, did not attempt to control for differences in student background characteristics and because students are not randomly allocated to schools this means that no firm conclusions about school effects on student mental health can be drawn from these results.

The lack of empirical evidence to support the assumption that schools powerfully influence student mental health was recently acknowledged by the NHMRC Health

Advancement Standing Committee in a major review of what constituted effective practice for promoting health in the school setting. The committee (NHMRC, 1996, p. 6) found that:

> While there is evidence to demonstrate causality between certain determinants and health and learning outcomes (e.g. excessive exposure to the sun and skin cancer), in many cases, such strong empirical evidence does not exist (e.g. a supportive psychosocial environment and mental health).

One possible reason for this lack of evidence for disorders such as depression or anxiety is the relative difficulty of detecting internalising mental health concerns compared with the more obvious problems created by externalising behaviour disorders.

School effects research carried out by educational researchers has focused on students' academic achievements (see Bosker & Witziers 1995, for a meta-analysis), disruptive behaviours such as absenteeism (Bryk & Thum, 1989; Rothman, 2001), and so-called 'dropping out' (Rumberger, 1995). The majority of this research has been carried out using a statistical technique known alternatively as hierarchical linear modelling (HLM: Bryk & Raudenbush, 1992) or multi-level modelling (MLN: Goldstein, 1987). The use of this statistical technique (for preference the term HLM is used) allows researchers to find out how much variation in an outcome variable lies within and between schools. It is generally accepted (but see Thrupp, 2001) that the use of HLM has produced a much better way of comparing schools (Goldstein, & Thomas, 1996).

The results from the school effectiveness research using this advanced statistical technique have shown that the background characteristics of students explain a considerable proportion of the variation in outcome variables related to educational achievement. But small independent school effects have also been evident. These small school effects are important because they act on very large populations and produce differential effects across an entire population of children and younger adolescents (Bosker & Witziers, 1995). School factors which have been identified as important for educational achievement and behavioural problems include the social composition of schools, the organisational effectiveness of the school, and the quality of the social relationships in the school.

That the school has not received attention by mental health researchers is perplexing given the large number of child and adolescent mental health research studies carried out with students clustered within schools. While the focus of the research effort has been at the level of the individual and family, speculatively schools themselves may contain both risk and protective factors. Examples of possible school protective factors would include supportive relationships with peers, counselling and pastoral care, structured sporting and academic activities offering opportunities for pleasure and mastery and overall a sense of belonging to the school community. Levels of bullying and drug use on the other hand would be possible risk factors.

Some time ago in a seminal review of the school effectiveness literature Rutter (1983) arrived at two main important conclusions. The first conclusion was that schools differed significantly on relevant measures of student success including social functioning. The second was that these differences were not merely a function of student intake but rather due to certain qualities of schools themselves. Rutter's conclusions suggest that student mental health could be substantially improved by interventions at the school level modelled on the practices found in the most successful schools. A substantive aim of the present study is to take the first steps

towards quantifying the magnitude of school effects on student levels of depressive symptomatology.

This substantive aim also has a methodological importance. Most CES-D studies carried out with samples of adolescents have recruited their samples through schools. For example, in Roberts et al. (1990a) the participants were recruited from four schools in Oregon while in Garrison et al. (1991a, 1991b) the participants were recruited from six schools in South Carolina. These school based mental health samples have been popular among adolescent depression researchers because they avoid the well known difficulties of generalising from young people presenting to hospitals or clinics and have an intuitive appeal of representativeness about them. Cooperative school staff can facilitate the testing and studies with large numbers of subjects can be implemented at relatively low cost.

The sampling design employed in most large scale school based studies of adolescent depression is known as a 'two stage cluster sample design' (Ross, 1988). In this design schools are first selected and then in the second stage students from within those schools are recruited. Although the samples produced from this types of design are often described as 'community samples' (as opposed to a clinic based sample) in the mental health literature, they are not equivalent to a simple random sample of adolescents. This is because students are not allocated to schools at random and within each school they will share common experiences. Consequently the results from a two stage cluster (school) sample design will be more homogenous than those of a simple random sample of students drawn from the population of all schools (Aitkin & Longford, 1985).

The magnitude of the dependence or clustering in school based samples is commonly estimated by what is known as the 'intraclass correlation coefficient'. From a methodological perspective a failure to take account of non-zero intraclass correlation coefficients within sampling units (e.g. schools) in statistical analyses results in biased standard errors of parameter estimates usually in the direction which exaggerates that parameter's significance (Norton, Bieler, Ennett & Zarkin, 1996). Estimates of this dependency in clusters (the intraclass correlation coefficient) vary for different populations and outcome variables but correlations of approximately 0.2 have been found to be reasonably accurate estimates of student homogeneity for achievement variables within schools (Ross, 1988).

For mental health variables, including CES-D scores, researchers have not yet calculated intraclass correlation coefficients. If it is the case that school effects on student mental health are significant then it follows that the results from previous standard statistical analyses which have ignored this clustering might be in error. The present study uses data generated from a two stage cluster sampling design and the extent of clustering in the data can provide a guide as to the degree to which previous analyses might be in error. In summary, possible school effects on student mental health may have important implications for the provision of mental health programs in schools as well as for the analysis of mental health data collected from school students. As far as the author is aware this study is the first to systematically investigate school effects on student levels of depressive symptomatology.

3

Research Design and Questions

Research design

The present study uses data collected in a longitudinal study carried out by the Southern Child and Adolescent Mental Health Service in South Australia (CAMHS) called the Early Detection of Emotional Disorders (EDED) program. The EDED program had two goals. First, it sought to increase knowledge about the development of emotional disorders and suicidal behaviours in young people, and second, to identify those students in the program experiencing emotional disorders or engaging in suicidal behaviours, so that assistance could be provided to them. The EDED program ran for three years during which the same group of students were tested each year as they progressed from Year 8 (the first year of high school) to Year 10 of high school. Overall around 2500 students from 26 public and private high schools took part in EDED each year for the three years of the program.

The design for the EDED program is known as a prospective time series study because it involved following the same sample at successive time points, with corresponding increases in the age of the group under survey. Data were collected prospectively which means that subjects were followed in time, as opposed to retrospective data collection which usually involves extracting multiple measurements on each person from historical records. The program was approved in 1994 by the Committee on Clinical Investigations (Ethics) of Flinders Medical Centre and by the Department of Education and Children's Services. An overview of the EDED program is available in Martin et al. (1997).

The present author was responsible for the line management of the EDED program between 1995 and 1997. This included, with the assistance from research, clinical and clerical staff, administration of the questionnaires in schools, scanning and checking questionnaires and developing the criteria to identify students requiring follow-up. Lists of the initials and dates of birth of students identified as at risk of developing emotional disorders were provided to school counsellors and the present

author was responsible for liaising with school staff and parents to ensure that the most appropriate response was provided to these students. The present author reported to the Chief Investigators of the EDED program, Professor Graham Martin (then Director of Southern CAMHS) and Dr Stephen Allison (Senior Consultant Psychiatrist, Southern CAMHS). These personnel are also supervisors of the present study.

Data in the program were collected by way of a 16 page self-report numbered questionnaire called the Youth Assessment Checklist (YAC). Students identified themselves on the questionnaire by recording their initials (first, middle and last) and their date of birth. The YAC consisted of items relating to demographic characteristics, self assessed academic performance, music preferences, physical and sexual abuse, alcohol and drug use and life events, including experiences of suicide.

The YAC contained quite a number of social-psychological scales including the: *General Functioning subscale of the Family Assessment Device* (FAD-GF: Byles, Byrne, Boyle & Offord, 1988); *Parental Bonding Instrument* (PBI: Parker, Tupling & Brown, 1979); *Adolescent Suicide Questionnaire* (ASQ: Pearce & Martin, 1994); *Self-Reported Delinquency Scale* (SRDS: Rushton & Chrisjohn, 1981) and the *Beck Hopelessness Scale* (BHS: Beck, Weissman, Lester & Trexler, 1974).

The CES-D scale was included in the YAC. It was provided early in the questionnaire following the PBI and the FAD-GF scales. The CES-D played an important role in the screening aspect of the EDED program with CES-D scores weighted heavily in the calculation of a student's individual risk score. Risk scores determined whether or not a student received a follow-up interview with a school counsellor and data checking (particularly of the CES-D) was very thorough.

The present study uses three main pieces of information from the YAC. These are the gender of the student, the school they attended and their CES-D responses. In the next section details about school and student selection, questionnaire administration, the treatment of missing data and other methodological issues are set out. The EDED program comprised a myriad of complexities mainly revolving around the identification of so-called 'at risk' students and ensuring an appropriate response to them. The research design of the present study on the other hand is relatively straightforward and can be expressed simply by saying that CES-D data were collected from the same 2500 odd students from 26 high schools each year for three years starting from when they were in Year 8 and running through to when they were in Year 10.

Research questions

The broad aims of the present study are to: (a) examine whether the CES-D measures depressive symptomatology equivalently across gender for young adolescents; and (b) to determine whether high schools exert effects on student levels of depressive symptomatology independently of individual level characteristics. These broad aims are developed into a series of research questions organised around the main statistical techniques employed in the present study.

With simple descriptive statistics:

1. What overall levels of depressive symptomatology will be reported by Australian adolescents compared with their American counterparts?

2. Do girls show higher total CES-D scores than boys and do CES-D scores increase during early adolescence (Years 8 to 10: Ages 13 to 15 years)?
3. Are gender and year level (age) differences at the total score level reflected in differences at the factor, item and response option level?

With IRT models:

4. Are individual CES-D item scores equivalent across gender and year level at equal levels of depressive symptomatology?
5. If item scores are not equivalent across gender and year level, what impact does this have on total scores?
6. Are there gender or year level differences for CES-D items at the response option level, controlling for levels of depressive symptomatology?
7. What is the relative quality of the information provided by the CES-D across different levels of depressive symptomatology?

With SEM techniques:

8. From the variety of factor models proposed for the CES-D which provides the best fit to the data?
9. Does the CES-D exhibit unidimensionality in an adolescent population?
10. To what extent might previous SEM analyses which have ignored the ordinal nature of the CES-D be in error?
11. Do boys and girls and students across year levels employ the same conceptual frame of reference to the construct hypothesised to underlie the CES-D (configural invariance)?
12. Are the CES-D SEM measurement model parameters (factor loadings & thresholds) equivalent across gender and year level (metric & scalar invariance)?
13. Are the CES-D SEM structural model parameters (factor variances, item residual variances & latent means) equivalent across gender and year level?
14. What is the impact of any lack of gender or age measurement invariance on CES-D total scores?

With HLM techniques:

15. What is the extent of clustering for school based CES-D data?
16. Does the extent of clustering for school based CES-D school increase during the first three years of high school consistent with a school effect on student depressive symptomatology?

4

Method

School and student recruitment

In 1994, 24 State funded high schools within the catchment area of CAMHS were invited to join the EDED program. These government schools with secondary students (Years 8 to 12) comprise an enrolled population of nearly 20,000 students and constitute around a third of South Australia's total secondary school population. Of the 24 schools approached, 16 agreed to participate and in 1995, Year 8 students from these schools completed questionnaires. The 16 State funded schools principally drew students from lower to upper middle socioeconomic areas. In 1995, funding became available to conduct the program in Independent School Board (ISB) schools. Expressions of interest were sought from 85 private schools with 10 schools agreeing to take part. The 10 ISB or private schools included some of South Australia's most expensive schools with tuition fees in the order of $10,000 per annum at the time the study was conducted.

Public schools participated in the EDED program between 1994 – 1996 and private schools between 1995 – 1997. All the public schools were co-educational with both male and female students. Four of the private schools were single sex only (two girls only and two boys only). During the course of the EDED program two schools withdrew. One was a public school which closed in 1996 – the third year of data collection. The second was a private school (one of the single sex schools) which experienced difficulties in arranging class time for the questionnaire to be completed by students. This school also withdrew in the third year of data collection (1997). This means that data were available for 24 schools for the full three years and two schools for two years.

Each year the parents of students taking part in the EDED program received a written explanation of the program and its purpose, with a clear explanation of the processes adopted to secure confidentiality. A permission form was enclosed to be returned if consent was **not** granted - a process known as 'assent'. The lack of a permission form was taken to indicate that parents did not object to their child taking part in the program. Letters to parents were distributed by school staff (to maintain the confidentiality of student addresses) and it was not possible to determine now the

exact number of parents and students who were initially approached to participate. In addition, on testing day some students were absent, had other engagements or chose not to complete the questionnaires. These numbers are also not known but based on school student enrolment figures the best estimate is that approximately 85 per cent of eligible students participated each year.

The number of students completing questionnaires by gender is shown for each wave of EDED in Table 3. At each wave (or year level) more boys than girls took part in the program. In South Australian high schools (up to and including Year 10) there are approximately the same number of boys enrolled as girls. The gender discrepancy in the present sample therefore raises the possibility of a differential gender response rate to the EDED program.

Table 3 Number of students by year level and gender

	Year 8	Year 9	Year 10
Boys	1422	1362	1310
Girls	1125	1090	961
Missing gender	5	9	6
Total	2552	2461	2277

Data presented much later in this book (see Chapter 9, Table 54) shows that in each coeducational school approximately the same number of boys completed questionnaires as did girls. Of the four single sex private schools two were boys only and two were girls only. While the two boy only schools were fairly large the two girl only schools were relatively small. This sampling anomaly at school level and not a differential gender response rate accounts for the difference between the number of boys and girls in the sample.

Table 3 shows that although some attrition is evident in terms of overall numbers of students taking part in the EDED program the extent of this is relatively minor (Year 8: 2552; Year 9: 2461; Year 10: 2277). The gender ratio across the three year levels remained stable except for Year 10 where a slightly greater proportion of boys took part (Year 8: 55.8%; Year 9: 55.5%; Year 10: 57.7%). The most likely reason for the Year 10 increase in the proportion of boys in the program is that in the third wave of the program one of the two schools to withdraw was a girls only school.

Data collection

With some minor exceptions each year the YAC was administered in Term 2 or Term 3 of the school year and over two lessons of morning school time. Questionnaires were completed under the supervision of teachers. Teachers were instructed to inform students that their participation was voluntary and that non-participation

would have no consequences whatsoever. Students were asked to record their initials and dates of birth and informed that, if they appeared to be experiencing personal problems, these initials and dates of birth would be given to school counsellors who would then seek to conduct a clarification interview with them. Students were told that not every one might be able to complete the questionnaire but to do the best that they could. The questionnaire took about one and a half hours to complete for those who experienced the most difficulty.

The EDED program was both a research study and an intervention program. Students were made aware that their responses to the questionnaire were confidential but if they scored highly on some scales then their initials and dates of birth would be passed on to school counsellors. In the first year of the program this aspect of the program was not widely understood but in the second and third years a high proportion of students did not record their correct initials and dates of birth on their questionnaires. This meant that many students could not be identified even if their responses indicated that they might benefit from some assistance. It also meant that matching of questionnaires across the three waves of the study became virtually impossible for around nearly one half of the sample.

A comparison of those students able to be matched across the three waves and those not able to be matched revealed that the non-matched group scored more highly (showed higher levels of pathology) on many of the scales in the questionnaire. A dataset comprised only of those students able to be matched across all three waves therefore would be significantly biased towards low scoring students. As a consequence even though there was very little attrition across the three waves, since overall numbers in the EDED program remained high, the analysis of the longitudinal aspect of the EDED program is quite problematic. For present purposes longitudinal analyses are not required to answer the study questions.

Missing data

In any large study based on self-report questionnaires there will inevitably be some missing data and the EDED program is no exception. In this section the method used for dealing with missing data is outlined. At each data collection point a relatively small number (Year 8: 5; Year 9: 9; Year 10: 6) of students did not record their gender. Given that gender is necessary for nearly all the analyses and only a small number of students did not record their gender these cases are deleted.

Using the sample of students with their gender recorded (i.e. Year 8: 2547; Year 9: 2452; Year 10: 2271) missing data on the CES-D is examined. The percentage of students with missing CES-D items is shown in Table 4 for each year level and by gender. Most (around 80%) students provided complete data to all 20 CES-D items but a sizeable proportion (nearly 10%) missed only one item and a smaller number skipped between 2 and 20 items. The pattern of missing data for each item is also examined and is shown in Table 5.

Table 5 shows that for boys the proportion of missing data for each CES-D item varied between around 5 to 8 per cent. For girls the proportion of missing data for each CES-D item varied between around 2 to 4 per cent. Importantly, no single item for boys or girls appears to have a markedly higher probability of not being completed than any other item.

Table 4 Number of CES-D items not completed by gender and year level

% recorded	Boys			Girls		
Number of items missed	Year 8	Year 9	Year 10	Year 8	Year 9	Year 10
0	78.8	83.1	85.0	82.4	86.3	89.9
1	10.3	8.8	8.4	9.9	7.5	7.4
2	2.6	1.8	1.5	2.8	1.3	1.4
3	0.8	0.3	0.5	1.0	0.6	0.0
4	0.6	0.3	0.0	0.1	0.4	0.1
5	0.4	0.2	0.2	0.2	0.0	0.1
6-10	0.6	0.4	0.4	1.2	0.7	0.0
11-19	2.9	1.1	1.2	1.6	1.0	0.0
20	3.0	3.9	3.1	1.2	2.3	1.0
N	1432	1362	1310	1125	1090	961

Table 5 Per cent of CES-D items not completed by gender and year level

% recorded	Boys			Girls		
CES-D items	Year 8	Year 9	Year 10	Year 8	Year 9	Year 10
Bothered	4.9	4.6	3.6	2.2	2.9	1.7
Appetite	5.1	5.2	4.0	1.9	3.0	1.7
Blues	7.0	5.3	4.9	3.6	3.5	1.2
Good	6.3	6.2	5.3	4.2	3.8	2.1
Mind	6.0	5.4	4.7	3.4	3.6	1.4
Depress	6.3	5.2	4.3	3.4	3.6	1.8
Effort	7.0	6.2	5.0	3.8	3.9	1.7
Hopeful	7.7	6.5	4.7	4.8	4.0	2.1
Failure	6.5	5.7	5.0	3.6	3.9	1.5
Fearful	7.9	6.0	5.4	4.5	4.9	2.1
Sleep	6.6	5.5	4.7	3.6	3.8	1.4
Happy	6.0	5.5	4.8	3.6	3.9	1.8
Talk	6.5	5.8	5.0	4.1	3.9	1.6
Lonely	6.3	5.9	4.9	3.3	4.3	1.7
Unfriendly	6.1	6.1	4.5	3.9	4.5	1.7
Enjoy	6.8	6.0	5.0	3.7	4.2	1.6
Cry	7.7	6.5	5.2	4.4	4.2	1.4
Sad	7.5	6.2	4.8	4.4	4.1	2.1
Dislike	7.5	5.5	4.7	4.0	3.8	1.5
Get-going	7.9	6.1	5.0	4.5	4.2	2.3

Several options for dealing with the missing item level CES-D data are available. A simple strategy would be to delete all cases with any missing CES-D items at all. This would have the effect of reducing the sample size by around 20 per cent and discarding a considerable number of valid CES-D item responses (recall that about 10% of total sample missed only one item). Second, the convention (Radloff, 1977) of only including respondents who missed fewer than five items could be adopted. This convention is recommended for epidemiological studies where it is assumed that item responses are summed to produce a total score but in the present study most analyses are performed at the item level. In addition, several of the more complex statistical procedures used in the present study do not allow cases with any missing data.

One approach to handling missing data is to use a statistical procedure to calculate missing item level data based on the pattern of other non-missing item responses. These procedures are available in many SEM software packages but they are complex and typically make strong data distributional assumptions. In the present study a simpler approach was adopted. Based on the earlier examination of the pattern of missing CES-D data it is clear that very little would be gained (in terms of sample size) and much could be lost (in terms of introducing bias) from attempting to estimate responses to CES-D items for students who had missed many items.

It was decided therefore to estimate the item responses for students who had missed only one item. This approach avoids discarding the responses of nearly 10 per cent of the sample (who had completed 19 out of 20 items) and arguably the benefits of this approach outweigh the potential disadvantages. For each student who had missed only one CES-D item, the other 19 items were summed, divided by 19 and rounded to the nearest whole number. This value was used as the value for the one missing item.

The number of students receiving an 'imputed' value for one CES-D item were as follows (Year 8: 258; Year 9: 202; Year 10: 181). The effect on item, factor and total mean scores of this missing data procedure is examined for each gender separately and is shown in Table 6 (Boys) and Table 7 (Girls). Very small, mostly negligible differences, are evident for both boys and girls. In effect this means that students who missed more than one CES-D item (Year 8: 241; Year 9: 177; Year 10: 113) were excluded from all analyses.

This produced a final sample of Year 8: 2306 (Boys: 1268; Girls: 1038), Year 9: 2275 (Boys: 1252; Girls: 1023) and Year 10: 2158 (Boys: 1223; Girls: 935) all of whom (following data imputation) had complete data for gender and CES-D items. It is this single data set which will be used in all of the statistical analyses of this report. A key benefit of using a single data set is that this will allow the results across different (and complex) statistical procedures to be compared.

Table 6 Mean of CES-D items, factors and total score by year level before and after data imputation (Boys)

CES-D items	Year 8		Year 9		Year 10	
	Before	After	Before	After	Before	After
Bothered	.31	.31	.29	.30	.37	.37
Appetite	.28	.28	.26	.27	.29	.29
Blues	.32	.33	.31	.31	.28	.28
Good	1.08	1.07	.97	.96	.88	.88
Mind	.86	.86	.91	.91	.93	.92
Depress	.50	.50	.46	.46	.45	.45
Effort	1.21	1.20	1.05	1.05	.93	.92
Hopeful	1.30	1.28	1.17	1.16	1.16	1.16
Failure	.30	.29	.26	.26	.21	.21
Fearful	.25	.26	.24	.24	.22	.22
Sleep	.54	.55	.55	.55	.52	.52
Happy	.80	.80	.82	.82	.75	.75
Talk	.54	.54	.50	.51	.50	.50
Lonely	.36	.36	.33	.34	.33	.33
Unfriendly	.55	.55	.44	.44	.38	.38
Enjoy	.83	.83	.83	.83	.80	.79
Cry	.16	.16	.12	.12	.09	.10
Sad	.38	.38	.32	.32	.31	.31
Dislike	.49	.49	.42	.42	.36	.36
Get-going	.51	.51	.52	.52	.52	.53
Total item mean	.58	.57	.54	.54	.52	.51
Factor mean						
Depressed Affect	.33	.33	.30	.29	.28	.27
Positive Affect	1.01	.98	.95	.94	.90	.89
Somatic	.60	.61	.59	.59	.58	.57
Interpersonal	.52	.51	.43	.43	.37	.37

Table 7 Mean of CES-D items, factors and total score by year level before and after data imputation (Girls)

Items	Year 8		Year 9		Year 10	
	Before	After	Before	After	Before	After
Bothered	.47	.47	.57	.57	.64	.64
Appetite	.58	.58	.57	.57	.66	.66
Blues	.53	.53	.57	.57	.62	.62
Good	1.21	1.20	1.15	1.14	1.11	1.11
Mind	.86	.86	1.02	1.02	1.10	1.10
Depress	.72	.72	.74	.74	.77	.77
Effort	.94	.94	.83	.83	.80	.80
Hopeful	1.28	1.28	1.23	1.23	1.24	1.23
Failure	.38	.39	.35	.35	.36	.36
Fearful	.36	.37	.33	.33	.34	.34
Sleep	.77	.77	.76	.76	.79	.79
Happy	.80	.80	.79	.78	.81	.81
Talk	.62	.62	61	.60	.62	.62
Lonely	.54	.54	.53	.53	.54	.54
Unfriendly	.48	.48	.41	.41	.36	.37
Enjoy	.89	.89	.88	.88	.90	.89
Cry	.38	.38	.41	.41	.40	.41
Sad	.64	.64	.61	.61	.67	.68
Dislike	.66	.66	.57	.57	.56	.57
Get-going	.59	.59	.59	.59	.65	.66
Total item mean	.69	.69	.68	.67	.70	.70
Factor mean						
Depressed Affect	.51	.51	.51	.50	.53	.53
Positive Affect	1.05	1.04	1.01	1.00	1.01	1.01
Somatic	.69	.70	.71	.70	.75	.75
Interpersonal	.57	.57	.50	.49	.47	.47

Sample characteristics

The YAC included quite a number of general social-demographic questions, such as, the country of birth of the student, their family type, and the number of close friends the student had. Descriptive statistics from several of these questions are provided in Table 8.

Table 8 Sample demographics by year level and gender

	Year 8		Year 9		Year 10	
	Boys	Girls	Boys	Girls	Boys	Girls
N	1268	1038	1252	1023	1223	935
Age (years)	13.61	13.59	14.60	14.58	15.61	15.60
	%	%	%	%	%	%
Country of Birth						
Australia	92.7	93.8	92.4	93.1	90.8	92.1
UK.	2.7	2.0	2.7	2.3	2.9	2.3
Europe	0.2	0.5	0.5	0.6	0.5	0.9
New Zealand	0.6	0.3	0.7	0.4	0.7	0.6
Asia	0.4	0.4	0.4	1.1	0.4	0.6
Other	3.4	3.0	3.4	2.6	4.7	3.4
Live with						
Two natural parents	73.7	71.6	72.9	70.4	72.5	66.6
Mother alone	10.3	13.9	11.1	14.2	10.9	15.6
Mother and stepfather	9.0	7.5	9.5	8.8	8.5	8.8
Other						
Closeness of family						
Just a group of people	0.6	0.9	1.0	1.4	1.4	1.0
Not very close	6.6	6.8	7.5	10.3	8.6	11.5
Close	53.5	56.9	61.4	59.9	60.1	62.6
Very close	39.3	35.5	30.1	28.4	29.9	24.8
Academic performance						
Failing	1.5	0.9	2.3	1.6	2.7	1.6
Below average	6.9	4.4	9.1	6.1	9.2	5.3
Average	67.9	72.2	65.4	70.8	64.1	71.6
Above average	23.6	22.5	23.1	21.5	24.0	21.5
Number of close friends						
None	1.1	0.3	1.6	0.7	1.5	0.3
One	3.3	3.0	2.2	1.7	2.5	1.6
Two – three	26.4	20.8	21.8	20.4	25.5	24.8
Four or more	69.1	75.9	74.4	77.3	70.5	73.2
FAD (mean)	1.84	1.79	1.87	1.84	1.88	1.90
SD	0.45	0.49	0.47	0.52	0.47	0.51

Table 8 shows that in the first year of the EDED program the average age of students is around 13.5 years. Satisfyingly this increased by one year each year of the program. The majority (over 90%) of students taking part in the program were born in Australia and three quarters lived with both their natural parents. Less than 10 per cent of students indicated that they felt that their family was 'not very close' or 'just a group a people living together'.

The level of family functioning, as measured by the General Functioning subscale of the Family Assessment Device, is similar to ratings obtained in comparable overseas population based studies (Byles et al., 1988) and similar (although slightly higher showing worse functioning) to estimates from a community survey of children living in Western Australia (Silburn et al., 1996). Over the three year levels the majority of students rated their academic performance as average (around 70%) or above average (around 20%). Less than 10 per cent of students felt that their academic performance was below average or failing.

Definition of a high scoring CES-D case

Most of the statistical analyses in the present study are conducted using the CES-D as a continuous measure of depressive symptomatology. For a small number of analyses, students need to be categorised into low and high scoring groups. The first of these analyses concerns a comparison of the mean CES-D value for low and high scoring boys and girls. These results provide preliminary information about whether some CES-D items are more effective than others in discriminating between low and high scoring students and whether this varies by gender.

In the second analysis students need to be categorised into low or high scorers to investigate the possibility that school CES-D differences might be evident at severe levels of depressive phenomena. In order to carry out these analyses a cut-point for the CES-D needs to be determined so that students can be categorised as 'high scorers' if they score above a certain CES-D total score. In this section various methods for forming CES-D high scoring groups are canvassed and the approach adopted in the present study is outlined.

In the original CES-D reliability and validity study Radloff (1977) investigated whether CES-D scores discriminated between adults drawn from psychiatric inpatient samples and the general population. Radloff found that the average CES-D score for the psychiatric inpatient sample was substantially higher than the average score for the general population sample. In addition, Radloff examined the proportion of people in these samples scoring above what she termed 'an arbitrary cut-off score of 16'. Around 75 per cent of the psychiatric inpatient sample but only 21 per cent of the general population scored above this cut-point of 16. Subsequent to Radloff many researchers in community samples of adults have used either a score of 16 or alternatively the top 20 per cent to designate respondents as 'high scorers'.

During the 1990s the CES-D was used extensively with samples of adolescents and it was observed that mean scores were nearly twice those reported in most adult samples (Roberts et al., 1990a). As a consequence when a score of 16 and above was applied the prevalence of 'high scorers' or 'cases' equated to around 50 per cent of the sample. Given that the prevalence of clinical depression among adolescents has been estimated from epidemiological surveys to be around 5 per cent the criterion of a score of 16 or above appeared to be set too low.

Recognising this problem Garrison et al. (1991b) and Roberts et al. (1991) attempted to determine the best cut-point for the CES-D when used as a screening measure for clinical depression among adolescents. These researchers (as discussed in an earlier section) found that optimal screening cut-points were different for boys and girls. For girls, both research teams provided similar scores of 22 and 24 but for boys Garrison et al. (1991b) proposed a score of 12 and above while Roberts et al. recommended a score of 22 and above.

In the general CES-D adolescent research literature a variety of different cut-off scores have been used to create comparison groups of low and high scoring respondents. These include scores of 16 and above (Wells et al., 1987), 22 or above (Garrison et al., 1989), scores of 30 or above (Garrison et al., 1990), and the referred to earlier Garrison - Roberts screening cut-points (Gore et al., 1993). Barnes and Prosen (1985) suggested that adolescent depression scores be classified as mild, moderate or severe on the basis of scores of 16, 24 and 31 respectively.

An alternative approach, more closely aligned with a mental health disorder perspective, is to compute CES-D total scores ignoring the first two (Garrison et al., 1989) or three (Schoenbach, 1982) levels of CES-D item response formats. This means that unless an item symptom is reported at least 'a lot of the time' or 'most or all of the time' it is not counted. A cut-point can then be used with these new scores to create a group of respondents experiencing persistent symptoms of depression.

Table 9 Number and per cent of high scoring cases by gender and year level

	Boys		Girls		Total	
CES-D score	<22	≥22	<22	≥22	<22	≥22
Year 8						
N	1121	147	829	209	1950	356
%	88.4	11.6	79.9	20.1	84.6	15.4
Year 9						
N	1119	133	830	193	1949	326
%	89.4	10.6	81.1	18.9	85.7	14.3
Year 10						
N	1087	136	735	200	1822	336
%	88.9	11.1	78.6	21.4	84.4	15.6
Total N	3327	416	2394	602	5721	1018
%	88.9	11.1	79.9	20.1	84.9	15.1

Earlier, it was noted that Gotlib et al. (1995) had found that adolescents with high CES-D scores, even in the absence of clinical depression, experienced considerable impairment. The level of impairment was similar to that expressed by individuals with clinical depression. In that study the cut-point on the CES-D used to define a 'high score' was one standard deviation above the mean, which in their sample of older adolescents (17 years of age) equated to a CES-D score of 27.

In the present sample, which on average is slightly younger, the overall (both boys and girls) mean CES-D score across the three year levels was 12.08 with a standard deviation of 9.97. Using the one standard deviation rule this equates to a cut-point of 22. This figure is similar to other cut-points commonly used in the literature for adolescent samples and is adopted in the present study.

Table 9 shows the number of students with scores equal to or above 22 across the three year levels.

Table 9 shows that on the basis of scores equal to or above 22, around 15 per cent of the sample are designated as 'high scorers'. It is also evident that nearly twice as many girls as compared to boys are classified as 'high scorers'. In the analyses to follow a CES-D 'high scorer' is defined as a student scoring 22 CES-D points or above. For these categorical analyses a binary variable identified as 'CES-D22' is used.

Statistical software

Four main statistical software packages are used in the present study. These are: SPSS for Windows version 10.05 to carry out basic descriptive analyses; TestGraf (Ramsay, 2000) to perform the IRT analyses of the CES-D; M*plus* version 2.1 (Muthén & Muthén, 1998) for SEM analyses and HLM for Windows version 5.00 (Bryk, Raudenbush & Congdon, 1996) for multilevel analyses of possible school effects of depressive symptomatology. SPSS is a basic statistics package which is very widely used in the social sciences and needs no further description. The remaining packages are more specialised and are not as commonly used. A detailed description of each of these packages is provided in the results chapters as follows: TestGraf (Chapter 6), M*plus* (Chapter 7) and HLM (Chapter 9).

5

Descriptive Analyses of Gender and Year Level Differences

This chapter presents the results from a number of simple basic descriptive analyses of CES-D gender differences at the total score, factor, item and response option level. At the total score and factor level mean values are calculated for boys and girls across year levels (Year 8, 9 & 10) to show overall CES-D levels and the extent of possible gender differences in the EDED data set. The mean value of items comprising the four factors of the CES-D is also examined to explore the possibility that a group(s) of items might be responsible for the observed gender difference in total CES-D scores.

At the item level, the importance of individual items to boys and girls is assessed by ranking the means of items for each gender. If these two rank orders prove to be similar then this suggests that the saliency of CES-D items (symptoms) are equivalent for boys and girls. On the other hand, items deviating sharply from the general pattern of rank orders raises the suspicion of bias. In addition, for each gender the mean value of items is compared between low and high scoring students.

The purpose of this analysis is to examine whether high total CES-D scores arise from a relatively small increase in the mean value of a large number of items or alternatively from very high scores on a few selected items. It is also possible that the individual contribution of items to overall CES-D scores might vary by gender and this may help explain gender differences at the CES-D total score level.

At the response option level a technique outlined by Santor and Coyne (1997) is used to identify the CES-D items which best discriminate between boys and girls. For every item the proportion of boys and girls endorsing each response option is computed. Using the ordinal nature of the CES-D response format, a cumulative probability for each item is then calculated. Large differences in these cumulative probabilities between boys and girls indicate a different response option pattern. This analysis is repeated to examine the CES-D items that best discriminate between low

and high scoring students. This technique at the response option level provides information about possible CES-D item gender differences that might not be evident from analyses at the item mean value level. It will also assist in the interpretation of the results from the IRT and SEM analyses which are similarly based at the item response option level.

Descriptive analyses at the total score and factor level

The overall mean total CES-D score across gender and year levels is 12.08 (SD = 9.97). The mean total CES-D score is lower for boys (mean 10.80, SD = 8.76) compared with girls (mean = 13.67, SD = 11.10). shows mean CES-D total scores and factor scores by year level and gender. At each year level, girls on average show higher CES-D total scores than boys.

Higher total scores for girls are also reflected in higher mean values on each of the four factors which are presumed to comprise the CES-D. At each year level the variance in CES-D scores is greater for girls than it is for boys. Values of skewness are positive (indicating a predominance of low scores) but remained under two for both boys and girls across year levels.

An effect size for the gender difference in CES-D total scores is calculated for each year level and for all year levels combined. This is calculated in the same manner as the effect size calculated in the review of previous CES-D studies in adolescent samples (see xxx). In the present data the effect sizes are as follows: Year 8: 0.20; Year 9: 0.24; Year 10: 0.34; all year levels: 0.26. These effect sizes, using the Cohen (1977) convention, are classified as small.

In the final two row sections of **Table 10** the number of students classified as low or high scorers is shown using the CES-D22 variable (see Method chapter). Predictably the mean CES-D score of students classified as high scorers is substantially higher than students classified as low scorers. Of more interest is the fact that for low scoring students the mean gender difference is small (for the most part less than one CES-D point. For high scoring students the gender difference is larger with girls scoring two to three CES-D points higher than boys.

Table 10 CES-D total score and factor means by year level and gender

	Year 8		Year 9		Year 10		Total	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
N	1268	1038	1252	1023	1223	935	3743	2996
Total score								
Mean	11.47	13.75	10.73	13.35	10.18	13.94	10.80	13.67
SD	8.89	11.47	8.80	10.76	8.52	11.04	8.76	11.10
Median	9.00	10.00	8.00	11.00	8.00	11.00	8.00	11.00
Variance	78.96	131.64	77.60	115.68	72.65	121.98	76.68	123.15
Skewness	1.60	1.33	1.63	1.36	1.59	1.13	1.60	1.28
Depressed Affect								
Mean	2.29	3.58	2.01	3.48	1.88	3.70	2.06	3.58
SD	3.50	4.80	3.34	4.41	3.17	4.46	3.35	4.56
Somatic								
Mean	4.24	4.87	4.10	4.91	4.01	5.27	4.12	5.01
SD	3.30	4.01	3.41	3.94	3.43	4.05	3.38	4.00
Positive Affect								
Mean	3.91	4.15	3.76	3.98	3.56	4.04	3.75	4.06
SD	3.09	3.22	3.16	3.09	2.95	3.14	3.07	3.15
Interpersonal								
Mean	1.02	1.14	0.85	0.98	0.73	0.94	0.87	1.02
SD	1.47	1.51	1.28	1.39	1.22	1.35	1.33	1.42
Low scorers								
N	1121	829	1119	830	1087	735	3327	2394
Mean	8.93	8.85	8.36	9.18	7.81	9.12	8.37	9.05
SD	5.13	5.27	5.27	5.67	5.04	5.61	5.17	5.52
High scorers								
N	147	209	133	193	136	200	416	602
Mean	30.83	33.20	30.66	31.32	29.18	31.65	30.24	32.08
SD	7.52	8.36	7.21	8.76	6.65	7.44	7.17	8.23

Figure 1 shows a frequency distribution of CES-D total scores by gender for the dataset of all year levels combined. Because there are unequal numbers of boys and girls in the sample (there were more boys than girls) the frequency distribution is plotted as a percentage of cases to facilitate a comparison of the two series. The first striking feature of this figure is the commonly reported reverse J shape of the two distributions consistent with the positive skewness value in Table 10. Second, quite clearly there is a greater proportion of boys with CES-D scores less than 10 compared to girls and conversely there is a greater proportion of girls with scores above 15 compared with boys.

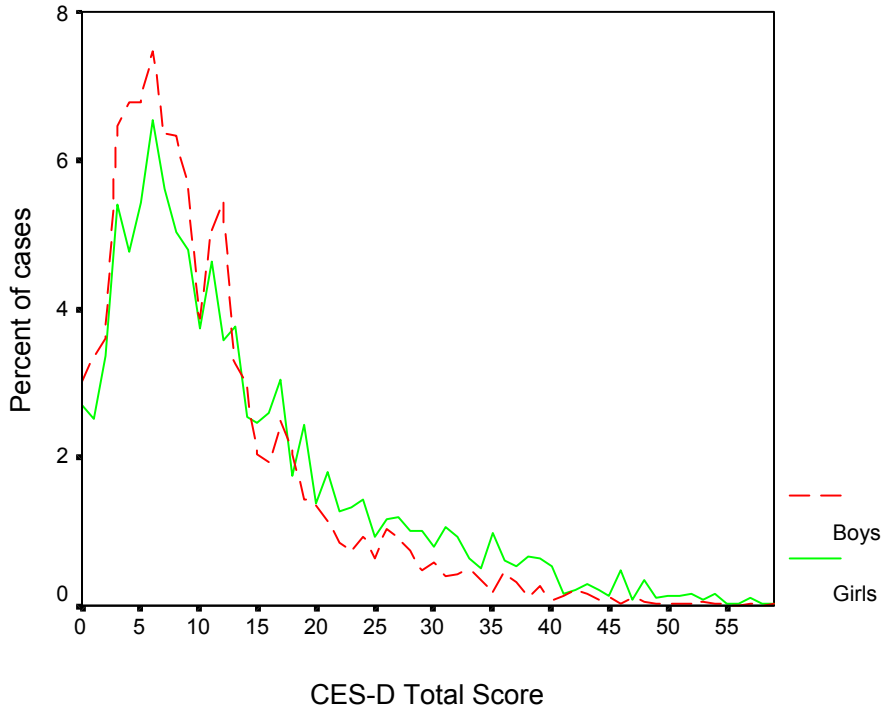


Figure 1 Frequency distribution of CES-D scores by gender

Table 11 shows the percentile rank values of CES-D total scores for boys and girls for the dataset of all year levels combined. At the 25th percentile the CES-D total score is five for boys and six for girls. The difference between the percentile values for boys and girls is shown in the fourth column and at the 25th percentile the difference between the boy and girl total CES-D score is one. In the final column of Table 11 the percentile values for the sample overall (both boys and girls combined) is shown.

For nearly the first half of the sample of boys and the first half of the sample of girls the difference in CES-D total scores is small (one CES-D point). At the 50th percentile and above this difference increases gradually so that by the 80th percentile the gender difference in CES-D total scores is six points and at the 95th percentile the difference is eight points.

Table 11 CES-D percentile scores by gender

Percentile	Boys Score	Girls Score	Boy – Girl Difference	Total
5	1	1	0	1
10	2	3	1	3
15	3	4	1	3
20	4	5	1	4
25	5	6	1	5
30	5	6	1	6
35	6	7	1	7
40	7	8	1	7
45	8	9	1	8
50	8	11	3	9
55	9	12	3	10
60	11	13	2	11
65	12	15	3	13
70	12	17	5	14
75	14	19	5	16
80	16	22	6	18
85	19	26	7	22
90	23	30	7	27
95	29	37	8	33

A series of standard *t* tests are performed to compare mean gender difference at the total score and factor level. At the total score level these mean gender differences are statistically significant at each year level (see Table 12). At the factor level inspection of the size of the *t* ratios suggests that the differences are most pronounced for the Depressed Affect and Somatic factors. In contrast, gender mean differences on the Positive Affect and Interpersonal factors are not statistically significant in some year levels and overall the *t* ratios are smaller in size.

The results from the *t* tests shown in Table 12 are calculated without taking into account the fact that the EDED data set is clustered and is not a simple random sample of students. To examine the potential bias on the estimates (particularly the standard errors) from ignoring this clustering, a series of HLM analyses are performed. These HLM analyses test gender mean differences with respect to CES-D total scores taking into account the school based clustering in the data.

Table 12 *T* tests of CES-D gender mean total score and factor differences

	Coefficient	Se	*t* ratio	p
Standard T tests				
Total score				
Year 8	2.28	0.42	5.38	<0.01
Year 9	2.62	0.41	6.40	<0.01
Year 10	3.76	0.42	8.92	<0.01
All years	2.87	0.24	11.87	<0.01
Depressed Affect				
Year 8	1.29	0.17	7.44	<0.01
Year 9	1.46	0.16	9.01	<0.01
Year 10	1.82	0.16	11.06	<0.01
All years	1.52	0.10	15.75	<0.01
Somatic				
Year 8	0.63	0.15	4.13	<0.01
Year 9	0.81	0.15	5.25	<0.01
Year 10	1.25	0.16	7.78	<0.01
All years	0.89	0.09	9.88	<0.01
Positive Affect				
Year 8	0.24	0.13	1.85	0.07
Year 9	0.22	0.13	1.71	0.09
Year 10	0.48	0.13	3.62	<0.01
All years	0.31	0.08	4.12	<0.01
Interpersonal				
Year 8	0.12	0.06	1.92	0.06
Year 9	0.13	0.06	2.23	0.03
Year 10	0.21	0.05	3.69	<0.01
All years	0.15	0.03	4.49	<0.01

The results are shown in Table 13 and reveal virtually identical estimates as those shown earlier. This suggests that the extent of the clustering in the EDED data set is minor and is unlikely to cause biased estimates if it is ignored in standard statistical analyses. Further details (consistent with this interpretation) regarding the extent of clustering in the EDED data set are provided in Chapter 9 which examines possible school effects on student CES-D levels.

Table 13 HLM *t* tests of CES-D gender mean total score differences

	Coefficient	Se	*t* ratio	p
Year 8	2.33	0.43	5.42	<0.01
Year 9	2.72	0.42	6.50	<0.01
Year 10	3.76	0.42	8.92	<0.01
All years	2.87	0.24	11.87	<0.01

Descriptive analyses at the item level

The mean and standard deviation of each CES-D item is calculated for boys and girls separately by year level. These are shown in Table 14 and Table 15. For most items the mean value is less than one. Of particular note are the consistently low mean values on Item 17 (*Cry*) for boys at all year levels and the relatively high mean values for boys and girls on the four positively worded items: *Good* (4), *Hopeful* (8), *Happy* (12) and *Enjoy* (16).

Boy and girl mean values are contrasted using two methods. In the first method the mean values for boys and girls are ranked (from highest to lowest) and these two rank orders compared. In the second method a mean ratio is calculated by dividing the girl item mean value by the boy mean value. The results from these analyses are shown in Table 16 and Table 17.

Table 16 shows CES-D mean item ranks by year level and gender. By way of example the Year 8 boy rank value for Item 8 (*Hopeful*) is one. This is because the mean value (1.26) for this item is the highest mean value of any item for boys in that year level. A Spearman Rank Correlation is calculated between the rank orders produced from the boys and girls data. The two rank orders are significantly (all at $p < 0.01$) related at each year level (Year 8: $r = 0.91$; Year 9: $r = 0.91$; Year 10: $r = 0.85$) and overall (all year levels: $r = 0.90$).

Table 17 shows gender differences in the rank order of each CES-D items and the gender mean ratio. Using Item 6 (*Depress*) by way of example the results are interpreted as follows. At Year 8 the gender mean rank difference for Item 6 (*Depress*) is two. This can be confirmed from the results presented in the earlier tables. Table 14 shows that the mean value for this item at Year 8 for boys is 0.51 and for girls it is 0.73. For boys the value for this item is their 10th highest and for girls it is their 8th highest.

Table 14 CES-D item means by year level and gender

	Year 8		Year 9		Year 10		Total	
CES-D Items	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
Depressed Affect								
3 Blues	.33	.54	.30	.57	.27	.62	.30	.57
6 Depress	.51	.73	.46	.73	.44	.76	.47	.74
9 Failure	.29	.39	.25	.35	.21	.36	.25	.37
10 Fearful	.26	.37	.24	.32	.22	.34	.24	.34
14 Lonely	.37	.54	.33	.53	.33	.53	.34	.53
17 Cry	.15	.38	.11	.40	.09	.40	.12	.39
18 Sad	.38	.64	.32	.60	.31	.67	.34	.64
Somatic								
1 Bothered	.31	.47	.29	.57	.37	.64	.33	.56
2 Appetite	.28	.59	.27	.56	.27	.66	.27	.60
5 Mind	.86	.87	.91	1.02	.92	1.11	.90	.99
7 Effort	1.20	.95	1.05	.83	.92	.80	1.06	.86
11 Sleep	.54	.78	.55	.75	.52	.79	.54	.77
13 Talk	.55	.62	.50	.60	.49	.62	.51	.61
20 Getgoing	.50	.60	.53	.59	.52	.66	.51	.61
Positive Affect								
4 Good	1.06	1.19	.96	1.13	.87	1.11	.96	1.14
8 Hopeful	1.26	1.27	1.16	1.22	1.16	1.23	1.19	1.24
12 Happy	.79	.80	.81	.76	.74	.81	.78	.79
16 Enjoy	.82	.90	.83	.87	.79	.89	.81	.89
Interpersonal								
15 Unfriendly	.54	.48	.44	.41	.37	.37	.45	.42
19 Dislike	.48	.67	.41	.57	.36	.57	.42	.60

Table 15 CES-D item standard deviations by year level and gender

	Year 8		Year 9		Year 10		Total	
CES-D Items	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
Depressed Affect								
3 Blues	.73	.91	.70	.90	.66	.92	.70	.91
6 Depress	.83	.99	.81	.94	.78	.93	.81	.95
9 Failure	.72	.84	.64	.75	.59	.76	.66	.79
10 Fearful	.63	.74	.59	.67	.58	.70	.60	.71
14 Lonely	.75	.92	.71	.85	.69	.83	.72	.87
17 Cry	.52	.77	.46	.73	.41	.74	.47	.75
18 Sad	.72	.92	.67	.83	.65	.85	.68	.87
Somatic								
1 Bothered	.68	.77	.64	.83	.76	.83	.70	.81
2 Appetite	.64	.90	.65	.87	.66	.89	.65	.89
5 Mind	.95	.97	.95	1.00	.93	.98	.94	.99
7 Effort	1.11	1.03	1.08	.94	1.03	.94	1.08	.98
11 Sleep	.87	1.00	.88	.96	.85	.96	.86	.97
13 Talk	.83	.89	.79	.83	.75	.80	.79	.84
20 Getgoing	.81	.87	.82	.82	.78	.85	.80	.85
Positive Affect								
4 Good	1.12	1.10	1.07	1.05	1.02	1.03	1.07	1.06
8 Hopeful	1.10	1.04	1.04	1.01	1.03	1.01	1.06	1.02
12 Happy	.95	.99	.94	.91	.87	.91	.92	.94
16 Enjoy	1.00	1.04	1.00	1.00	.94	.99	.98	1.01
Interpersonal								
15 Unfriendly	.83	.79	.72	.73	.70	.72	.76	.75
19 Dislike	.81	.93	.72	.84	.68	.83	.74	.87

Table 16 CES-D item mean ranks by year level and gender

	Year 8		Year 9		Year 10		Total	
CES-D items	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
Depressed Affect								
3 Blues	15	15	15	13	17	13	16	14
6 Depress	10	8	10	8	10	8	10	8
9 Failure	17	18	18	19	19	19	18	19
10 Fearful	19	20	19	20	18	20	19	20
14 Lonely	14	14	13	16	14	16	13	16
17 Cry	20	19	20	18	20	17	20	18
18 Sad	13	10	14	10	15	9	14	9
Somatic								
1 Bothered	16	17	16	12	11	12	15	15
2 Appetite	18	13	17	15	16	10	17	12
5 Mind	4	5	4	3	2	2	4	3
7 Effort	2	3	2	5	3	6	2	5
11 Sleep	8	7	7	7	7	7	7	7
13 Talk	7	11	9	9	9	14	9	11
20 Getgoing	11	12	8	11	8	11	8	10
Positive Affect								
4 Good	3	2	3	2	4	3	3	2
8 Hopeful	1	1	1	1	1	1	1	1
12 Happy	6	6	6	6	6	5	6	6
16 Enjoy	5	4	5	4	5	4	5	4
Interpersonal								
15 Unfriendly	9	16	11	17	12	18	11	17
19 Dislike	12	9	12	14	13	15	12	13

The difference in rank orders (taking the girl rank order from the boy rank order) is two. This indicates that Item 6 (*Depress*) is more salient to girls than to boys. Also shown is the mean item ratio calculated from dividing the girl mean value by the boy mean value. For Year 8, Item 6 (*Depress*) this is calculated as 1.43 (0.73/0.51). Mean ratios greater than one indicate that the mean value for girls is higher than for boys. The figures for all year levels combined are shown in Table 17.

Table 17 CES-D item rank differences and gender mean ratios

	Year 8		Year 9		Year 10		Total	
CES-D items	Difference	Ratios	Difference	Ratios	Difference	Ratios	Difference	Ratios
Depressed Affect								
3 Blues	0	1.62	+2	1.89	+4	2.26	+2	1.90
6 Depress	+2	1.43	+2	1.60	+2	1.72	+2	1.57
9 Failure	-1	1.33	-1	1.37	0	1.74	-1	1.45
10 Fearful	-1	1.42	-1	1.32	-2	1.55	-1	1.42
14 Lonely	0	1.47	-3	1.59	-2	1.61	-3	1.55
17 Cry	+1	2.55	+2	3.47	+3	4.41	+2	**3.31**
18 Sad	+3	1.67	+4	1.88	+6	2.18	**+5**	1.88
Somatic								
1 Bothered	-1	1.51	+4	1.95	-1	1.74	0	1.72
2 Appetite	+5	2.12	+2	2.12	+6	2.41	**+5**	**2.21**
5 Mind	-1	1.00	+1	1.12	0	1.20	+1	1.11
7 Effort	-1	0.79	-3	0.78	-3	0.87	-3	**0.81**
11 Sleep	+1	1.44	0	1.35	0	1.51	0	1.43
13 Talk	-4	1.14	0	1.19	-5	1.25	-2	1.19
20 Getgoing	-1	1.19	-3	1.12	-3	1.27	-2	1.19
Positive Affect								
4 Good	+1	1.13	+1	1.17	+1	1.27	+1	1.19
8 Hopeful	0	1.01	0	1.05	0	1.06	0	1.04
12 Happy	0	1.02	0	0.94	+1	1.09	0	1.01
16 Enjoy	+1	1.10	+1	1.06	+1	1.13	+1	1.09
Interpersonal								
15 Unfriendly	-7	0.89	-6	0.94	-6	1.00	**-6**	**0.94**
19 Dislike	+3	1.37	-2	1.37	-2	1.57	-1	1.43

Table 17 shows the largest gender difference (shown in bold) in rank orders are for: Item 18 (*Sad*), Item 2 (*Appetite*) and Item 15 (*Unfriendly*). Two of these items favoured girls: *Sad* (18) and *Appetite* (2) with much higher rankings compared with boys but *Unfriendly* (15) favoured boys. Other than for these three items the remaining item rank differences are less than five.

Mean gender ratios for nearly all items are above one indicating that item mean values were higher for girls than for boys. Mean gender ratio's greater than two (shown in bold) are reported for Item 2 (*Appetite*) and Item 17 *(Cry*). Mean gender ratios less than one (indicating higher boy mean values) are found for Item 7 (*Effort*) and Item 15 (*Unfriendly*) across all three year levels.

Item to total score correlations and the coefficient alphas are shown in Table 18 for each gender and year level separately. For each gender and year level the coefficient alpha is high (above 0.80) indicating good internal consistency for the scale. Generally item to total score correlations are above 0.40 except for the items: *Effort* (7), *Hopeful* (8) and *Appetite* (2) (Boys only). Using the data from all year levels combined it can be seen, with exception of Item 15 (*Unfriendly*), that item-scale correlations are higher for girls than boys.

Using a data set comprising all year levels combined item mean values are calculated for low and high (į 22 CES-D points) scoring students. These mean values for boys and girls are shown in Table 19. Also shown is the ratio between the mean value of high scorers to low scorers. For example, Item 3 (*Blues*) for boys the mean value for low scorers is 0.16 and for high scorers it is 1.44. This produced a ratio of 9.00 (1.44 / 0.16). For boys and girls it is clear that the mean value for every item is higher for high scoring students compared with low scoring students.

In terms of discriminating between high and low scoring students, the Depressed Affect items record the largest mean ratios and the Positive Affect items the lowest. Of note, for both boys and girls, are very high mean ratios for Item 9 (*Failure*) and Item 17 (*Cry*) and low ratios for Item 7 (*Effort*) and Item 8 (*Hopeful*).

Table 18 Item to total score correlations by year level and gender

	Year 8		Year 9		Year 10		Total	
CES-D items	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
Depressed Affect								
3 Blues	.56	.66	.60	.69	.61	.71	.59	.68
6 Depress	.69	.80	.69	.77	.70	.78	.69	.78
9 Failure	.59	.68	.61	.68	.60	.66	.60	.67
10 Fearful	.44	.59	.52	.57	.51	.55	.49	.57
14 Lonely	.66	.72	.65	.66	.65	.70	.66	.70
17 Cry	.44	.62	.40	.59	.36	.56	.41	.59
18 Sad	.67	.78	.70	.75	.66	.74	.68	.76
Somatic								
1 Bothered	.48	.52	.47	.58	.56	.55	.50	.55
2 Appetite	.33	.47	.38	.46	.39	.44	.36	.46
5 Mind	.47	.56	.47	.54	.47	.57	.47	.55
7 Effort	.02	.21	.10	.25	.18	.37	.10	.27
11 Sleep	.37	.49	.47	.48	.43	.49	.42	.49
13 Talk	.46	.52	.46	.52	.47	.51	.47	.52
20 Getgoing	.52	.63	.52	.58	.56	.66	.53	.62
Positive Affect								
4 Good	.34	.50	.42	.49	.38	.54	.38	.51
8 Hopeful	.26	.32	.30	.36	.27	.41	.28	.36
12 Happy	.55	.68	.57	.65	.54	.66	.55	.66
16 Enjoy	.53	.64	.58	.64	.56	.66	.55	.65
Interpersonal								
15 Unfriendly	.53	.44	.46	.44	.48	.42	.49	.43
19 Dislike	.61	.65	.58	.60	.58	.65	.59	.63
Alpha	.86	.91	.87	.91	.88	.91	.87	.92

Table 19 CES-D item means by gender and low versus high scorers

	Boys			Girls		
CES-D items	Low	High	Ratio	Low	High	Ratio
Depressed Affect						
3 Blues	.16	1.44	9.00	.28	1.75	6.25
6 Depress	.29	1.92	6.62	.40	2.10	5.25
9 Failure	.12	1.34	11.17	.12	1.33	11.08
10 Fearful	.14	1.04	7.43	.16	1.09	6.81
14 Lonely	.19	1.61	8.47	.25	1.66	6.64
17 Cry	.04	.68	17.00	.20	1.18	5.90
18 Sad	.19	1.55	8.16	.34	1.82	5.35
Somatic						
1 Bothered	.21	1.21	5.76	.36	1.35	3.75
2 Appetite	.19	.92	4.84	.42	1.34	3.19
5 Mind	.77	1.91	2.48	.75	1.96	2.61
7 Effort	1.01	1.49	1.48	.73	1.35	1.85
11 Sleep	.42	1.49	3.55	.56	1.60	2.86
13 Talk	.40	1.41	3.53	.42	1.36	3.24
20 Getgoing	.38	1.58	4.16	.37	1.56	4.22
Positive Affect						
4 Good	.85	1.87	2.20	.90	2.11	2.34
8 Hopeful	1.11	1.86	1.68	1.08	1.88	1.74
12 Happy	.64	1.94	3.03	.50	1.94	3.88
16 Enjoy	.66	2.03	3.08	.59	2.08	3.53
Interpersonal						
15 Unfriendly	.33	1.41	4.27	.27	1.02	3.78
19 Dislike	.28	1.52	5.43	.35	1.59	4.54

Descriptive analyses at the response option level

In the previous section, among other things, item mean values were compared between boys and girls. These item means are weighted averages affected by two components: the frequency of the symptom and the duration of the symptom among those who experienced it (Wells et al., 1987). In order to examine more closely the gender differences in item mean values a method outlined by Santor and Coyne (1997) is used to calculate the probability of endorsing CES-D item response options for boys and girls separately for each year level and for all year levels combined.

The pattern of results is very similar across year levels and in this chapter only the results from Year 8 are presented. These results are shown in Table 20 with the results for the remaining year levels provided in Appendix B.

Table 20 Proportion of boys and girls endorsing response options (Year 8)

	CES-D Item	Response Option	% Boys	% Girls	Cumulative % Boys	Cumulative % Girls	Difference Total
1	Bothered	0	78.1	66.8			
		1	15.4	22.2	21.9	33.3	**11.4**
		2	3.5	8.0	6.5	11.1	4.6
		3	3.0	3.1	3.0	3.1	0.1
2	Appetite	0	80.5	63.2			
		1	13.2	21.3	19.5	36.9	**17.4**
		2	4.3	9.0	6.3	15.6	9.3
		3	2.0	6.6	2.0	6.6	4.6
3	Blues	0	78.6	68.7			
		1	13.4	15.6	21.4	31.4	**10.0**
		2	4.2	9.2	8.0	15.8	7.8
		3	3.8	6.6	3.8	6.6	2.8
4	Good	0	42.8	35.3			
		1	25.7	28.0	57.2	64.7	7.5
		2	14.6	19.1	31.5	36.7	5.2
		3	16.9	17.6	16.9	17.6	0.7
5	Mind	0	45.0	45.7			
		1	32.1	31.1	55.0	54.4	-0.6
		2	14.5	14.3	22.9	23.3	0.4
		3	8.4	9.0	8.4	9.0	0.6
6	Depress	0	65.9	56.8			
		1	22.0	22.5	34.1	43.2	9.1
		2	7.2	11.6	12.1	20.7	8.6
		3	4.9	9.1	4.9	9.1	4.2
7	Effort	0	36.4	45.3			
		1	23.7	25.5	63.5	54.7	-8.8
		2	23.2	18.4	39.8	29.2	**-10.6**
		3	16.6	10.8	16.6	10.8	-5.8
8	Hopeful	0	32.4	28.3			
		1	28.5	32.9	67.6	71.7	4.1
		2	20.3	22.6	39.1	38.8	-0.3
		3	18.8	16.2	18.8	16.2	-2.6
9	Failure	0	82.6	78.4			
		1	9.0	10.1	17.3	21.6	4.3
		2	4.8	5.7	8.3	11.5	3.2
		3	3.5	5.8	3.5	5.8	2.3

Table 20 Proportion of boys and girls endorsing response options (Year 8) (continued)

CES-D Item	Response Option	% Boys	% Girls	Cumulative % Boys	Cumulative % Girls	Difference Total
10 Fearful	0	82.6	75.9			
	1	11.1	15.1	17.4	24.1	6.7
	2	4.3	5.4	6.3	9.0	2.7
	3	2.0	3.6	2.0	3.6	1.6
11 Sleep	0	65.7	53.5			
	1	20.1	25.0	34.3	46.5	**12.2**
	2	8.6	11.7	14.2	21.5	7.3
	3	5.6	9.8	5.6	9.8	4.2
12 Happy	0	49.5	52.3			
	1	31.0	24.2	50.5	47.7	-2.8
	2	10.7	14.5	19.5	23.5	4.0
	3	8.8	9.0	8.8	9.0	0.2
13 Talk	0	63.5	60.3			
	1	22.6	22.6	36.6	39.7	3.1
	2	9.9	11.7	14.0	17.1	3.1
	3	4.1	5.4	4.1	5.4	1.3
14 Lonely	0	76.1	68.6			
	1	14.9	16.1	23.9	31.4	7.5
	2	5.1	8.1	9.0	15.3	6.3
	3	3.9	7.2	3.9	7.2	3.3
15 Unfriendly	0	63.6	67.4			
	1	23.3	21.0	36.4	32.6	-3.8
	2	8.6	7.9	13.1	11.6	-1.5
	3	4.5	3.7	4.5	3.7	-0.8
16 Enjoy	0	50.6	47.7			
	1	27.2	26.5	49.4	52.4	3.0
	2	12.2	14.1	22.2	25.9	3.7
	3	10.0	11.8	10.0	11.8	1.8
17 Cry	0	90.5	75.2			
	1	5.8	15.3	9.6	24.7	**15.1**
	2	2.1	5.4	3.8	9.4	5.6
	3	1.7	4.0	1.7	4.0	2.3
18 Sad	0	72.8	59.2			
	1	19.0	25.0	27.2	40.8	**13.6**
	2	5.2	8.4	8.2	15.8	7.6
	3	3.0	7.4	3.0	7.4	4.4
19 Dislike	0	67.7	57.8			
	1	20.5	25.6	32.4	42.2	**9.8**
	2	7.6	8.8	11.9	16.6	4.7
	3	4.3	7.8	4.3	7.8	3.5
20 Get-going	0	65.8	60.8			
	1	22.6	24.5	34.2	39.3	5.1
	2	7.4	9.1	11.6	14.8	3.2
	3	4.2	5.7	4.2	5.7	1.5

Cumulative difference totals ≥10 shown in bold

Taking Item 1 (*Bothered*) by way of example the results shown in Table 20 can be interpreted as follows. The majority of boys (78.1%) and girls (66.8%) indicate that for the last two weeks they had been bothered by things that don't usually bother them 'rarely or none of the time' (Option 0). Around one fifth of the sample endorse Option 1 (some or a little of the time) followed by a smaller group endorsing Option 2 (occasionally or a moderate amount of the time) or Option 3 (most or all of the

time). Note that a much higher proportion of girls endorse Option 1 (22.2%) and Option 2 (8.0%) compared with boys (Option 1: 15.4%; Option 2: 3.5%). Similar proportions of boys and girls endorse Option 3.

Cumulative probabilities are reported in the next two columns. For Item 1 (*Bothered*) the cumulative probability of endorsing either Options 1, 2 or 3 (i.e. reporting the presence of a symptom) is larger for girls (33.3%) than boys (21.9%). Similarly the cumulative probability of endorsing either Options 2 or 3 is larger for girls (11.1%) than boys (6.5%). Finally the probability of endorsing Option 3 is nearly identical for boys (3.0%) and girls (3.1%). The difference between these cumulative probabilities are reported in the last column of Table 20. For Item 1 (*Bothered*) the largest difference between boys and girls is observed for the cumulative probability to Option 1 (11.4). A large difference indicates that the item discriminates well between boys and girls. Differences equal to or greater than 10 are shown in bold face.

Table 20 shows that for the majority of items the most frequent response option chosen for both boys and girls is zero. For example the majority of boys (91%) and girls (75%) endorse Option 0 for Item 17 (*Cry*). A small number of items such as Item 8 (*Hopeful*) are endorsed fairly evenly across Options 0, 1 and 2. This item (e.g. 'I felt hopeful about the future') is one of four positively worded items that has been reversed scored. This means that subjects have been scored as if they were answering the question 'I have **not** felt hopeful about the future'. To reflect this, subjects endorsing Option 4 are coded as Option 0, those endorsing Option 3 are scored as Option 1 and so on. Interestingly a lack of positive affect (Options 1, 2 or 3) is also evident for the other positively worded items: *Good* (4), *Happy* (12) and *Enjoy* (16). The difference between boys and girls on these items, however, is not large.

From Table 20 it can be seen that the items showing the greatest difference between boys and girls are: *Bothered* (1), *Appetite* (2), *Blues* (3), *Effort* (7), *Sleep* (11), *Cry* (17), *Sad* (18) and *Dislike* (19). With the exception of Item 7 (*Effort*) the differences are positive indicating that the probability of endorsing Options 1, 2 or 3 are higher for girls than for boys. Consistent with the fact that the mean value for Item 7 (*Effort*) is higher for boys than for girls (see Table 14) a greater proportion of boys endorsed Option 2 or 3 on this item than did girls.

For the remainder of items the probability of endorsing either Option 1, 2 or 3 (in effect signalling the presence of a symptom) is greater for girls than it is for boys. The items identified in this manner are largely consistent with the earlier analyses of item means. That is, the items shown to be effective in discriminating between boys and girls on the basis of differences at the response option level, viz: *Bothered* (1), *Appetite* (2), *Blues* (3), *Effort* (7), *Sleep* (11), *Cry* (17), *Sad* (18) and *Dislike* (19) are the same items showing relatively large gender differences with respect to their mean values. In

Table **21** the group difference totals in the proportion of boys and girls endorsing response options by each year level is shown. For Year 8 these figures are identical to those presented in the final column of the previous table (Table 20). Differences equal to or greater than 10 are shown in bold face. A good deal of consistency is shown across the three year levels with items that best discriminate between boys and girls at Year 8 also showing large differences at Years 9 and 10.

Table 21 Group differences between the proportion of boys and girls endorsing response options by year level

	CES-D Item	Response Option	Difference Year 8	Difference Year 9	Difference Year 10	Difference All year levels
1	Bothered	1	**11.4**	**17.9**	**19.8**	**16.1**
		2	4.6	7.8	6.4	6.2
		3	0.1	2.1	1.1	1.1
2	Appetite	1	**17.4**	**18.4**	**25.0**	**20.0**
		2	9.3	8.8	**10.9**	**9.5**
		3	4.6	2.7	3.0	3.4
3	Blues	1	**10.0**	**15.9**	**19.6**	**15.1**
		2	7.8	7.3	**10.3**	8.4
		3	2.8	3.6	4.4	3.6
4	Good	1	7.5	**9.8**	**12.4**	**9.8**
		2	5.2	6.7	9.1	7.0
		3	0.7	0.2	2.0	1.0
5	Mind	1	-0.6	3.4	8.0	3.3
		2	0.4	4.8	6.9	3.9
		3	0.6	3.0	3.7	2.4
6	Depress	1	9.1	**15.7**	**19.5**	**14.7**
		2	8.6	8.6	9.2	8.8
		3	4.2	2.7	3.4	3.4
7	Effort	1	-8.8	-4.7	-2.4	-5.3
		2	**-10.6**	**-12.5**	-6.4	**-9.8**
		3	-5.8	-5.8	-3.3	-5.0
8	Hopeful	1	4.1	4.2	3.1	3.8
		2	-0.3	3.1	4.9	2.6
		3	-2.6	-1.1	-0.9	-1.5
9	Failure	1	4.3	5.4	9.4	6.3
		2	3.2	2.1	3.9	3.1
		3	2.3	1.6	2.3	2.1
10	Fearful	1	6.7	5.1	7.9	6.6
		2	2.7	1.6	2.6	2.3
		3	1.6	1.0	1.4	1.4
11	Sleep	1	**12.2**	**11.5**	**16.1**	**13.2**
		2	7.3	5.1	6.8	6.4
		3	4.2	2.6	3.9	3.6
12	Happy	1	-2.8	-2.1	1.7	-1.2
		2	4.0	-0.7	5.1	2.8
		3	0.2	-1.9	0.2	-0.5
13	Talk	1	3.1	6.5	**9.6**	6.3
		2	3.1	2.3	1.6	2.4
		3	1.3	0.5	1.1	1.0
14	Lonely	1	7.5	**11.9**	**13.1**	**10.8**
		2	6.3	5.7	5.3	5.8
		3	3.3	1.9	2.1	2.5
15	Unfriendly	1	-3.8	-3.0	-1.1	-2.7
		2	-1.5	0.1	1.0	-0.3
		3	-0.8	0.2	0.2	-0.2
16	Enjoy	1	3.0	2.6	4.2	3.1
		2	3.7	2.3	4.1	3.3
		3	1.8	-0.1	1.7	1.1
17	Cry	1	**15.1**	**19.8**	**22.3**	**19.1**
		2	5.6	6.9	6.9	6.5
		3	2.3	1.2	2.0	1.9
18	Sad	1	**13.6**	**19.0**	**24.4**	**18.7**
		2	7.6	7.3	**9.5**	8.0
		3	4.4	1.6	2.7	2.9
19	Dislike	1	**9.8**	7.9	**12.1**	**9.9**
		2	4.7	4.6	5.8	5.0
		3	3.5	2.6	2.3	2.8
20	Get-going	1	5.1	5.3	8.4	6.2
		2	3.2	1.4	4.8	3.1
		3	1.5	-0.3	0.9	0.7

Cumulative difference totals ≥10 shown in bold

Table 22 Group differences between the proportion of high and low scorers endorsing response options by gender (all year levels)

CES-D Item	Response Option	Boys Difference	Girls Difference	Difference Girls – Boys
1 Bothered	1	47.6	45.0	-2.6
	2	35.1	38.5	3.4
	3	16.9	16.2	-0.7
2 Appetite	1	38.9	39.6	0.7
	2	23.6	34.4	**10.8**
	3	10.2	17.8	7.6
3 Blues	1	63.0	65.0	2.0
	2	44.7	54.5	**9.8**
	3	19.9	27.3	7.4
4 Good	1	36.4	35.2	-1.2
	2	44.1	50.5	6.4
	3	22.2	35.5	**13.3**
5 Mind	1	38.0	38.0	0
	2	48.6	52.4	3.8
	3	27.6	30.4	2.8
6 Depress	1	68.5	63.9	-4.6
	2	63.0	69.8	6.8
	3	31.7	36.2	4.5
7 Effort	1	26.0	33.7	7.7
	2	17.4	19.7	2.3
	3	5.3	8.4	3.1
8 Hopeful	1	20.7	20.8	0.1
	2	35.3	37.3	2.0
	3	19.8	22.0	2.2
9 Failure	1	63.9	61.4	-2.5
	2	40.7	38.9	-1.8
	3	18.3	21.1	2.8
10 Fearful	1	51.1	51.0	-0.1
	2	30.5	30.3	-0.2
	3	8.7	11.9	3.2
11 Sleep	1	45.0	38.3	-6.7
	2	39.4	41.1	1.7
	3	23.2	25.2	2.0
12 Happy	1	45.1	55.8	**10.7**
	2	59.8	63.6	3.8
	3	25.7	24.5	-1.2
13 Talk	1	48.3	44.4	-3.9
	2	38.0	33.3	-4.7
	3	14.7	15.4	0.7
14 Lonely	1	68.8	64.0	-4.8
	2	51.7	53.2	1.5
	3	21.9	24.1	2.2
15 Unfriendly	1	49.2	39.3	**-9.9**
	2	39.8	25.9	**-13.9**
	3	18.8	9.2	**-9.6**
16 Enjoy	1	46.4	51.3	4.9
	2	59.9	63.8	3.9
	3	30.8	33.9	3.1
17 Cry	1	35.0	50.4	**15.4**
	2	20.0	33.1	**13.1**
	3	8.5	14.8	6.3
18 Sad	1	69.3	62.9	-6.4
	2	48.0	60.6	**12.6**
	3	19.7	24.7	5.0
19 Dislike	1	57.4	52.3	-5.1
	2	44.9	47.2	2.3
	3	21.6	25.1	3.5
20 Get-going	1	55.3	55.3	0
	2	43.1	44.2	1.1
	3	21.1	19.5	-1.6

Cumulative difference totals ≥10 shown in bold

The most effective items for discriminating between boys and girls include the items found at Year 8, namely: *Bothered* (1), *Appetite* (2), *Blues* (3), *Effort* (7), *Sleep* (11), *Cry* (17), *Sad* (18) and *Dislike* (19), as well as the items: *Good* (4), *Depress* (6) and *Lonely* (14). The direction of this difference is greater than zero (positive). Generally the largest difference in cumulative probabilities between boys and girls occurs for Option 1. This indicates that the probability of endorsing either Options 1, 2 or 3 (in effect signalling the presence of these symptoms) for these items is greater (because the difference is greater than zero) for girls and greater than the difference between boys and girls with respect to the probability of endorsing either Options 2 or 3 combined or Option 3 alone.

Table 22 shows the group difference totals in the proportion of low and high scorers endorsing response options. These figures are calculated for boys and girls separately using a data set comprising all year levels. The full tables showing the actual response proportions are contained in Appendix B. Consistent with the earlier results that examine item mean values,

Table 22 shows that for both boys and girls all items are to some degree effective in discriminating between high and low scorers. In the third and final column the difference value for boys is subtracted from the difference value for girls. Values equal to or greater than 10 are shown in bold face.

From Table 22 it can be seen that there is considerable consistency between boys and girls in terms of the items that best discriminate between low and high scorers. For example, Item 9 (*Failure*) shows a large cumulative difference for Option 1. For boys this value is 63.9 and for girls it is 61.4. This indicates that the probability endorsing either Options 1, 2 or 3 for this item is larger for high scoring students (Boys & Girls) than low scoring students. The items *Effort* (7), *Hopeful* (8) and *Cry* (17) (for boys only) appear quite poor at discriminating between low and high scoring students.

Although a good degree of consistency is evident across gender, some differences for the items: *Appetite* (2), *Blues* (3), *Good* (4), *Happy* (12), *Unfriendly* (15), *Cry* (17) and *Sad* (18) are apparent. Generally the largest difference between boys and girls occurred with respect to Options 1 and 2. Except for Item 15 (*Unfriendly*) these differences favour girls in the sense that the items showed better discrimination (larger differences in cumulative probabilities) for girls compared with boys.

Year level descriptive analyses

In this section, CES-D mean total scores and item mean values for boys and girls are examined across year levels. The results are presented in Table 23. Total scores for boys and girls are fairly stable across year levels but overall show a slight decrease over time. Boys show a slight decrease from 11.47 in Year 8 down to 10.18 in Year 10, while total scores for girls decrease between Year 8 to Year 9 (13.75 to 13.35) but then increases between Year 9 and Year 10 (13.35 to 13.94).

Overall, item means are fairly stable across year levels for both boys and girls but some item differences are evident. These are highlighted in bold in Table 23 using a difference of greater than 0.20 between any two year levels as a rough guide to indicate possible DIF. Using this rule of thumb two items are identified as showing possible DIF. Mean values for Item 5 (*Mind*), particularly for girls, increase across year levels while mean values for Item 7 (*Effort*), particularly for boys, decrease across year levels.

Table 23 CES-D item means by gender and year level

		Boys			Girls			Boys + Girls		
	CES-D Items	Year 8	Year 9	Year 10	Year 8	Year 9	Year 10	Year 8	Year 9	Year 10
1	Bothered	.31	.29	.37	.47	.57	.64	.39	.42	.49
2	Appetite	.28	.27	.27	.59	.56	.66	.42	.40	.44
3	Blues	.33	.30	.27	.54	.57	.62	.42	.42	.42
4	Good	1.06	.96	.87	1.19	1.13	1.11	1.12	1.03	.97
5	Mind	.86	.91	.92	**.87**	**1.02**	**1.11**	.86	.96	1.00
6	Depress	.51	.46	.44	.73	.73	.76	.61	.58	.58
7	Effort	**1.20**	**1.05**	**.92**	.95	.83	.80	**1.09**	**.95**	**.87**
8	Hopeful	1.26	1.16	1.16	1.27	1.22	1.23	1.26	1.19	1.19
9	Failure	.29	.25	.21	.39	.35	.36	.34	.29	.28
10	Fearful	.26	.24	.22	.37	.32	.34	.31	.28	.27
11	Sleep	.54	.55	.52	.78	.75	.79	.65	.64	.64
12	Happy	.79	.81	.74	.80	.76	.81	.79	.79	.77
13	Talk	.55	.50	.49	.62	.60	.62	.58	.55	.55
14	Lonely	.37	.33	.33	.54	.53	.53	.44	.42	.42
15	Unfriendly	.54	.44	.37	.48	.41	.37	.51	.43	.37
16	Enjoy	.82	.83	.79	.90	.87	.89	.85	.85	.83
17	Cry	.15	.11	.09	.38	.40	.40	.25	.24	.23
18	Sad	.38	.32	.31	.64	.60	.67	.50	.44	.47
19	Dislike	.48	.41	.36	.67	.57	.57	.57	.48	.45
20	Get-going	.50	.53	.52	.60	.59	.66	.54	.55	.58
	Total	11.47	10.73	10.18	13.75	13.35	13.94	12.50	11.91	11.81

Summary

In this chapter the results from a number of simple descriptive analyses have been presented. A variety of analyses were performed to reflect the fact that CES-D gender differences are possible at a number of different levels: namely, total score, factor, item and response option. The key findings presented in this chapter can be summarised as follows:

At each year level girls on average showed higher CES-D total scores than boys. This difference although statistically significant was not large in magnitude. Generally higher girl item mean values were most notable for items comprising the Depressed Affect and Somatic factors.

The mean value for nearly every CES-D item was higher for girls than for boys. The exceptions to this were higher boy mean values for Item 7 (*Effort*)

and Item 15 (*Unfriendly*). The mean values on Item 17 (*Cry*) and Item 2 (*Appetite*) for girls were more than double the corresponding mean values for boys. There was a large rank order difference in means for Item 18 (*Sad*) with this item more salient to girls than to boys.

The most frequent response option chosen for both boys and girls for all items (with one minor exception for Item 8 (Hopeful)) was a zero. The most effective items in terms of discriminating between boys or girls were: *Bothered* (1), *Appetite* (2), *Blues* (3), *Good* (4), *Depress* (6), *Effort* (7), *Sleep* (11), *Lonely* (14), *Cry* (17), *Sad* (18) and *Dislike* (19). With the exception of Item 7 (*Effort*) these items were effective because a greater proportion of girls reported the presence (as opposed to the intensity) of these symptoms.

Item 7 (*Effort*) appeared to be effective in discriminating between boys or girls because a greater proportion of boys compared with girls endorsed either Option 2 or 3.

Across year levels mean total CES-D scores decreased slightly. Item means remained relatively stable but the means for Item 5 (*Mind*) increased over time and the mean for Item 7 (*Effort*) decreased over time.

The analyses presented in this chapter were simple to calculate and the results easy to interpret. As such they provide useful information to help guide the more complex statistical analyses, based around IRT and SEM techniques, which follow in the next three chapters.

6

IRT Analyses of Gender and Year Level Effects

This chapter presents the results from a series of item response theory (IRT) analyses of the CES-D. Initially, and primarily to assist readers not familiar with IRT, a brief non-technical background to IRT models and their use in detecting differential item functioning (DIF) is provided. This is followed by a more detailed description of the specific statistical IRT technique and software used in the present study.

The main section of this chapter comprises a graphical IRT analysis of the CES-D. Option characteristic curves (OCC) and item characteristic curves (ICC) plots for each CES-D item separately for boys and girls are presented. These plots provide the foundation for quantifying the magnitude of gender DIF in the CES-D and for investigating the source of the DIF at the response option level. The impact of gender DIF at the total score level and the possibility of DIF across year levels (from Year 8, 9 & 10) is also examined. In the final section, IRT psychometric data for the CES-D at the scale level are presented.

Background to IRT models

IRT provides a class of models for describing the relationship between item responses and the construct being measured by the test (Thissen, Steinberg & Wainer, 1993). In the present study IRT models are used to describe the relationship between responses to CES-D items and the unobservable (or latent) variable which is labelled 'depressive symptomatology'. Many different kinds of IRT models have been developed ranging from simple one parameter Rasch models (Rasch, 1960) to complex multi-dimensional three parameter models (Reckase, 1997).

A basic concept in all IRT models is that respondents and items can be located along a common underlying dimension or latent trait. Because IRT models have mainly been applied to educational testing this underlying dimension is usually called 'ability' but it can be any construct that is being measured by a test. This construct is

traditionally symbolised using the Greek small letter θ (theta). The relationship between item responses and the construct is defined by the item response function – most commonly a S-shaped trace line (e.g. Figure 2(C)) of the proportion of individuals at the same level of θ who correctly endorse an item.

The IRT trace line, termed an item characteristic curve (ICC), is defined by a small number of parameters. One of the most commonly used IRT models is the two parameter logistic model (2PL) for dichotomous data. The formula (Hays, 1998, p. 184) for the 2PL model is:

$$Pr(x = 1 : \theta = \theta) = \left| 1 / \{1 + EXP[-\alpha(\theta - \beta)]\} \right|$$

Where,

$Pr(x = 1 : \theta = \theta)$ = the conditional probability of item endorsement

θ = the examinee trait level parameter

α = the item discrimination parameter

β = the item difficulty parameter

This formula indicates that the probability that a person with a latent trait θ correctly endorses an item is a function of two parameters: α the discrimination parameter which represents the steepness of the ICC and β the difficulty parameter which represents the point on the latent trait continuum where the respondent has a 0.50 probability of endorsing the item.

Other popular IRT models are the one parameter logistic (1PL) model and the three parameter (3PL) model. In the 1PL model the discrimination parameters are held constant across items, which in a factor analytic sense is analogous to assuming that test items have equivalent factor loadings. The 3PL model includes a pseudo-guessing parameter (called *c*) to take into account that in multiple choice tests, respondents have a non-zero probability (because of chance) of correctly answering questions.

The most appropriate logistic IRT model for the investigation of DIF is a complex topic. It does seem clear that despite the theoretical parameter invariance achieved by the 1PL model and the ready availability of computer software (Scheuneman and Bleistein, 1997), the 1PL model may be less than ideal in many applications. This is because if there are group differences in the α parameters (which are not modelled in the 1PL model) then these differences can result in misleading DIF values. As Angoff (1993, p. 9) explains:

> Because it [the Rasch 1PL] assumes that there are no differences in the α or *c* parameters, any such real differences are therefore not detectable by means of the Rasch model. Even more serious is the fact that any real differences in the α or *c* prameters are likely to result in artifactual DIF values.

A similar conclusion was reached by Osterlind (1983, p.59) who argued that while there is considerable debate about the relative advantages of logistic IRT models for the detection of DIF the "… three-parameter logistic model appears to receive the most favourable attention for bias item work. This may be because it comes closest to describing psychometrically multiple choice tests as they are presently constructed and used".

ICCs are ideally suited to investigating DIF (Lord, 1980). The basic idea for this is very simple. At each level of θ the value of the ICC is the probability of item endorsement given that level of ability. If the ICCs from two groups are very similar then this indicates that the item is functioning in a similar fashion in both groups and that there is little or no DIF. On the other hand if the ICCs for the two groups differ then this indicates that there is DIF.

Examples of ICCs for two groups (Boys & Girls) are shown in Figure 2(D) and (E). A more detailed explanation of these graphs is provided later but for now it is sufficient to see that in graph (D) the ICCs are very similar (indicating no DIF) while in graph (E) the two ICCs are quite divergent with girls showing higher item scores than boys indicating possible DIF.

Observant readers will have noticed that the Y axis in Figure 2(D) and (E) is labelled 'Item Score' with a scale from zero to three rather as might be expected an axis showing the probability between zero and one for correct item endorsement. This is because the graphs show ICCs for CES-D items estimated from an IRT package specially developed for rating scale analysis. Early researchers confronted by rating scales comprising three or more response options were forced to collapse response categories and then examine changes between pairs of response options separately. In terms of analysing data which did not match the original response scale and the loss of information inherent in collapsing categories this approach was less than ideal.

A number of polytomous IRT models have been specially developed for the analysis of test items with more than two categories or response options. These include models for nominal data where the ordering of response options within items is not known *a priori* (Block, 1972), the partial credit model (Masters, 1982) which is a polytomous generalisation of the 1PL model, the graded response model (Samejima, 1969) which is a generalisation of the 2PL model, and the rating scale model (Andrich, 1978). Separate chapters describing these, and other IRT models, can be found in van der Linden and Hambleton (1996). The IRT models described to this point are based on the parametric logistic response function. Although this approach is very popular it does have its limitations.

Parametric IRT models assume that the relationship between responses to items and their underlying construct can be described by a small number of parameters, most typically these are α (the discrimination parameter) and β (the difficulty parameter). A key advantage of relying on just a few parameters is that the calculated estimates will be stable. However, when items have not been constructed with the logistic model in mind (as would apply to most psychological scales) then these items may not be modelled efficiently and the resulting parameter estimates may be misleading (Santor & Ramsay, 1998). In fact, items which are not modelled efficiently with the logistic function might still be useful items either across a very narrow range of the underlying trait or in special samples (e.g. a group of clinically depressed outpatients).

The second main limitation of parametric IRT models is that these models assume that the calculated parameters are meaningful for the entire sample (Santor & Ramsay, 1998). In reality, data points in the less dense region of the data distribution will be estimated less efficiently than data points in the more dense regions. For clinical researchers it is individuals at the extremes who are often of most interest and arguably it might be advantageous to sacrifice some parameter stability for more accurate modelling of data in the less dense regions of a sample distribution.

Figure 2 Example IRT curves

It is important to note that the CES-D scale was not constructed with a parametric IRT model as the basis for item selection. For the reasons outlined above the 1PL model or 2 PL model may represent a less than ideal statistical approach for the IRT analysis of CES-D items. In response to concerns about the appropriateness of using logistic IRT models for data items which have not been constructed with the logistic model in mind, alternative nonparametric IRT models have been developed. It is to these nonparametric IRT models that attention is now given.

Non-parametric IRT models: TestGraf

A key feature of nonparametric IRT models is that they estimate response curves directly from the data and make no *a priori* assumptions about the underlying distribution of responses and their relation to the underlying trait. The main advantage of a nonparametric model is that these models generally achieve a better fit to the data when response option curves change rapidly with changes in the latent trait or show departures from monotonicity or unity. In this sense nonparametric models are consistent with a strategy emphasising exploratory data analysis prior to confirmatory analysis (Santor & Ramsay, 1998).

Given that there have been few IRT analyses of the CES-D generally, and with adolescent samples in particular, a nonparametric IRT approach is well suited to the current data and should provide new information about the CES-D. TestGraf (Ramsay, 2000) is a specialised statistical software program designed for use with psychometric test data. It has been used to perform IRT analyses of the BDI and CES-D referred to earlier and provided many useful insights into the psychometric properties of these scales. In light of its demonstrated utility to analyse the CES-D, TestGraf is used for the IRT analyses in the present study.

TestGraf is based on a kernel smoothing technique (or local averaging) to estimate the relation between choosing a response option to the value of the latent trait. A general introduction to kernel smoothing can be found in Altman (1992) and a quite detailed technical description of the method and algorithm as implemented in TestGraf can be found in Ramsay (1991, 2000). More didactic treatments of the TestGraf method can be found in the substantive papers of Santor and colleagues (Santor & Coyne, 1997; Santor et al., 1994; Santor et al., 1995).

In essence, TestGraf begins by ranking individuals on their raw total test scores. These rank values are converted to standard normal scores. Along a set of equally spaced evaluation points, dichotomous values for signifying whether or not an option was endorsed are then differently weighted for each individual. These weights are defined by a Gaussian kernel function which together with a bandwidth parameter determines the rate at which these weights fall to zero.

In most applications better precision in estimating an individual's score and the response curves is obtained by using ML estimation techniques. These ML estimates take into account the relative effectiveness of items and as such provide better estimates of an individual's true score on the latent construct. These ML estimates can then be fed back into TestGraf to rank order individuals for a second time.

The TestGraf program was developed by Ramsay from McGill University in Canada. The program and manual are available free of charge and can be down-loaded by the ftp communications utility from ego.psych.mcgill.ca/pub/ramsay/testgraf (accessed 6 February 2002). IRT analyses in the psychological literature are still relatively novel

and to assist in the replication of the findings from the present study the procedures involved in the TestGraf analyses carried out in this study are outlined below.

1. The main dataset for the study was stored in a SPSS file. An ASCI file was written out of SPSS using the Table command. This ASCI file was 20 columns wide (for 20 CES-D items) and in total (Boys & Girls) contained 6739 records.
2. The ASCI file was edited and two new lines inserted at the beginning of the file. The first line indicated the number of items and the number of characters used to identify each respondent. In the ASCI data file individual students were not identified so the number of characters was zero. The second line contains the answer key for multiple choice tests which have a correct answer. Even though the CES-D does not have a so-called 'correct' answer this line is still important because it is a prototype for how the data for each respondent is organised. An example of the first three lines of a ASCI file for this study is as follows:

   20 0

   33333333333333333333

   00012000200221113322
3. TestGraf was opened and the dialog window for a new job selected. The appropriate ASCI data file was chosen and item type 'Scale' as opposed to 'Multiple Choice' was selected. The data were then read into TestGraf. This step creates a file with an extension 'itm'. This file contains descriptive statistics for the sample to enable simple data checks.
4. The next step 'Analyse' actually performs the TestGraf analysis. The Analyse dialog screen asks for the number of display values and the smoothing parameter. The number of display values is changed from the default value of 51 to the maximum value of 101. This is done to improve the quality of the ML estimates which are based on the number of evaluation points selected in this step. The smoothing parameter is set at 0.30 which is slightly higher than the default value of 0.28.
5. ML estimates are then calculated using the 'Score' step. This step computes a total score for each respondent using ML estimation conditional on the curves computed previously in the Analyse step. The resulting file containing the ML estimates for each respondent is then fed back into the Analyse step and with the same settings as shown above the process is repeated. In all two ML iterations are performed.

The results presented in this chapter are produced following the process and settings as outlined above. A number of ASCI data files are processed because separate data files are required for each year level and for all year levels combined by gender. The TestGraf package is described as a program for the graphical analysis of questionnaire data and the primary output from the package are graphs. These graphs can be saved as postscript plots and then edited manually using the intermediate plotting language that is included with TestGraf. This process is quite tedious but it does produce high resolution plots of good quality. Figure 2 shows six graphs illustrating different features of the graphical results which are presented in this chapter.

The first graph in Figure 2 (graph A) shows four Option Characteristic Curves (OCCs) for a effective CES-D item. This is in fact Item 6 (*Depress*) for girls. OCCs

show both how discriminating response options are relative to other options and over what range of scores each option discriminates. It will be remembered that the CES-D asks respondents to indicate the frequency with which he or she experienced each item during the past week by checking one of four alternatives: rarely or none of the time (less than 1 day), some or a little of the time (1-2 days), occasionally or a moderate amount of the time (3-4 days), or most or all of the time (5-7 days). These response options are coded from zero to three respectively. Because the CES-D has four possible response options, four OCCs curves are shown in Figure 2(A).

In Figure 2(A) each OCC shows the probability that an option is endorsed (vertical axis) as a function of different levels of a latent variable (the horizontal axis labelled 'Score'). This latent variable, as outlined earlier, is derived from ML estimation procedures and in the present study is taken to reflect levels of depressive symptomatology. Traditionally the latent construct is referred to using the Greek small letter θ (theta). The vertical dashed lines shown in Figure 2(A) indicate the 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} quantiles. It can be seen that at the 50^{th} quantile the expected score, θ, is about 11. This closely corresponds to the raw total percentile score of 11 which is reported in the previous chapter.

Figure 2(A) shows that at low (< 10) levels of θ, Option 0 (experiencing the symptom rarely or none of the time), is the response option most likely to be endorsed. For example at an expected score of around five this option is nearly certain (> 0.95) to be endorsed. As expected, as θ increases the probability that Option 0 is endorsed decreases rapidly. For example, at an expected score of around 10 the probability that Option 0 is endorsed is only slightly greater than 50 per cent. Option 1 (experiencing the symptom some or a little of the time), increases in the low to moderate range of θ but then decreases rapidly. Option 2 (experiencing the symptom occasionally or a moderate amount of the time) increases in the moderate range of CES-D total expected scores but then decreases rapidly at expected scores of around 30. The probability of endorsing Option 3 (experiencing the symptom most or all of the time) begins to increase markedly for expected scores greater than 25 and for expected scores greater than around 35, it is the most likely (> 0.50) option to be endorsed.

The OCCs curves shown in Figure 2(A) illustrate several important features of an effective item. These are (a) the range in which options are endorsed can be fairly easily identified; (b) the OCCs change rapidly with changes in θ; and finally (c) the region in which options are endorsed corresponds with the ordinal position indicated by a weight assigned to the option. The second graph (Figure 2 (B)) on the other hand shows OCCs for a CES-D item exhibiting poor psychometric properties. This is in fact Item 17 (*Cry*) for boys. Across all levels of θ the most likely response option endorsed is Option 0. Even at an expected score of 30 (around the 95^{th} quantile) boys are more likely (nearly 0.60 probability) to report that they 'rarely or none of the time had crying spells' than to acknowledge the presence of this symptom (Options 1, 2 or 3). Further problems with this item are evidenced by the OCC for Options 1 and 2 which are very similar to each other and have a low probability of being endorsed at any level of θ.

From a set of OCCs an expected item score curve can be derived. In TestGraf this is calculated by summing the probability each option is endorsed with the *a priori* weight assigned to each option for each point where the curve is evaluated. Options that are endorsed more frequently and with larger weights contribute more strongly to the expected item score than do options endorsed less frequently and with smaller weights. These expected item scores are plotted across different levels of θ. These plots are referred to as Item Characteristic Curves (ICCs) and they show how

effective or discriminating items are to changes in levels of the latent variable. The third graph (Figure 2(C)) shows the ICC for Item 6 (*Depress*) for girls.

The ICC plotted in Figure 2(C) shows a very steep rate of increase in expected item score commensurate with increases in θ. This very steep rate of increase indicates that scores on this item are very sensitive to changes in θ (levels of depressive symptomatology). Because of this, Item 6 (*Depress*) for girls is considered to be an effective item. The error bars shown indicate 95 per cent confidence intervals at each of the evaluation points. As expected, with the relatively small number of observations available at the upper end of the range of expected scores (above 40) the standard errors of estimates in this range are much larger than those shown for low and moderate levels of θ. Nonetheless even at high levels of θ, the standard errors of estimate are less than ± 0.2 of predicted item scores.

Figure 2(D) shows the Item 6 (*Depress*) ICC for boys and girls plotted together on the same graph. It can be readily seen that the difference between the two ICCs across all levels of θ is minimal. This indicates that expected scores on Item 6 (*Depress*) for boys and girls are nearly identical for equal levels of depressive symptomatology. It would be concluded therefore that this item shows very little evidence of DIF. Figure 2(E) on the other hand plots the ICCs for Item 17 (*Cry*). Quite clearly the expected item score for girls is higher than for boys at all levels of θ. The size of this difference is about one half of a CES-D point and this difference is fairly constant or uniform across all levels of depressive symptomatology.

Figure 2(F) plots the ICCs for Item 1 (*Bothered*). For this item the expected item score for girls is slightly higher than boys at low to moderate ranges of depressive symptomatology but at very high scores (> 38) the expected item score for boys is higher than the corresponding score for girls. The DIF shown in Figure 2(F) therefore is not uniform across different levels of θ.

The graphs in Figure 2 illustrating DIF also include a summary statistic of the degree of bias. For example in Figure 2(D) the degree of bias is calculated as 0.055. This summary statistic is the weighted square difference between the ICCs for boys and girls with the difference at each evaluation point weighted by the proportion of boys and girls at each of these points. Larger values for this DIF summary statistic indicate a greater amount of discrepancy between the two curves. The summary statistic allows the amount of DIF to be compared across items but it does not in itself indicate whether the DIF is significant or not. Deciding whether any observed difference in ICCs is significant or not "… is a relative question and one that should be addressed with respect to both other items or samples of interest and the investigator's goals" (Santor et al., 1994, p. 258).

TestGraf analyses of CES-D items and response options

Figure 3 presents the TestGraf results for each CES-D item for the dataset of all year levels combined. Five graphs are presented for each item and each item is presented separately on its own page. The first two graphs for each item show the OCC plots – one for boys and one for girls. The second two graphs show the ICC plots – again one each for boys and girls. In the final graph (the fifth) the ICC for boys and the ICC for girls are plotted together (without their error bars) to facilitate a visual assessment of possible DIF. The summary statistic for the degree of DIF is also shown in this graph.

A considerable amount of information is contained in these 100 graphs (20 CES-D items by 5 graphs) and to assist readers the results are described in summary form below.

(a) Highly effective CES-D items: *Depress* (6), *Happy* (12), *Lonely* (14), *Enjoy* (16) and *Sad* (18).

Five CES-D items: *Depress* (6), *Happy* (12), *Lonely* (14), *Enjoy* (16) and *Sad* (18) have very good OCCs and ICCs. In each case the range in which response options are endorsed is able to be readily identified, the OCCs changed rapidly with changes in θ and the region in which response options are endorsed corresponded with the ordinal position indicated by the weight assigned to the option. Inspection of the ICCs showed that these items are very sensitive to changes in levels of depressive symptomatology across the majority of the sample. Generally these items are less effective in the bottom 25 per cent and top 5 per cent of the sample.

(b) Moderately effective CES-D items: *Blues* (3), *Good* (4), *Mind* (5), *Talk* (13), *Dislike* (19) and *Getgoing* (20).

Six CES-D items: *Blues* (3), *Good* (4), *Mind* (5), *Talk* (13), *Dislike* (19) and *Getgoing* (20) have OCCs and ICCs characteristic of moderately effective items. Very effective items produce OCCs with steep slopes which increase and decrease over a fairly narrow band of θ. These six items on the other hand yield OCCs, which while overall are satisfactory, failed to discriminate sharply across a narrow range of the expected score.

This point can be illustrated by contrasting the OCC for Option 1, Item 5 (*Mind*) for boys with the OCC for Option 1, Item 6 (*Depress*) for boys. For Item 6 (*Depress*) the OCC for Option 1 increased sharply across expected scores of 10 to 18 and then showed a similar rate of decrease across expected scores between 20 and 30. For Item 5 (*Mind*) the probability of endorsing Option 1 was relatively constant across expected scores of 10 to 25. Compared with the highly effective items identified in the previous section the poor OCCs for the items: *Blues* (3), *Good* (4), *Mind* (5), *Talk* (13), *Dislike* (19) and *Getgoing* (20) are reflected in less steep (less discriminating) ICCs.

(c) CES-D items dominated by Option 0: *Bothered* (1), *Appetite* (2), *Failure* (9), *Fearful* (10), *Sleep* (11), *Unfriendly* (15) and *Cry* (17).

The OCCs for seven CES-D items: *Bothered* (1), *Appetite* (2), *Failure* (9), *Fearful* (10), *Sleep* (11), *Unfriendly* (15) and *Cry* (17) indicate that for the majority of the sample the response option most likely to be endorsed is Option 0. This problem can be seen very clearly in the OCC plot for Item 1 (*Bothered*) for boys. For the first 75 per cent of the sample the probability that Option 0 would be endorsed is greater than 50 per cent. The probability that either Options 1 or 2 are endorsed remained relatively low across all levels of θ and the probability that Option 3 is endorsed increases only at very high levels of θ.

As a result of these problems these seven items are generally not very discriminating across low to moderate levels of depressive symptomatology. In fact the ICCs for these items show that the steepest increase in slope is in the region of the top 5 per cent of the sample or those with expected scores above 30.

(d) CES-D items which discriminate for only low levels of depressive symptomatology: *Effort* (7) and *Hopeful* (8).

Two items: *Effort* (7) and *Hopeful* (8) produce ICCs which show that they are highly discriminating for low levels of depressive symptomatology but relatively ineffective

across moderate and high levels of θ. The ICC for Item 7 (*Effort*) shows a steep increase in expected score between 0 and 5 but then actually a decrease between an expected score of between 5 and 10. For expected scores greater than 10, item scores increase very marginally. A similar, although less pronounced, pattern is observed for Item 8 (*Hopeful*). Inspection of the OCCs for both these items show that at very low levels of θ the probability of endorsing Option 1 increases markedly and that this response option remained relatively strong across the majority of sample scores.

Figure 3 CES-D OCCs and ICCs for boys and girls.

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

Figure 3 CES-D OCCs and ICCs for boys and girls (continued)

TestGraf gender DIF at the item level

The final graph for each item in Figure 3 shows the ICC for boys and the ICC for girls plotted together along with the summary statistic for the degree of DIF. These analyses were based on the dataset of all year levels combined. Table 24 presents these summary DIF statistics for each year level separately and for all year levels together.

Table 24 Gender DIF at the item level by year level

Item		Year 8	Year 9	Year 10	All Years
1	Bothered	0.091	0.092	**0.123**	0.092
2	Appetite	**0.136**	**0.121**	**0.186**	**0.138**
3	Blues	0.048	0.065	0.090	0.062
4	Good	0.054	0.056	0.066	0.051
5	Mind	0.097	0.028	0.044	0.045
6	Depress	0.041	0.062	0.085	0.055
7	Effort	**0.169**	0.097	**0.126**	**0.119**
8	Hopeful	0.070	0.041	0.058	0.039
9	Failure	0.025	0.024	0.017	0.019
10	Fearful	0.042	0.016	0.018	0.021
11	Sleep	**0.116**	0.072	**0.107**	0.091
12	Happy	**0.122**	0.094	**0.138**	**0.113**
13	Talk	0.031	0.031	0.031	0.017
14	Lonely	0.034	0.035	0.020	0.024
15	Unfriendly	0.072	0.073	0.077	0.069
16	Enjoy	0.054	0.059	0.079	0.058
17	Cry	0.094	**0.109**	**0.132**	**0.106**
18	Sad	0.081	0.095	**0.137**	0.092
19	Dislike	0.067	0.032	0.065	0.042
20	Get-going	0.040	0.054	0.042	0.041

DIF statistics ≥ 0.10 shown in bold

A considerable degree of consistency is shown across year levels with items showing low or high levels of DIF at one year level exhibiting a similar degree of DIF in the remaining year levels. DIF statistics greater than or equal or 0.10 are shown in bold text. Using this criterion of 0.10, seven items: *Bothered* (1), *Appetite* (2), *Effort* (7), *Sleep* (11), *Happy* (12), *Cry* (17) and *Sad* (18) show DIF in one or more year levels. Inspection of the ICCs show that five of these items: *Bothered* (1), *Appetite* (2), *Sleep* (11), *Cry* (17) and *Sad* (18) serve to increase scores for girls while the other two items: *Effort* (7) and *Happy* (12) increase scores for boys. The results for these items are described in further detail below.

Inspection of the ICCs for Item 1 (*Bothered*) shows that across most levels of depressive symptomatology (θ), item scores are slightly higher for girls than boys. At total expected scores of around 35 (roughly the top 5% of the sample) this difference is reversed with boys showing higher item scores than girls. Overall, it is clear that this item serves to increase slightly scores for girls.

The ICCs for Item 2 (*Appetite*) show clear evidence of DIF. In addition, the ICCs for boys and girls are most steep (most discriminating) only in the very severe range of θ. Girls show higher item scores than boys across all ranges of θ, a difference that is fairly uniform. The size of this difference is around just under one half of a CES-D point.

The ICCs for Item 7 (*Effort*) are quite unusual for both boys and girls. The ICCs for both boys and girls are most steep (most discriminating) in the very low range of θ. Boys show a higher score than girls across all ranges of θ, a difference that is fairly uniform, and equal to around just under one third of a CES-D point.

The ICCs for Item 11 (*Sleep*) exhibit DIF with item scores across most levels of θ (with exception of total scores between 25-35) slightly favouring girls by around just under one third of a CES-D point.

Item 12 (*Happy*) is a very effective item. The ICCs for boys and girls show that this item is very sensitive to changes in levels of θ across the majority of the sample and is particularly discriminating between levels of θ in the range 0 to 20. For the majority of the sample, item scores are higher for boys than girls although this difference is not uniform. The size of this difference is not large (less than one third of a CES-D point) and it is most pronounced for expected scores of around 10 (i.e. the median).

Item 17 (*Cry*) is not a particularly effective item for either boys or girls. This is because for the majority of the sample the response option most likely to be endorsed is Option 0. Girls show just over one half of a CES-D point higher item score, a difference which is relatively uniform across levels of θ.

The ICCs for Item 18 (*Sad*) exhibit features characteristic of an effective item. Nonetheless item scores across most levels of θ (with exception of total scores between 25-35) slightly favour girls by around less than one third of a CES-D point.

TestGraf gender DIF at the response option level

DIF can also be assessed at the response option level. For these analyses the dataset comprising all year levels combined is used. Table 25 presents the results from these analyses.

Table 25 Gender DIF at the option characteristic curve level

	Item	Option 0	Option 1	Option 2	Option 3
1	Bothered	**0.064**	**-0.055**	-0.011	0.003
2	Appetite	**0.098**	**-0.069**	-0.023	-0.005
3	Blues	0.036	-0.030	-0.006	-0.000
4	Good	0.033	-0.034	-0.023	0.024
5	Mind	-0.008	-0.007	0.009	0.006
6	Depress	0.029	-0.027	-0.005	0.003
7	Effort	**-0.077**	-0.009	0.038	**0.048**
8	Hopeful	0.003	-0.015	-0.016	0.027
9	Failure	-0.002	-0.000	0.002	0.000
10	Fearful	0.006	-0.011	0.004	0.000
11	Sleep	**0.057**	**-0.048**	-0.006	-0.003
12	Happy	**-0.067**	**0.044**	0.007	0.015
13	Talk	-0.000	-0.006	0.002	0.004
14	Lonely	0.017	-0.014	-0.002	0.000
15	Unfriendly	-0.039	0.029	0.008	0.002
16	Enjoy	-0.028	0.010	0.005	0.012
17	Cry	**0.058**	**-0.050**	-0.009	0.001
18	Sad	**0.058**	**-0.053**	-0.004	-0.000
19	Dislike	0.025	-0.024	0.000	-0.001
20	Get-going	0.003	-0.015	0.005	0.006

DIF statistics ≥ 0.05 shown in bold

Taking Item 1 (*Bothered*) by way of example the results presented in Table 25 are interpreted as follows. The direction of the summary DIF statistic (0.064) for Option 0 is positive. This indicates (because the boys served as the focus group) that on average across levels of θ the probability of endorsing this option is higher for boys than it is for girls. The results for Options 1 and 2 on the other hand suggest (because the summary DIF statistics are negative) that girls are more likely to endorse these options than are boys. The summary DIF statistic for Option 3 is positive but small (0.003).

In Table 25 DIF summary statistics greater than 0.05 are shown in bold face. As expected the seven items: *Bothered* (1), *Appetite* (2), *Effort* (7), *Sleep* (11), *Happy* (12), *Cry* (17) and *Sad* (18) showing large DIF values at the item level also show relatively large DIF values at the response option level. That is, the DIF results for these items at the response option level are consistent with the data reported earlier for DIF at the item level (see Table 24).

Five items: *Bothered* (1), *Appetite* (2), *Sleep* (11), *Cry* (17) and *Sad* (18) serve to increase scores for girls. The reason for this is that for these items boys are more likely to endorse Option 0 than girls and less likely than girls to endorse Option 1. The difference between boys and girls for these items with respect to Options 2 and 3

is quite small. In other words, for these items girls are more likely (at equal levels of θ) to acknowledge the presence of these symptoms than are boys.

Two items: *Effort* (7) and *Happy* (12) serve to increase scores for boys. For these items, boys are less likely than girls to endorse Option 0. For Item 12 (*Happy*) boys are more likely to endorse Option 1 than girls and the gender difference with respect to Options 2 and 3 is quite small. For Item 7 (*Effort*) on the other hand the gender difference with respect to Options 1 and 2 is quite small but boys are more likely than girls to endorse Option 3.

Impact of CES-D gender DIF at the total score level

In the previous section seven items were identified with relatively high levels of DIF. Five items: *Bothered* (1), *Appetite* (2), *Sleep* (11), *Cry* (17) and *Sad* (18) served to increase scores for girls while the other two items: *Effort* (7) and *Happy* (12) served to increase scores for boys. The magnitude of the DIF for each item is relatively minor but because more items serve to increase scores for girls it might be expected that the overall impact of gender DIF is to artificially increase scores at the total scale level for girls.

The actual impact of these items on total CES-D scores is examined by simply omitting these items when calculating a total raw score. Earlier, descriptive analyses were presented showing mean total and item scores for boys and girls separately (Table 10 and Table 14). For the data set of all year levels combined the boy mean raw CES-D total score is 10.80 and the girl mean raw CES-D total score is 13.67. The overall mean difference therefore between boys and girls is 2.87 CES-D points.

Taking Item 17 (*Cry*) by way of example the gender difference for this item is 0.27 (Boy mean 0.12; Girl mean 0.39). Therefore if Item 17 (*Cry*) is omitted from the calculation of the CES-D total raw score then the overall gender difference would reduce to 2.60. Conversely if Item 7 (*Effort*) is omitted, this would increase the overall gender difference by 0.20 to 3.07 CES-D points. This is because for this item the boy mean (1.06) is higher than the girl mean (0.86).

If all seven items showing DIF: *Bothered* (1), *Appetite* (2), *Effort* (7), *Sleep* (11), *Happy* (12), *Cry* (17) and *Sad* (18) are omitted when calculating CES-D total scores the boy mean CES-D score is 7.35 and the girl mean CES-D score is 9.05. Multiplying these scores by 1.54 (20 items / 13 items) to give comparable estimates with the 20 item total score gives a boy mean of 11.32 and a girl mean of 13.94 for a difference of 2.62. Comparing the difference of 2.62 and 2.87 it can be seen that the impact of gender DIF on total CES-D scores is around only one quarter (0.25) of a CES-D point.

TestGraf analyses of CES-D DIF across year levels

A series of TestGraf analyses are performed to examine the possibility of DIF across year levels. The summary DIF statistics for each CES-D item for boys and girls separately are presented in

Table 26. DIF statistics greater than or equal to 0.10 are shown in bold text. Using this criteria of 0.10, two items, Item 5 (*Mind*) and Item 7 (*Effort*) show the largest amount of DIF across year levels.

Table 26 Year level DIF by item and gender

Item		Boys	Girls
1	Bothered	0.044	0.087
2	Appetite	0.030	0.048
3	Blues	0.017	0.042
4	Good	0.030	0.051
5	Mind	**0.120**	**0.111**
6	Depress	0.023	0.044
7	Effort	**0.125**	**0.095**
8	Hopeful	0.088	0.050
9	Failure	0.025	0.013
10	Fearful	0.020	0.028
11	Sleep	0.033	0.029
12	Happy	0.085	0.073
13	Talk	0.039	0.047
14	Lonely	0.023	0.034
15	Unfriendly	0.046	0.060
16	Enjoy	0.075	0.049
17	Cry	0.026	0.038
18	Sad	0.036	0.048
19	Dislike	0.025	0.058
20	Get-going	0.055	0.055

DIF statistics ≥ 0.10 shown in bold

A number of graphs are produced for these two items showing OCCs and ICCs across year levels. For both items the plots are similar for boys and girls. Graphs are presented for Item 5 (*Mind*) using the girl dataset and for Item 7 (*Effort*) using the boy dataset. These graphs are shown in Figure 4 and Figure 5.

Figure 4 shows five graphs for Item 5 (*Mind*). The first four graphs plot the year level OCCs for each CES-D response option while the final graph plots the year level ICCs. In each graph three curves are shown reflecting the three year levels (8, 9 & 10). Turning to the ICC graph it can be seen that between Year 8 and Year 10 expected item scores tend to increase for those in the sample with low total expected scores. Inspection of the OCCs show that across year levels Option 0 becomes less likely to be endorsed in the sample while Option 1 becomes more likely to be endorsed.

Figure 4 Item 5 (*Mind*): DIF across year levels

Figure 5 Item 7 (*Effort*): DIF across year levels

In Figure 5 for Item 7 (*Effort*) between Year 8 and Year 10 the ICC graph shows that expected item scores decrease for those in the sample with low total expected scores. At the 25th percentile (or an expected score of around five) item scores decrease across year levels from about one and three quarters of a CES-D point to about one CES-D point. The OCCs show that across year levels Option 0 becomes more likely to be endorsed while Option 3 becomes less likely to be endorsed.

TestGraf analyses of the psychometric properties of the CES-D

In this section results relating to the psychometric properties of the CES-D as a measure of depressive symptomatology are presented. Item-total correlations and Cronbach's alpha presented in Chapter 5 (see Table 18) provide useful but limited information about the psychometric properties of the CES-D. These statistics (derived from classical test theory) are limited because they represent an average across levels of individual variation and do not take into account that test performance may vary across different levels of the trait or ability.

A major improvement provided by IRT is that the psychometric properties of a test can be examined at different levels of the underlying trait or ability. These results are presented in a series of four graphs shown in Figure 6 and in each graph two series of estimates (Boys & Girls) are plotted.

In Figure 6(A) the reliability coefficient for the CES-D as a function of θ is plotted. At median levels of depressive symptomatology (scores of around 10) the reliability of the CES-D in this sample is quite high (around 0.92). For total expected scores between 20 and 35 reliability estimates decrease sharply, down to around 0.87. For total expected scores of less than 20 (75% of the sample) reliability estimates are higher for girls than boys although this difference is not large.

The reliability estimates shown in Figure 6(A) are roughly comparable with the estimate of Cronbach's alpha (0.87 for boys and 0.92 for girls) provided earlier. It should be noted, however, that reliability coefficients measure both population heterogeneity as well as test quality. Later a better measure of test quality, the TIF (test information function) is presented.

In Figure 6(B) the standard deviation of observed score is shown. For the majority of sample with CES-D total observed scores between around 5 and 40 (90% of the sample) the standard deviation of score is higher for boys than for girls. For the sample with median levels of depressive symptomatology (observed total scores of around 10) the standard deviation of observed score is about 2.5 for girls and about 3.0 for boys. A 95 per cent confidence interval can be constructed by adding and subtracting twice this standard deviation to a specific score value.

For example the 95 per cent confidence interval for a girl scoring 10 on the CES-D is between 5 and 15. The 95 per cent confidence interval for a boy scoring 10 on the CES-D is between 4 and 16. These are clearly very wide confidence intervals but according to Ramsay (2000) a lack of measurement precision is typical of tests in which the total test score is calculated by simply adding together item scores.

In Figure 6(C) the standard error of expected score is shown. Expected score is the ML estimation of depressive symptomatology. Because the ML estimation makes use of much more information it would be expected that using this ML estimate

confidence limits will be smaller. At an expected score of 10, the 95 per cent confidence interval for boys is between 4.5 and 15.5 and for girls it is about the same.

Standard errors increase between expected scores of 10 to 30 to reach a maximum of four. For the sample scoring above the 95 per cent quantile standard errors are around three. The increase in standard errors for expected total scores between 10 and 30 indicates a deterioration in the quality of information provided by the CES-D for a key section of the sample.

The deterioration in the quality of information provided by the CES-D for the sample scoring between 10 and 30 is shown more clearly in Figure 6(D). This graph indicates the amount of information provided by the CES-D across levels of depressive symptomatology and is known as a Test Information Function (TIF). It will be recalled that TIFs take into account both how well tests discriminate between individuals and how precisely they measure the amount of the trait (in our case depressive symptomatology) an individual has. The better a test is able to discriminate between individuals and estimate those differences precisely the more information is provided by that test.

A formal definition of the information for an item, *j*, with *k* options is provided by Santor and Ramsay (1998, p. 357) as:

$$I_j(\theta) = \sum_{m}^{K} \frac{\left[P'_{jm}(\theta)\right]^2}{P_{jm}(\theta)}$$

This definition takes into account that the amount of information provided by an item is a function of the rate at which the probability of endorsing options changes, $P'_{jm}(\theta)$, the derivative of the slope of P_{jm} at a given level of θ and finally the amount of variance of error in estimating the response function for individual options. TIFs represent the average of item information functions for the entire measure.

Figure 6(D) shows very clearly that the CES-D (in this sample) provides high levels of information for median levels of depressive symptomatology (around a total score of 10) but relatively less information in the moderate range of θ. It can be noted that screening cut-points for the CES-D when used with adolescent samples are typically set in the moderate (between 20 and 30) range of total scores.

For the majority of the sample scoring less than 20 the quality of the information provided by the CES-D is higher for girls than boys. This difference is most pronounced for total expected scores of around 10. Interestingly, for the sample with total expected scores of between around 25 and 40 information estimates are higher for boys than for girls. This finding, however, should be interpreted cautiously because TestGraf information estimates are subject to sampling variation and can lack accuracy when there is not a substantial number of observations (Ramsay, 2000).

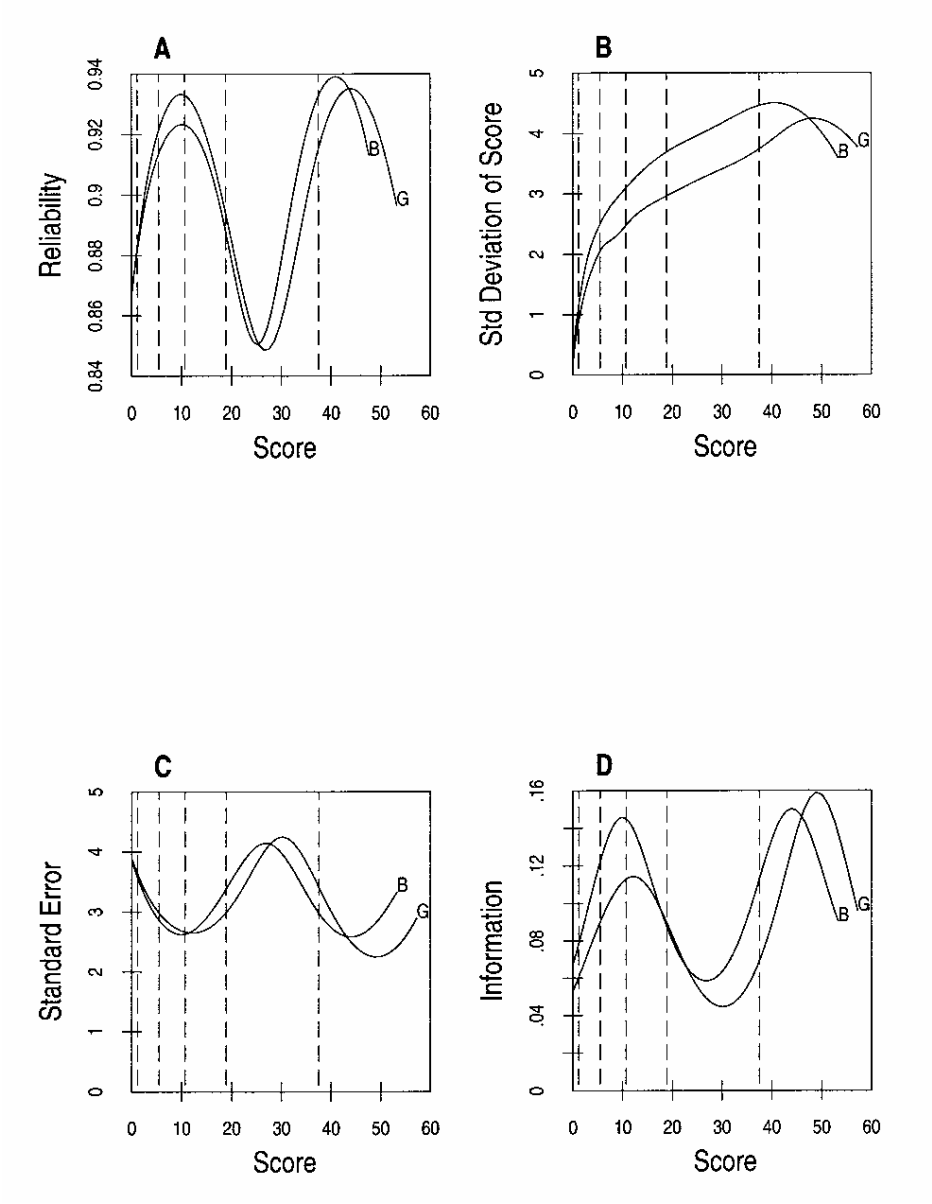


Figure 6 Psychometric properties of the CES-D

Summary

In this chapter the results from a series of non-parametric IRT analyses have been presented. These analyses were carried out with the TestGraf software package and a considerable amount of very useful graphical output was produced. The analyses were quick and easy to perform with few modelling decisions needed. In addition, no *a priori* assumptions of the underlying distribution of responses were required. As a consequence the results presented in this chapter, along with the earlier descriptive statistics, provide a solid foundation to guide the more complex SEM analyses that will follow in the next two chapters. The key findings presented in this chapter can be summarised as follows:

The CES-D items identified as most effective included: *Depress* (6), *Happy* (12), *Lonely* (14), *Enjoy* (16) and *Sad* (18). On a number of items: *Bothered* (1), *Appetite* (2), *Failure* (9), *Fearful* (10). *Sleep* (11), *Unfriendly* (15) and *Cry* (17) the majority of the sample (even those with moderately high total scores) endorsed Option 0. Two items: *Effort* (7) and *Hopeful* (8) were found to discriminate at only low levels of depressive symptomatology.

Seven items show relatively high levels of gender DIF. Five of these: *Bothered* (1), *Appetite* (2), *Sleep* (11), *Cry* (17) and *Sad* (18) served to increase scores for girls while the other two items: *Effort* (7) and *Happy* (12) served to increase scores for boys.

The reason the items: *Bothered* (1), *Appetite* (2), *Sleep* (11), *Cry* (17) and *Sad* (18) increased scores for girls can be seen at the response option level. For these items, girls were less likely than boys to endorse Option 0 and more likely to endorse Option 1.

Two items: *Effort* (7) and *Happy* (12) served to increase scores for boys. For these items, boys were less likely than girls to endorse Option 0. For Item 12 (*Happy*) boys were more likely to endorse Option 1 than girls. For Item 7 (*Effort*) the gender difference with respect to Options 1 and 2 was quite small but boys were more likely than girls to endorse Option 3.

Overall the impact on total scores of the gender DIF was small. An estimate of the impact the gender DIF at the total score level indicates that by omitting these seven items the mean gender difference might be reduced by about one quarter of a CES-D point.

The possibility of DIF across year levels was examined. Item scores for Item 7 (*Effort*) were found to decrease slightly across year levels primarily because Option 0 became more likely to be endorsed by students in Year 9 and Year 10. For Item 5 (*Mind*) the opposite trend was evident with item scores increasing across year levels. This occurred because Option 0 became less likely to be endorsed.

7

SEM Confirmatory Factor Analyses

This chapter presents the results of a series of structural equation modelling (SEM) confirmatory factor analyses (CFA) of the CES-D. The SEM analyses performed in the present study use the M*plus* software and a non technical (as far is as possible) introduction to the approach taken by M*plus* for the analysis of categorical data is provided. It is argued that the M*plus* software program is particularly suited to the present study because of its ability to estimate threshold measurement parameters that are allowed to vary across groups in a multiple group analysis. Details of the M*plus* syntax programs written for the present analyses are provided to allow future replication of the results by other researchers.

Initially, a number of different proposed CES-D factor models are tested in a CFA framework. As an important preliminary step to developing the measurement model to be used in the invariance analyses which follow in the next chapter, a series of higher order CFAs of the CES-D are also carried out. These higher order CFAs replicate previous analyses carried out in older samples and examine the extent to which the CES-D exhibits unidimensionality. Finally, a comparison of the Maximum Likelihood (ML) and Weighted Least Squares (WLS) estimation techniques as implemented in LISREL and M*plus* is made to identify possible problems in previous CES-D SEM analyses which have relied almost exclusively on LISREL ML techniques.

The M*plus* software program

For most of the 1990s specialised SEM software for the analysis of categorical data was not widely available or easy to use. In 1998, the M*plus* program (Muthén & Muthén, 1998) was released (see Maydeu-Olivares, 2000 for review). This program is much easier to use than the program it superseded (LISCOMP) and it incorporates a number of features which make it more flexible than conventional SEM software for modelling categorical data. A full technical description of the M*plus* approach to

categorical variables is provided in a number of papers (see Muthén, 1978, 1984, 1989a, 1989b; Muthén, Kao & Burstein, 1991) and here only a brief discursive account is given.

The general approach taken by M*plus* for analysing ordinal variables using latent variable models is known as an 'underlying response variable approach' (Jöreskog & Moustaki, 2001). The M*plus* technique is also referred to as a categorical variable methodology (CVM: Kaplan, 1991). In this approach it is assumed that for each ordinal variable y there is an underlying continuous variable y^*. It is these underlying y^* variables that are used in the analyses and not the observed ys. The underlying variable assigns a metric to the ordinal variable using parameters which are called thresholds. Each item with C categories contributes $C - 1$ thresholds. For example CES-D items would each contribute three thresholds (four CES-D response options provide three transition thresholds) to a model.

Thresholds represent the expected value of the latent variable at which an individual transitions from one response option to another response option. On a four point rating scale, the first threshold represents the expected value at which an individual would be most likely to transition from a value of zero to a value of one. The second threshold represents the expected value at which an individual would be most likely to transition from a value of one to a value of two on the outcome variable and so on through to the third threshold (change from response option 2 to 3). Thresholds therefore connect each observed variable to a latent continuous response variable which it is assumed would have been available if measurement had been more precise.

A threshold model addresses the problem that ordinal variables might not be normally distributed (Bollen, 1989). In addition to thresholds, for categorical models in M*plus* (and LISREL) a polychoric correlation matrix rather than a covariance matrix is used as input. A polychoric correlation is a correlation in the bivariate normal distribution from a pair of underlying y^* ordinal variables. These are calculated for every item pair. The polychoric correlation matrix (usually estimated by Maximum Likelihood (ML) techniques) calculated from the underlying y^* ordinal variables is used in place of the usual covariance matrix of the observed variables.

The underlying response variable approach is implemented in both M*plus* and LISREL although the computational methods are different. M*plus* uses a three stage approach outlined by Muthén (1984). In the first stage thresholds, means and variances are estimated by ML. This is followed by estimation of the polychoric correlation matrix by conditional ML given the earlier estimates. In the third stage the parameters of the structural part of the model are estimated using a variant of the weighted least squares method. Importantly, in a multiple group analysis separate sample estimates of thresholds and polychoric correlations are calculated. In this manner threshold parameters can be held equal or allowed to vary across groups.

When a polychoric correlation matrix is used rather than a covariance matrix as input different parameters enter the model. With a polychoric matrix, the variances of the outcome variables are not used and the residual variances of the outcome variables cannot be not identified and are not estimated. M*plus* models for categorical variables use scale factors to capture group differences in latent variable response variances. These scale factors refer to the inverted standard deviations of the latent response variable and are functions of loadings, factor variances and residual variances. By using scale factors the polychoric correlation parameters can vary across groups in a multiple group analysis and thus allow testing of both the equality of factor loadings and latent variable variances.

A key advantage of the M*plus* program over LISREL for the analysis of categorical data is that M*plus* allows modelling of categorical data with the less restrictive assumption of conditional normality of the latent response variables. This means that normality is only assumed for the residuals of each latent variable regressed on to the factor thereby avoiding the more restrictive assumptions of multivariate normality (MVN) required with ML estimation or underlying bivariate normality associated with using polychoric matrices in a LISREL weighted least squares (WLS) analysis.

M*plus* also includes an estimator known as 'WLSMV' which is an abbreviation for 'weighted least square parameter estimates using a diagonal weight matrix with robust standard errors and mean and variance adjusted chi-square statistic'. Preliminary Monte Carlo simulation analyses using this estimator suggest that with 12 observed binary indicators sample sizes as small as 200 are sufficient to produce reliable estimates (Muthén, 2001c). This is considerably less than the sample sizes (2000 to 5000) previously thought necessary for a SEM analysis with WLS estimation.

These two key advantages of M*plus*, namely, modelling with less restrictive normality assumptions and with smaller sample sizes, are the principal reasons that leading SEM commentators (see Rigdon, 2000; 2001) recommend M*plus* for the analysis of categorical data. There is a further advantage to M*plus* that is particularly important to the present study. This is, that M*plus* is unique among mainstream SEM software programs because it can estimate threshold measurement parameters in categorical variable models that are allowed to vary across groups in a multiple group analysis. In contrast, a multiple group LISREL analysis proceeds under the assumption of equal thresholds (see Jöreskog, 2001c; Muthén & Asparouhov, 2002 for details).

Why is modelling thresholds important to the present study? Earlier it was argued that previous SEM gender studies had inadequately tested for measurement invariance. This was because only group differences in factor loadings (metric invariance) had been examined. Scalar invariance, equality of intercepts for continuous variables or thresholds for categorical variables, is also necessary for mean group comparisons. Tests of scalar invariance have not yet been performed and therefore previous gender group mean CES-D comparisons might not be valid. By using M*plus*, scalar invariance can be tested by examining whether threshold parameters vary across groups. It is for this reason that M*plus* was selected for the SEM analyses performed in the present study.

M*plus* model estimation and output

M*plus* is described by its developers as a statistical modelling program with an easy-to-use-interface. True to this description, the user language for M*plus* is relatively straightforward to master although specifying models by writing syntax programs is required. Some prior familiarity with SEM is obviously an advantage. All the analyses reported in this, and in the next chapter, use M*plus* version 2.1. Copies of the key M*plus* syntax programs that were written for the SEM analyses are collated in Appendix C to enable replication by future researchers. These are identified in the text of this chapter.

A variety of different SEM analyses are performed but all share several common features. These include:

A common data set is used. This dataset is identical to that used for the descriptive and IRT analyses presented earlier. The dataset consists of 6739 observations. These observations comprise Boys: 3743; Girls: 2996 and by year levels Year 8: 2306, Year 9: 2275, Year 10: 2158.

The data-file is read into M*plus* in a similar manner for each analysis as follows:

```
DATA:
FILE IS bg123.DAT;
FORMAT IS 1F2.0 22F1.0;
VARIABLE:
NAMES ARE
    sch gender wave
    dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
    dep10 dep11 dep12 dep13 dep14 dep15 dep16
    dep17 dep18 dep19 dep20;
CATEGORICAL ARE
dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
    dep10 dep11 dep12 dep13 dep14 dep15 dep16
    dep17 dep18 dep19 dep20;
USEVARIABLES ARE
dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
    dep10 dep11 dep12 dep13 dep14 dep15 dep16
    dep17 dep18 dep19 dep20;
```

During the analyses a bug was discovered involving the USEVARIABLES command. If the order of variables on this command is different from that shown on the NAMES command the program produces incorrect latent mean estimates in a multiple group analysis. This bug was only picked from cross-checking the results with the earlier IRT analyses, highlighting one of the advantages of checking complex analyses by using different methods with a common dataset. This bug was reported to M*plus* product support and was rectified in an update to the program (version 2.12).

All analyses are performed using the weighted least squares (WLS) estimator for categorical variables.

No modifications in terms of correlated errors are used in any models to improve fit.

M*plus* produces a narrower range of goodness-of-fit statistics for categorical variable analyses compared with those available in continuous variable models. Five fit statistics for categorical models are shown in M*plus* output. The formulas for these are presented in Appendix 5 of the M*plus* users guide and they are briefly described below.

The chi-square (χ^2) statistic and its associated *p* value, are used to judge the goodness-of-fit of a SEM (Bagozzi, 1981; Bollen, 1989). The chi-square statistic

measures the discrepancy between the sample covariance (correlation) matrix and the fitted covariance (correlation) matrix. A model producing a small chi-square value indicates a better fitting model than one showing a larger chi-square value. A draw back of the chi-square statistic is that it is influenced by sample size. This means that even substantively trivial discrepancies can lead to rejection of an otherwise satisfactory model in large samples (Loehlin, 1998).

The comparative fit index (CFI) compares the covariance matrix predicted by the model to the observed covariance matrix to gauge the per cent of lack of fit which is accounted for by moving from the null model to the proposed SEM model. CFI values vary between zero and one with values closer to one indicating good fit. By convention, CFIs should be equal to or greater than 0.90 to accept the model, indicating that 90 per cent of the covariation in the data can be reproduced by the given model.

The Tucker-Lewis Index (TLI) is also known as the Non-Normed Fit Index. TLI values range from zero to one with higher values indicating better fit. The TLI, like the CFI, is classed as a incremental or comparative fit statistic and is derived from a comparison of a hypothesised model with a suitably defined null model. Marsh, Balla and Hau (1996) argue that the TLI is particularly useful in testing nested models because it takes into account goodness of fit as well as parsimony. In this respect, the TLI incorporates a penalty for model complexity and a reward (shown by higher values) for model simplicity.

The Root Mean Square Error of Approximation (RMSEA) fit index is a population-based index that is gaining in popularity among the plethora of fit indexes proposed for SEM (Loehlin, 1998). This index measures the amount of deviation between the hypothesised model and the observed data and the number of free parameters in the model (Byrne, 1998). Browne and Cudeck (1993) suggest that RMSEA values of about 0.05 or less indicate a close fit of the model, values of about 0.08 indicate a reasonable error of approximation and models showing RMSEA values greater than 0.10 should not be used. Confidence intervals to assess the precision of RMSEA values are not provided in the output of M*plus* categorical variable models.

The Standardised Root Mean Square Residual (SRMR) is the average difference between the predicted and observed variances and covariances in the model, based on standardised residuals. Standardised residuals are fitted residuals divided by the standard error of the residual. The smaller the SRMR the better the model fit. According to Hu and Bentler (1999) SRMR values should be below 0.08.

The use of fit statistics in SEM is quite controversial. Much depends on the use of a particular fit measure for a specific situation. In the present study the main use of fit statistics is to evaluate whether across group (gender or year level) restrictions in multiple group analyses are necessary for valid latent mean comparisons. The manner in which fit statistics are used in the present study to evaluate measurement invariance models is discussed in the next chapter.

Confirmatory factor analyses of the CES-D

In this section five simple factor models for the CES-D are tested in a confirmatory factor analysis framework. The models are tested for boys and girls separately and together. For each of the models tested, a path diagram is presented. The format for the SEM path diagrams follow a standard widely adopted in the literature. Readable accounts of the customary notation and symbols for SEM path diagrams are readily

available (Byrne, 1998; Hoyle & Panter, 1995; McDonald & Ho, 2002). Key aspects include:

> Ellipses represent unobserved latent factors and rectangles represent observed variables. The Greek letter Xi (ξ) is used to label and number the unobserved exogenous latent variables.
>
> Factor loadings are denoted by the Greek letter Lambda (λ). These indicate the strength of the correlation between the observed items and the latent variables.
>
> Subscripts associated with the lambda coefficients are keyed to the direction of the arrow. For example $\lambda_{15.2}$ in Figure 9 represents the regression of Item 15 (*Good*) to the second factor in the model (Positive Affect).
>
> Associated with each observed variable is an error term, denoted by the Greek letter Theta Delta (□). Associated with each factor being predicted is a disturbance term denoted by the Greek letter Zeta (□). For clarity, in most of the diagrams presented here, these factor residual error terms are not shown.
>
> Where more than one factor is included in the model the correlation between the factors is denoted by the Greek letter Phi (ϕ) and shown by a curved two way arrow.

In the models estimated the direction of the paths is from the latent variable to the observed variable. This implies that depressive symptomatology (the latent factor) is a general condition of the individual which is reflected by the presence and strength of symptoms of depression (the CES-D items). This is in contrast to a model which might posit that depressive symptomatology is simply a summary construct reflecting the degree to which a person experiences depressive symptoms. It is more likely that there will be positive correlations between CES-D items and that CES-D items are indicators of a more general or underlying condition termed depressive symptomatology. In other words, it is assumed that the observed variables are the effects of the latent variables and not the causes of them.

The five CES-D models tested are described below. The M*plus* syntax used to estimate the four factor CES-D model is provided in Appendix C (Program 3.1). The results tables presented for the CFAs show, for each model tested (and by gender), goodness-of fit statistics, factor loadings and where more than one factor is included in the model, factor correlations.

One factor model

The one factor model is the first model tested and it assumes that all CES-D items reflect one factor, namely depressive symptomatology (DS). This model is implicitly assumed when the CES-D is summed to produce a total CES-D score. From the perspective classical measurement theory this model is referred to as a 'true score model' because all the items are assumed to be true measures of the same underlying construct and nothing else. In giving each symptom item the same weight when manually scoring the CES-D it is also assumed that the regression coefficients or factor loadings (reflected in the path between observed item and the latent construct) are equivalent. The form of this model is shown in

Figure 7 and the results from the CFA presented in Table 27.

Two factor model

A two factor model, derived from a CFA in a sample of Hong Kong Chinese married couples by Cheung and Bagley (1998), is the second model tested. In this model items comprising the Depressed Affect (DA), Somatic (SOM) and Positive Affect (POS) factors are joined to form one factor. The two Interpersonal (INTER) items comprise the other factor. The rationale for this model is based around the notion that Chinese people minimise the difference between depressive and somatic symptoms and that the positive affect items are simply antonyms to depressed mood. The form of this model is shown in Figure 8 and the results in Table 28.

Three factor model

A three factor model has been proposed by a number of researchers (Beals et al., 1991; Dick et al., 1993; Manson et al., 1990; Prescott et al., 1998; Ying, 1988) particularly for when the CES-D is used with non-Western samples. In this model items comprising the Depressed Affect (DA) and Somatic (SOM) factors are joined to form one factor. This is based on the notion that non-Western people minimise the difference between depressive and somatic symptoms. The form of this model is shown in Figure 9 and the results from the CFA presented in Table 29.

Four factor model

The four factor CES-D model tested in the presented study is based on the traditional and most widely recognised factor structure for the CES-D. The form of this model is shown in Figure 10 and the results from the CFA presented in Table 30.

Five factor model

A five factor CES-D model was proposed by Thorson and Powell (1993) following an exploratory factor analysis of the CES-D in a sample 400 adults. A conceptual advantage of this model is that it includes a fifth factor (self worth: SW) that Radloff (1977) had designed in her original construction of the scale. There are a number of other small differences in this model compared to the four factor model including the inclusion of Item 13 (*Talk*) and Item 14 (*Lonely*) to the Interpersonal factor. The form of this model is shown in Figure 11 and the results from the CFA presented in Table 31.

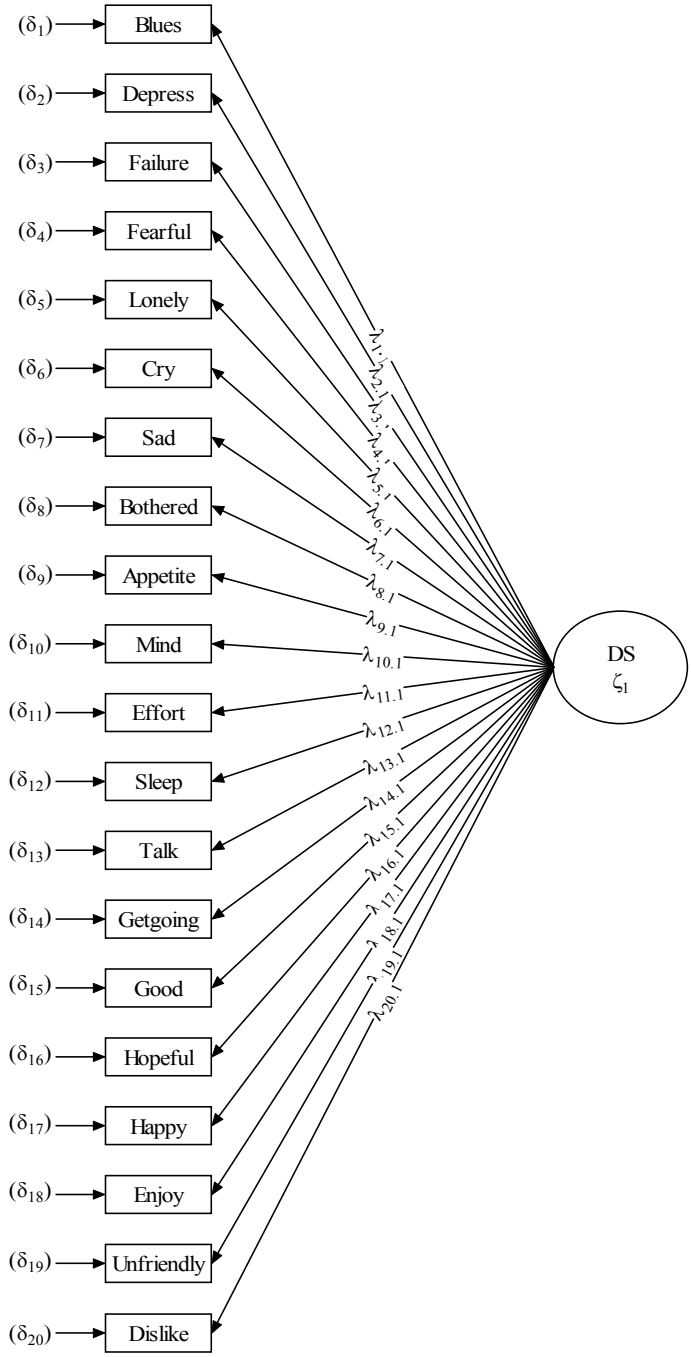


Figure 7 One factor CES-D model

Table 27 One factor CES-D model

	Boys	Girls	Boys + Girls
χ^2	1886.29	1654.00	3254.60
df	170	170	170
CFI	0.914	0.943	0.924
TLI	0.904	0.936	0.915
RMSEA	0.052	0.054	0.052
SRMR	0.172	0.183	0.170
3 Blues	0.87	0.89	0.88
6 Depress	0.91	0.94	0.93
9 Failure	0.90	0.92	0.90
10 Fearful	0.79	0.80	0.79
14 Lonely	0.90	0.91	0.90
17 Cry	0.75	0.85	0.83
18 Sad	0.93	0.94	0.93
1 Bothered	0.73	0.75	0.74
2 Appetite	0.55	0.64	0.63
5 Mind	0.65	0.72	0.68
7 Effort	0.14	0.37	0.21
11 Sleep	0.59	0.66	0.63
13 Talk	0.64	0.69	0.67
20 Getgoing	0.76	0.84	0.79
4 Good	0.65	0.77	0.71
8 Hopeful	0.52	0.62	0.57
12 Happy	0.85	0.90	0.86
16 Enjoy	0.87	0.91	0.89
15 Unfriendly	0.79	0.76	0.75
19 Dislike	0.91	0.91	0.91

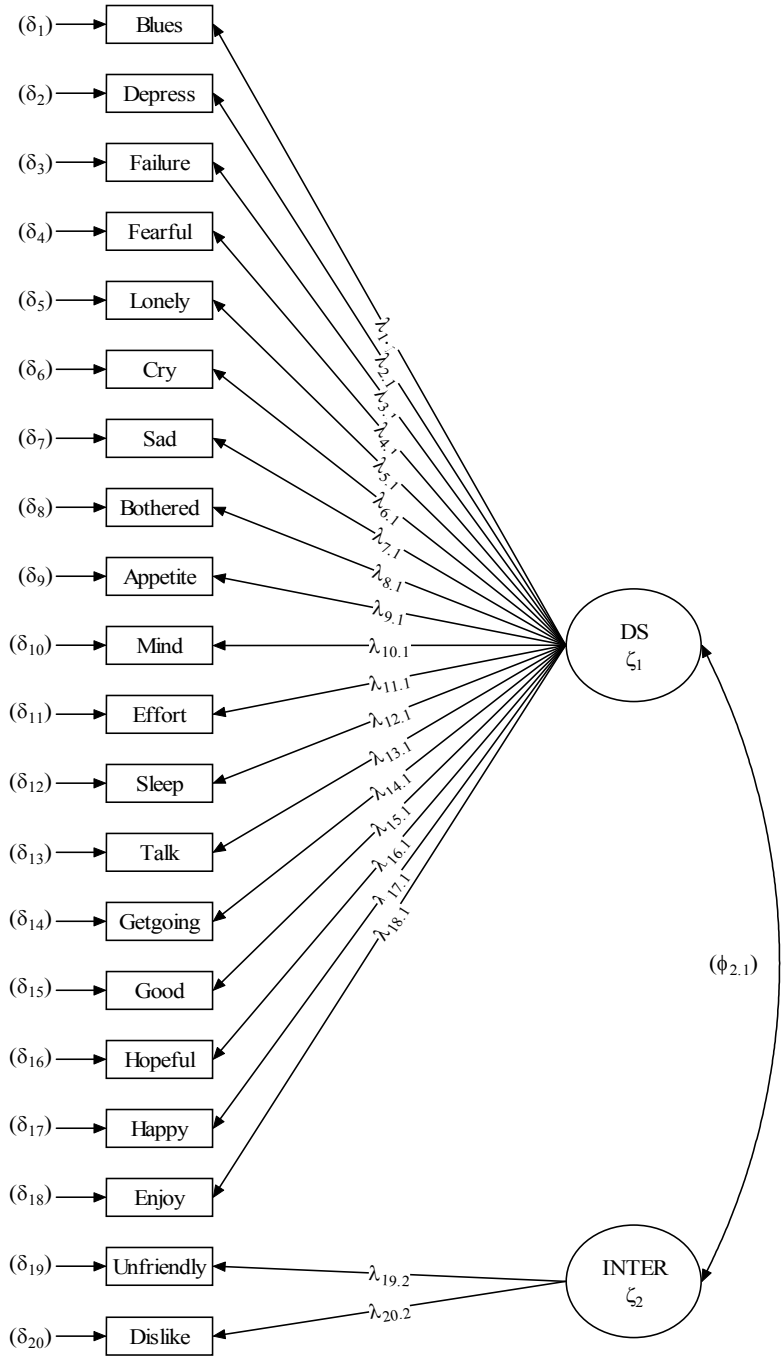


Figure 8 Two factor CES-D model

Table 28 Two factor CES-D model

	Boys	Girls	Boys + Girls
χ^2	1641.36	1457.31	2840.79
df	169	169	169
CFI	0.927	0.950	0.934
TLI	0.918	0.944	0.926
RMSEA	0.048	0.050	0.048
SRMR	0.152	0.164	0.152
Depressed Affect / Somatic / Positive Affect			
3 Blues	0.87	0.89	0.88
6 Depress	0.90	0.94	0.92
9 Failure	0.88	0.91	0.89
10 Fearful	0.78	0.80	0.78
14 Lonely	0.90	0.91	0.90
17 Cry	0.74	0.85	0.83
18 Sad	0.92	0.94	0.93
1 Bothered	0.72	0.74	0.74
2 Appetite	0.56	0.65	0.64
5 Mind	0.65	0.72	0.68
7 Effort	0.13	0.37	0.22
11 Sleep	0.59	0.66	0.63
13 Talk	0.64	0.69	0.66
20 Getgoing	0.74	0.83	0.78
4 Good	0.64	0.76	0.70
8 Hopeful	0.52	0.62	0.56
12 Happy	0.85	0.90	0.86
16 Enjoy	0.87	0.91	0.89
Interpersonal			
15 Unfriendly	0.78	0.75	0.74
19 Dislike	0.96	0.97	0.97
Factor Correlations			
Dep Aff/Inter	0.86	0.86	0.85

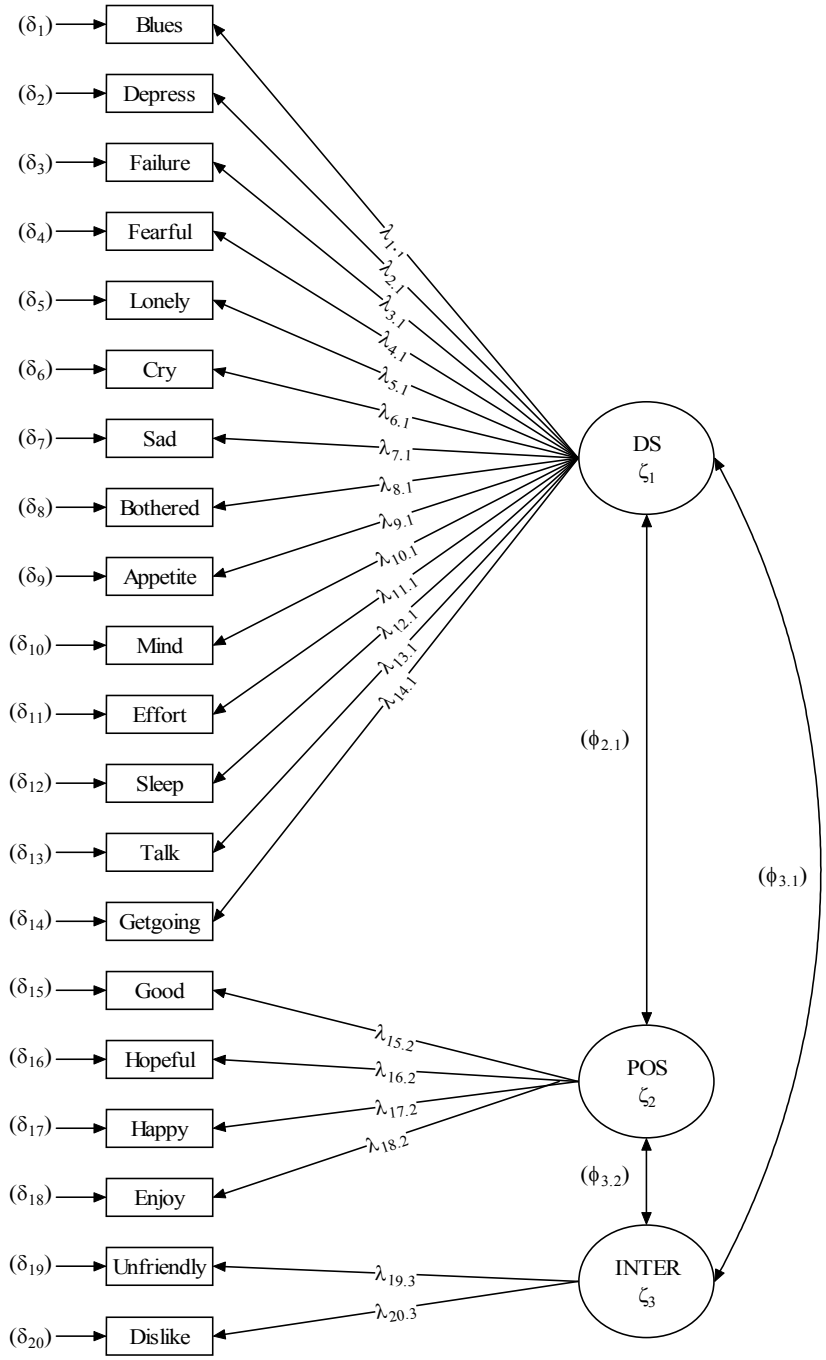


Figure 9 Three factor CES-D model

Table 29 Three factor CES-D model

	Boys	Girls	Boys + Girls
χ^2	1358.26	1204.53	2262.28
df	167	167	167
CFI	0.941	0.960	0.948
TLI	0.932	0.955	0.941
RMSEA	0.044	0.046	0.043
SRMR	0.112	0.135	0.114
Dep Aff/ Somatic			
3 Blues	0.86	0.88	0.87
6 Depress	0.89	0.94	0.91
9 Failure	0.87	0.90	0.88
10 Fearful	0.79	0.81	0.79
14 Lonely	0.89	0.90	0.89
17 Cry	0.75	0.86	0.84
18 Sad	0.91	0.93	0.92
1 Bothered	0.72	0.74	0.74
2 Appetite	0.56	0.65	0.64
5 Mind	0.65	0.73	0.68
7 Effort	0.15	0.39	0.24
11 Sleep	0.59	0.67	0.64
13 Talk	0.65	0.70	0.67
20 Getgoing	0.74	0.83	0.78
Positive Affect			
4 Good	0.64	0.76	0.69
8 Hopeful	0.54	0.63	0.58
12 Happy	0.85	0.90	0.86
16 Enjoy	0.88	0.91	0.89
Interpersonal			
15 Unfriendly	0.80	0.76	0.75
19 Dislike	0.95	0.97	0.97
Factor Correlations			
Dep Aff/Pos Aff	0.83	0.88	0.83
Dep Aff/Inter	0.86	0.85	0.85
Pos Aff/Inter	0.69	0.75	0.70

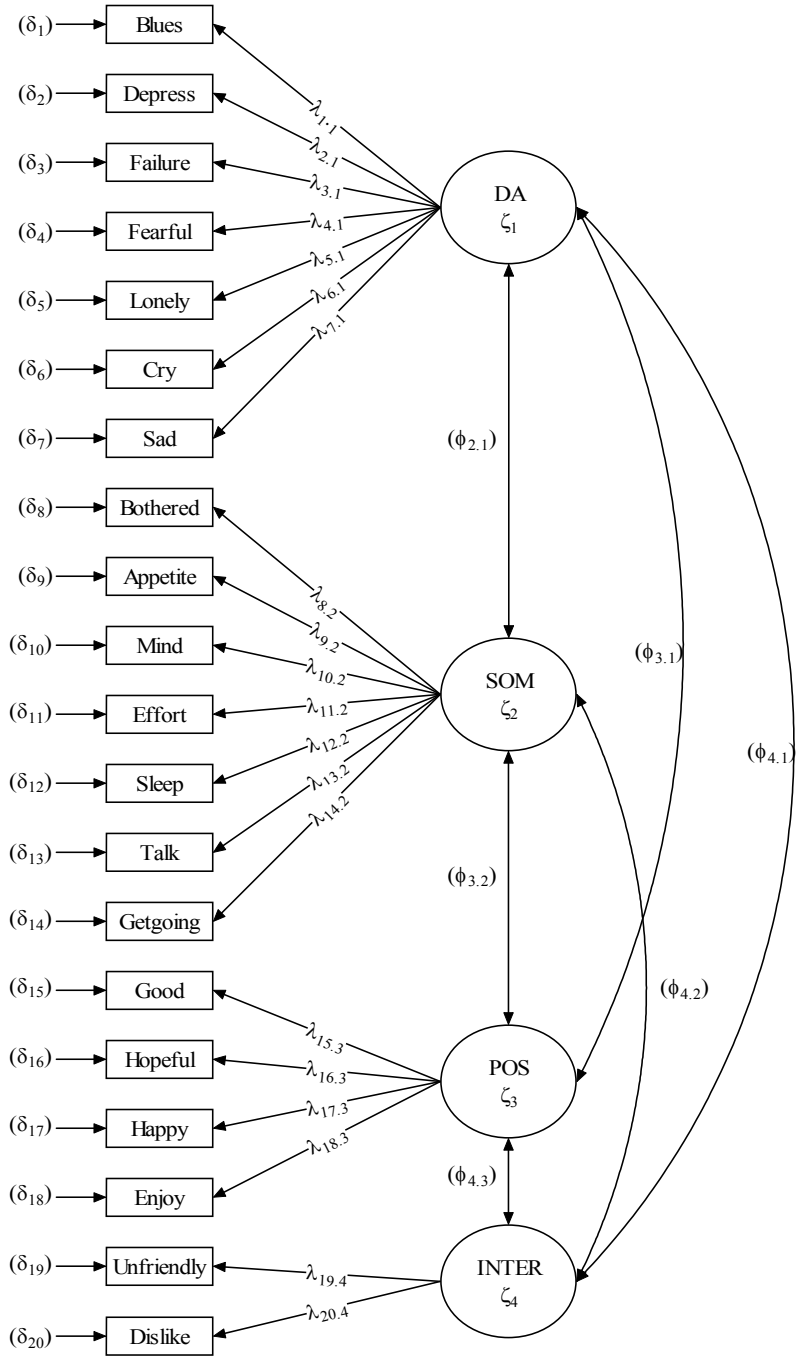


Figure 10 Four factor CES-D model

Table 30 Four factor CES-D model

	Boys	Girls	Boys + Girls
χ^2	1232.95	1131.67	2066.35
df	164	164	164
CFI	0.947	0.963	0.953
TLI	0.938	0.957	0.946
RMSEA	0.042	0.044	0.041
SRMR	0.103	0.127	0.106
Depressed Affect			
3 Blues	0.86	0.88	0.87
6 Depress	0.89	0.94	0.91
9 Failure	0.87	0.90	0.88
10 Fearful	0.78	0.80	0.79
14 Lonely	0.89	0.90	0.89
17 Cry	0.75	0.86	0.84
18 Sad	0.91	0.93	0.92
Somatic			
1 Bothered	0.74	0.75	0.75
2 Appetite	0.58	0.66	0.65
5 Mind	0.67	0.74	0.70
7 Effort	0.19	0.40	0.26
11 Sleep	0.61	0.67	0.65
13 Talk	0.67	0.70	0.68
20 Getgoing	0.76	0.84	0.80
Positive Affect			
4 Good	0.65	0.76	0.70
8 Hopeful	0.54	0.63	0.59
12 Happy	0.86	0.90	0.87
16 Enjoy	0.88	0.91	0.89
Interpersonal			
15 Unfriendly	0.79	0.76	0.75
19 Dislike	0.95	0.97	0.96
Factor Correlations			
Dep Aff./Som	0.94	0.96	0.96
Dep Aff/Pos Aff	0.83	0.88	0.84
Dep Aff/Inter	0.86	0.85	0.85
Som/Pos Aff	0.71	0.81	0.74
Som/Inter	0.78	0.81	0.79
Pos Aff/Inter	0.67	0.75	0.68

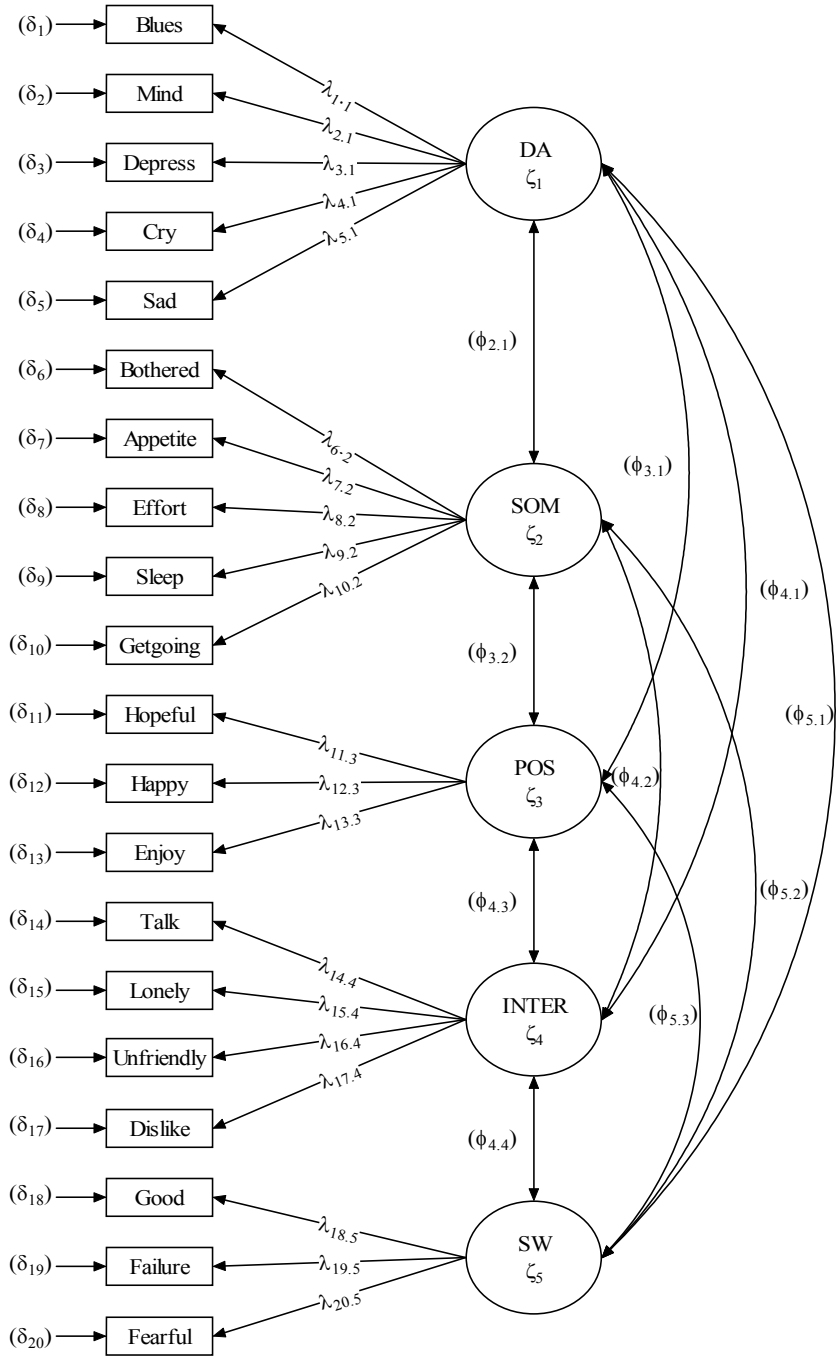


Figure 11 Five factor CES-D model

Table 31 Five factor CES-D model

	Boys	Girls	Boys + Girls
χ^2	1500.62	1338.31	2510.22
df	160	160	160
CFI	0.933	0.955	0.942
TLI	0.921	0.946	0.931
RMSEA	0.047	0.050	0.047
SRMR	0.132	0.145	0.128
Depressed Affect			
3 Blues	0.87	0.88	0.87
5 Mind	0.64	0.73	0.68
6 Depress	0.90	0.94	0.92
17 Cry	0.75	0.85	0.82
18 Sad	0.92	0.94	0.93
Somatic			
1 Bothered	0.72	0.74	0.73
2 Appetite	0.55	0.63	0.62
7 Effort	0.14	0.37	0.22
11 Sleep	0.59	0.65	0.63
20 Getgoing	0.74	0.82	0.78
Self worth			
4 Good	0.66	0.77	0.72
9 Failure	0.90	0.92	0.91
10 Fearful	0.81	0.80	0.80
Positive affect			
8 Hopeful	0.53	0.62	0.57
12 Happy	0.85	0.90	0.86
16 Enjoy	0.87	0.91	0.89
Interpersonal			
13 Talk	0.66	0.69	0.67
14 Lonely	0.92	0.92	0.91
15 Unfriendly	0.81	0.77	0.76
19 Dislike	0.92	0.92	0.91
Factor Correlations			
Dep Aff./Som	0.99	0.99	0.99
Dep Aff/SW	0.93	0.95	0.93
Dep Aff/Pos Aff	0.86	0.89	0.86
Dep Aff/Inter	0.95	0.93	0.94
Som/SW	0.93	0.96	0.94
Som/Pos Aff	0.86	0.91	0.86
Som/Inter	0.96	0.96	0.96
SW/Pos Aff	0.92	0.92	0.91
SW/Inter	0.90	0.93	0.91
Pos Aff/Inter	0.79	0.86	0.81

The results of CFA of the five models tested indicate that the four factor model provides the best fit to the data. A summary of the model fit statistics is presented in Table 32. For the dataset comprising both boys and girls the four factor model shows a lower chi-square statistic, RMSEA value and SRMR estimate (indicating better fit) and a higher CFI and TLI (also indicating better fit) than any other model. Using the rules of thumb outlined earlier, the RMSEA value of 0.041, the CFI value of 0.953 and the TLI value of 0.946 indicate the fit to data from the four factor model is good. Detracting from this is the SRMR value of 0.106 which is higher than the cut-off score of 0.08 recommended for good models.

The parameter estimates for the four factor model are sensible and for the most part standardised factor loading are above 0.60. The main exception to this is the factor loading for Item 7 (*Effort*) to the Somatic factor. For boys the factor loading for this item is very low (0.19) while for girls it is better (0.40). For the data set combining boys and girls this factor loading is only 0.26. Most factor correlations are in the range between 0.67 to 0.88. Of significance, for both boys and girls, the correlation between the Depressed Affect and Somatic factors is very high and around 0.96, indicating a possible lack of discriminant validity for this model.

Table 32 Summary of CES-D factor model fit statistics

	χ^2	df	CFI	TLI	RMSEA	SRMR
Boys						
One factor	1886.29	170	0.914	0.904	0.052	0.172
Two factor	1641.36	169	0.927	0.918	0.048	0.152
Three factor	1358.26	167	0.941	0.932	0.044	0.112
Four factor	1233.95	164	0.947	0.938	0.042	0.103
Five factor	1500.62	160	0.933	0.921	0.047	0.132
Girls						
One factor	1654.00	170	0.943	0.936	0.054	0.183
Two factor	1457.31	169	0.950	0.944	0.050	0.164
Three factor	1204.53	167	0.960	0.955	0.046	0.135
Four factor	1131.67	164	0.963	0.957	0.044	0.127
Five factor	1338.31	160	0.955	0.946	0.050	0.145
Boys and Girls						
One factor	3254.60	170	0.924	0.915	0.052	0.170
Two factor	2840.79	169	0.934	0.926	0.048	0.152
Three factor	2262.28	167	0.948	0.941	0.043	0.114
Four factor	2066.35	164	0.953	0.946	0.041	0.106
Five factor	2510.22	160	0.942	0.931	0.047	0.128

Higher order factor analyses of the CES-D

Based on the four factor model a second-order CFA is performed to determine whether or not the four first-order dimensions can be modelled using a single second-order depression factor. This model is shown in Figure 12 and essentially replicates previous (Hertzog et al., 1990; McCallum et al., 1995) second-order factor analyses of CES-D. The M*plus* syntax used to estimate this model is provided in Appendix C (Program 3.2). The results from this analysis are presented in Table 33.

The fit of the second-order model for both boys and girls is good with values for the fit indices (with the exception of the SRMR) within accepted cut-points. The loss of fit from the first-order four factor model with unconstrained factor correlations, however, is statistically significant. Using the dataset for both boys and girls this is calculated by taking the difference between the chi-square value from the four factor model (χ^2 = 2066.35, df = 164) from the chi-square value from the second-order model (χ^2 = 2144.62, df = 167). The difference in these values (78.27) for three degrees of freedom is highly significant ($p < 0.01$).

The large difference in chi-square values is, at least in part, simply a reflection of the statistical power of the sample. The other fit indices show very minor, perhaps best described as negligible, signs of deterioration of fit (e.g. the difference in RMSEA values was only 0.001). On this basis it can be concluded that nearly all the information in the correlations between the first-order factors is accounted for by the second-order factor loadings.

Following the lead of McCallum et al. (1995) in order to test better whether the four CES-D factors arise from one superordinate factor a version of a Schmid and Leiman second-order model is estimated. This model is referred to as a nested factor model (Gustafsson & Balke, 1993) and it is shown in Figure 13. This model is relatively uncommon in the literature (but see Newmann, 1984 for an early example) and so is specified in detail here. The M*plus* syntax used to estimate this model is provided in Appendix C (Program 3.3).

Using the notation of Gustafsson and Balke (1993) the form of the model is one general factor (in our case termed 'depressive symptomatology') with a relationship to all the observed variables (CES-D items). Included in the model are three specific factors with a more narrow range of influence. For example, the latent variable (ξ_1) corresponds to the Somatic factor. Preliminary modelling using four specific factors (i.e. including a Depressed Affect factor) produced standardised loading for the items comprising this factor of 0.00. This indicates (not surprisingly) that once the variation in the general factor 'depressive symptomatology' is accounted for there is very little left in common among these items. In the interests of parsimony therefore this specific factor was not included in the nested model.

Note that all factors in the model are orthogonal and that the variance estimates of all factors are fixed at one. This is shown in Figure 13 by the lack of any paths between the factors and in the equation (Ψ = 1*) alongside of each latent variable signalling the presence of a fixed parameter. These two features are necessary to partition item variance between the general factor, the specific factor and residual or error variance. In order to identify the model the factor loadings among each of the specific latent variables are constrained to be identical.

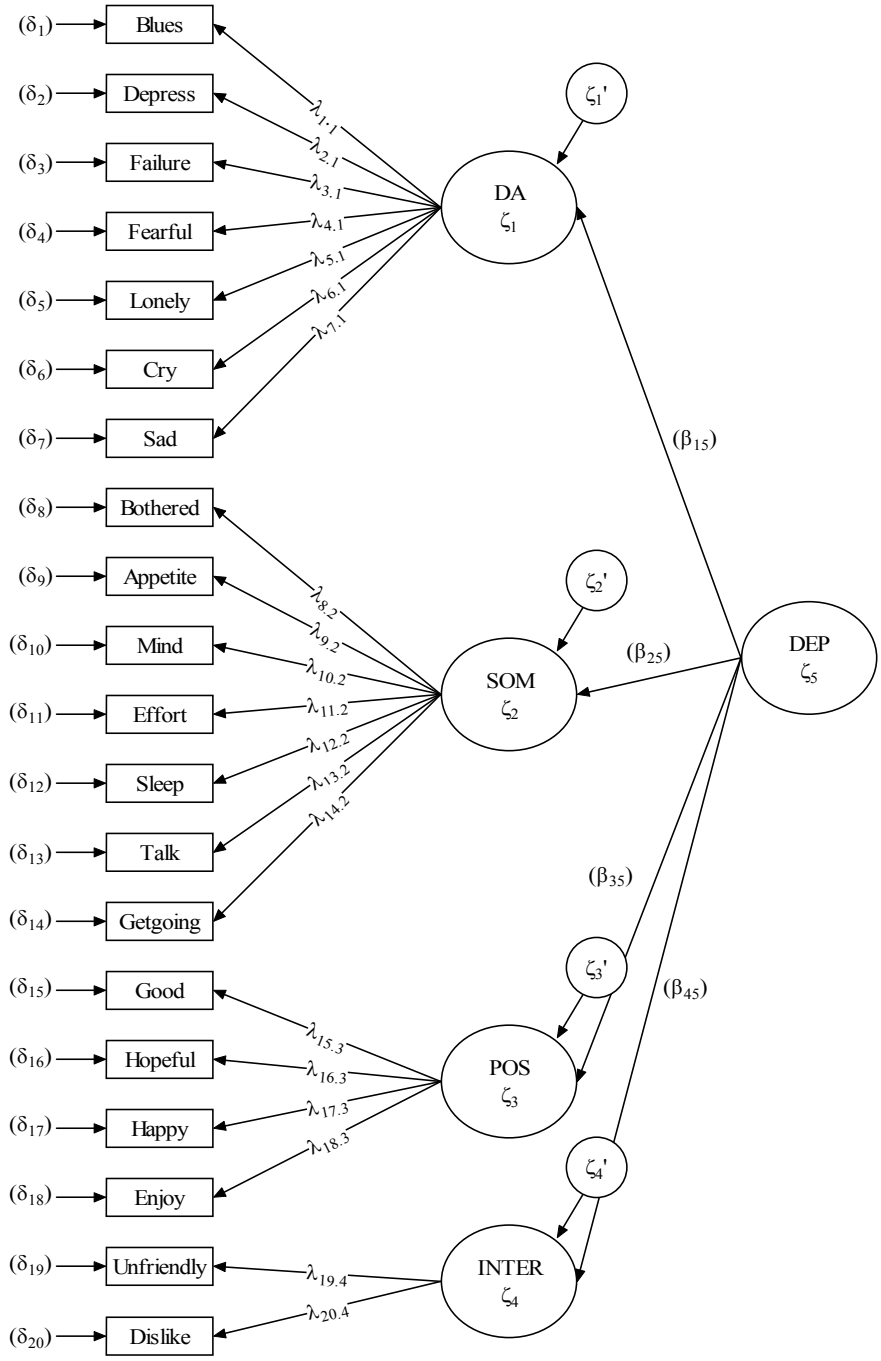


Figure 12 Second-order CES-D model based on the four factor solution

Table 33 Second-order CES-D model based on the four factor solution

	Boys	Girls	Boys + Girls
χ^2	1280.77	1147.87	2144.62
df	167	167	167
CFI	0.945	0.962	0.951
TLI	0.937	0.957	0.944
RMSEA	0.042	0.044	0.042
SRMR	0.102	0.127	0.105
Depressed Affect			
3 Blues	0.86	0.88	0.86
6 Depress	0.89	0.94	0.91
9 Failure	0.87	0.90	0.88
10 Fearful	0.78	0.80	0.79
14 Lonely	0.89	0.90	0.89
17 Cry	0.74	0.86	0.84
18 Sad	0.91	0.93	0.92
Somatic			
1 Bothered	0.73	0.75	0.75
2 Appetite	0.57	0.66	0.65
5 Mind	0.67	0.73	0.69
7 Effort	0.16	0.40	0.24
11 Sleep	0.61	0.67	0.65
13 Talk	0.66	0.70	0.67
20 Getgoing	0.75	0.84	0.79
Positive Affect			
4 Good	0.64	0.76	0.70
8 Hopeful	0.54	0.63	0.59
12 Happy	0.85	0.90	0.86
16 Enjoy	0.88	0.91	0.89
Interpersonal			
15 Unfriendly	0.79	0.76	0.74
19 Dislike	0.95	0.97	0.96
Second order factor loadings			
Depressed Affect	0.99	0.99	0.99
Somatic	0.93	0.93	0.96
Positive Affect	0.83	0.83	0.88
Interpersonal	0.85	0.85	0.85

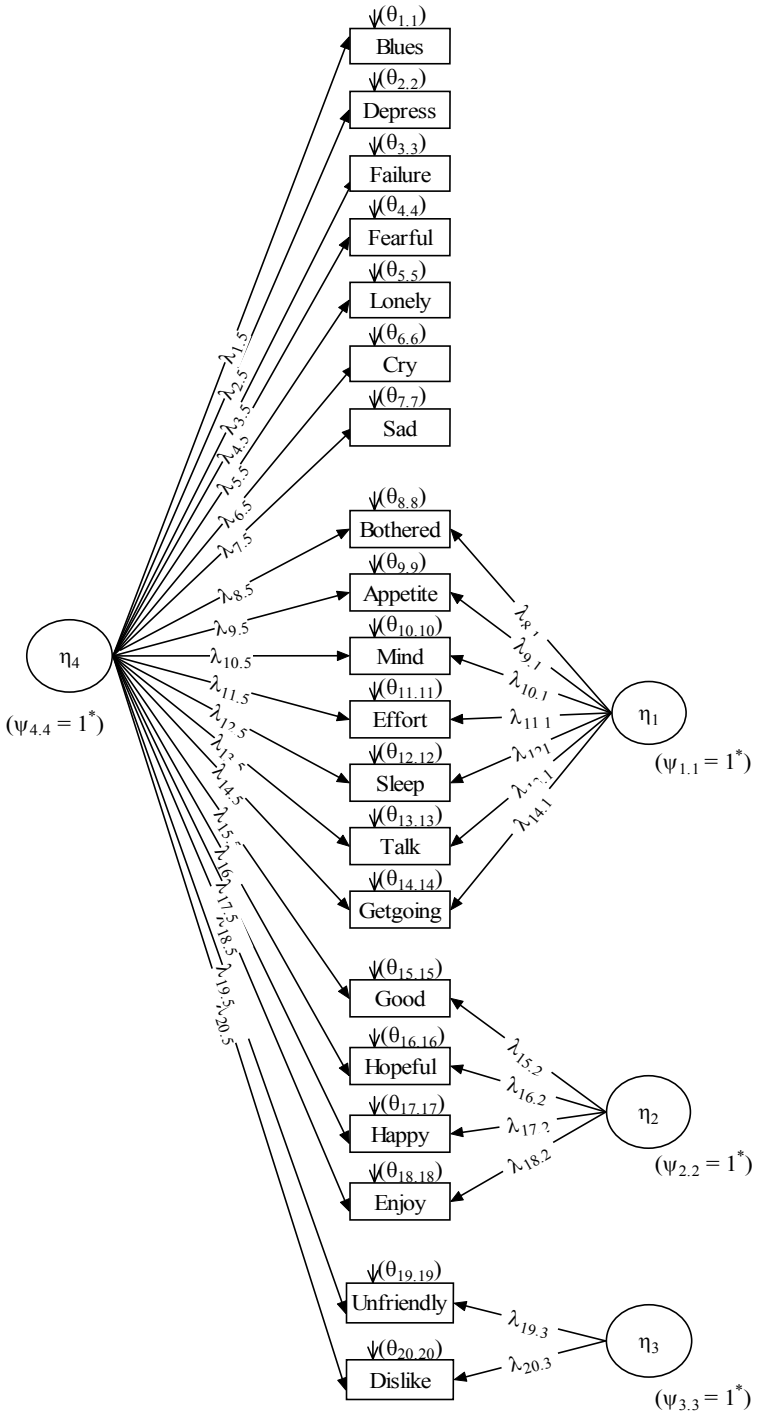


Figure 13 Nested three factor CES-D model

The calculation of the item variance accounted for in the model is relatively straightforward. Taking Item 1 (*Bothered*) by way of example the variance is calculated as:

$$\sigma_{8.8}^2 = \lambda_{8.5}^2 + \lambda_{8.2}^2 + \psi_{8.8}$$

The first terms represents the contribution of the general factor, the second term the contribution of the specific factor and the third term the error or residual variance. In the case of Item 1 (*Bothered*), using the figures shown in Table 34 the factor loadings are 0.72 (general factor), 0.22 (specific factor) and the residual variance is 0.43. Squaring the factor loadings gives 0.518 and 0.048 and summed to the residual variance (0.43) produces 0.996 (or unity allowing for rounding). The percentage variance accounted for by each of the latent variables is the sum of the squared factor loadings divided by 20 (for 20 CES-D items).

The results from the nested model are presented in Table 34 (for boys and girls), Table 35 (for boys) and Table 36 (for girls). The fit of the nested models in each of the datasets is good. For the dataset comprising only boys the fit indices are: (χ^2 = 1257.93, df = 166, $p < 0.001$, CFI = 0.946, TLI = 0.938, RMSEA = 0.042, SRMR = 0.103). For the dataset comprising only girls the fit indices are: (χ^2 = 1097.67, df = 166, $p < 0.001$, CFI = 0.964, TLI = 0.959, RMSEA = 0.043, SRMR = 0.122). For the dataset comprising both boys and girls the fit indices are: (χ^2 = 2085.07, df = 166, $p <$ 0.001, CFI = 0.953, TLI = 0.946, RMSEA = 0.041, SRMR = 0.104).

For each of the models estimated the general 'Depression' factor accounts for just over 50 per cent of the proportion of item variance. This proportion is slightly higher for girls (58%) compared with boys (50%). The factor loadings for this general factor are consistently high with the exception of Item 7 (*Effort*). This factor loading is particularly low for boys (0.13). Around 8 per cent of variance in total is accounted for by the specific factors.

Interestingly the specific factor accounting for most variation (around 4%) in item scores is Positive Affect. This suggests that the Positive Affect items are not simply antonyms to the depressed mood items and this in turn helps explain why models which combine the items from these two factors provide poorer fits to the data than models that maintain them as two separate factors.

The residual or error variation in each of the models is higher than what might be desired. For girls it is 35 per cent but for boys it is 42 per cent. This indicates that for boys almost one half of the variation in item scores is unaccounted for by the latent variables in the hypothesised model.

Table 34 Nested three factor CES-D model (Boys & Girls)

	General factor	Somatic	Positive Affect	Interpersonal	Residual Variance
3 Blues	0.87				0.25
6 Depress	0.92				0.16
9 Failure	0.88				0.23
10 Fearful	0.79				0.37
14 Lonely	0.89				0.20
17 Cry	0.83				0.31
18 Sad	0.93				0.14
1 Bothered	0.72	0.22			0.43
2 Appetite	0.61	0.22			0.57
5 Mind	0.65	0.22			0.52
7 Effort	0.22	0.22			0.90
11 Sleep	0.61	0.22			0.58
13 Talk	0.65	0.22			0.45
20 Getgoing	0.76	0.22			0.38
4 Good	0.57		0.45		0.47
8 Hopeful	0.41		0.45		0.62
12 Happy	0.74		0.45		0.24
16 Enjoy	0.76		0.45		0.22
15 Unfriendly	0.63			0.45	0.40
19 Dislike	0.82			0.45	0.13
Variance Explained	53.81%	1.74%	4.09%	2.03%	38.32%

Table 35 Nested three factor CES-D model (Boys)

	General factor	Somatic	Positive Affect	Interpersonal	Residual Variance
3 Blues	0.86				0.26
6 Depress	0.90				0.19
9 Failure	0.86				0.26
10 Fearful	0.79				0.38
14 Lonely	0.88				0.22
17 Cry	0.74				0.45
18 Sad	0.92				0.16
1 Bothered	0.70	0.25			0.45
2 Appetite	0.54	0.25			0.65
5 Mind	0.62	0.25			0.55
7 Effort	0.13	0.25			0.92
11 Sleep	0.56	0.25			0.62
13 Talk	0.62	0.25			0.56
20 Getgoing	0.71	0.25			0.44
4 Good	0.51		0.45		0.54
8 Hopeful	0.37		0.45		0.66
12 Happy	0.73		0.45		0.27
16 Enjoy	0.75		0.45		0.24
			0.45		
15 Unfriendly	0.67			0.45	0.35
19 Dislike	0.81			0.45	0.14
Variance Explained	50.37%	2.10%	3.98%	2.00%	41.50%

Table 36 Nested three factor CES-D model (Girls)

	General factor	Somatic	Positive Affect	Interpersonal	Residual Variance
3 Blues	0.88				0.22
6 Depress	0.94				0.12
9 Failure	0.90				0.20
10 Fearful	0.80				0.36
14 Lonely	0.91				0.18
17 Cry	0.86				0.26
18 Sad	0.93				0.13
1 Bothered	0.73	0.21			0.43
2 Appetite	0.62	0.21			0.57
5 Mind	0.70	0.21			0.47
7 Effort	0.37	0.21			0.81
11 Sleep	0.63	0.21			0.55
13 Talk	0.68	0.21			0.49
20 Getgoing	0.81	0.21			0.30
4 Good	0.65		0.42		0.40
8 Hopeful	0.48		0.42		0.59
12 Happy	0.80		0.42		0.18
16 Enjoy	0.80		0.42		0.19
15 Unfriendly	0.64			0.45	0.39
19 Dislike	0.83			0.45	0.12
Variance Explained	58.11%	1.60%	3.54%	1.98%	34.77%

CES-D normality tests

In this section a series of normality tests are applied to CES-D items. According to McDonald and Ho (2002) normality tests are required as a prerequisite to ML estimation. In the present study ML is not used and instead the normality tests are performed as a preliminary step prior to examining the extent to which previous SEM analyses (based on LISREL ML techniques) which have ignored the ordinal nature of CES-D items might be in error. The normality tests are based around a CFA of the four factor CES-D model using the dataset combining boys and girls across year levels.

There are a multitude of methods for assessing MVN (see Mecklin & Mundfrom, 2002 for review) but a series of tests developed by D'Agostino (1986) and Mardia (1970) are widely used. These tests are implemented in PRELIS (a pre–processor to LISREL) (see Bollen. 1989, p. 418 – 425 for details) and are applied to the present dataset. Initially, univariate normality is tested for each CES-D item individually. The results are presented in Table 37.

From Table 37 it can be seen that all CES-D items are positively skewed indicating a distribution with an asymmetric tail towards more positive values. In addition, with the exception of the four positive affect items: *Good* (4), *Hopeful* (8), *Happy* (12) and *Enjoy* (16) and the items: *Mind* (5) and *Effort* (7) the kurtosis values are positive suggesting a relatively peaked distribution for the majority of items. The statistical significance of these values is tested by the *Z* test statistic.

Table 37 Test of univariate normality for CES-D items

	Skewness			Kurtosis		
	Value	Z-Score	P	Value	Z-score	P
1 Bothered	1.84	42.48	0.00	2.79	46.84	0.00
2 Appetite	1.90	43.36	0.00	2.82	47.19	0.00
3 Blues	1.94	43.82	0.00	2.80	46.95	0.00
4 Good	0.62	19.12	0.00	-0.92	-15.49	0.00
5 Mind	0.72	21.89	0.00	-0.51	-8.55	0.00
6 Depress	1.40	35.90	0.00	0.95	15.99	0.00
7 Effort	0.67	20.47	0.00	-0.83	-13.91	0.00
8 Hopeful	0.39	12.58	0.00	-1.03	-17.23	0.00
9 Failure	2.54	50.74	0.00	5.71	95.65	0.00
10 Fearful	2.51	50.44	0.00	6.00	100.40	0.00
11 Sleep	1.30	34.17	0.00	0.62	10.35	0.00
12 Happy	0.99	28.02	0.00	-0.02	-0.27	0.78
13 Talk	1.38	35.57	0.00	1.11	18.61	0.00
14 Lonely	1.90	43.31	0.00	2.77	46.40	0.00
15 Unfriendly	1.79	41.81	0.00	2.61	43.70	0.00
16 Enjoy	0.92	26.60	0.00	-0.30	-5.01	0.00
17 Cry	2.87	53.97	0.00	8.03	134.62	0.00
18 Sad	1.71	40.71	0.00	2.24	37.51	0.00
19 Dislike	1.65	39.87	0.00	2.04	34.11	0.00
20 Getgoing	1.44	36.47	0.00	1.29	21.62	0.00

The very high *Z* scores for every item, with one exception for Item 12 (*Happy*), indicate the values of skewness and kurtosis are statistically different (at $p < 0.01$)

from those of a normal distribution. For many items these values exceed the limits (-1.0 to 1.0) proposed by Muthén and Kaplan (1985) for a ML factor analysis of non-normal categorical variables.

An omnibus test for the joint hypothesis of no multivariate skew or excess kurtosis is provided in PRELIS. Not surprisingly given that every item failed to exhibit univariate normality, the multivariate test of normality produced a chi-square value of 52322.08 which is statistically significant at $p < 0.001$. Overall the results from these normality analyses indicate that the assumption of MVN is not sustainable in the present data.

Given that the assumption of MVN is not met, a LISREL researcher, with an adequate sample size, could investigate the possibility of a LISREL WLS analysis. A LISREL WLS approach assumes that there is an underlying continuous variable for each ordinal variable which is reflected in the ordinal variables being bivariately normally distributed. This can be tested empirically using PRELIS. For each pair of item combinations PRELIS calculates a chi-square goodness of fit test of the model of an underlying bivariate normal distribution and also a variant to the RMSEA measure of population discrepancy.

The chi-square goodness of fit test is sensitive to sample size and therefore in large samples it can incorrectly reject the hypothesis of underlying bivariate normality. Jöreskog (2001b) argues that the preferred test of the assumption of underlying bivariate normality is shown by the PRELIS RMSEA statistic. This statistic is calculated for each item pair-wise comparison. According to Jöreskog (2001b) simulation studies have found that there are no serious effects of non-normality unless the RMSEA statistic is larger than 0.1. Associated with each PRELIS RMSEA value is a *p* value to test the hypothesis that the population value of RMSEA is less than 0.1.

In the present data set with 20 items, 190 tests are performed (one test for each pair of items). For the majority of item pairs PRELIS RMSEA values are between 0.020 and 0.050. The two highest RMSEA values are produced for the item pairs (*Enjoy* (16) – *Effort* (7): 0.089) and (*Enjoy* (16) – *Happy* (12): 0.098). The *p* value test for the hypothesis of approximate underlying bivariate normality is not rejected for any pair of variables. These results suggest that the hypothesis of approximate underlying bivariate normality holds for the item pairs in the present sample. This means that the polychoric correlation matrix estimated by PRELIS from the present data is suitable for further WLS analyses with LISREL.

Comparison of ML and WLS estimation techniques

In this section a comparison of ML and WLS estimation techniques is made. The purpose of this comparison is to gauge the extent of possible problems in previous CES-D SEM ML analyses. LISREL (version 8.52) and M*plus* (version 2.1) software packages are used for the analyses which are based on four factor CES-D model.

Four separate analyses are carried out. The first two analyses use the ML estimation technique implemented in LISREL and M*plus*. This technique treats CES-D items as continuous variables and replicates the approach used in the majority of SEM CES-D studies performed to date. Then, CES-D variables are treated as ordinal and the WLS techniques in LISREL and M*plus* are applied. The results are presented in Table 38.

Table 38 Comparison of ML and WLS in LISREL and M*plus*

	ML		WLS	
	LISREL	M*plus*	LISREL	M*plus*
χ^2	2808.49	2808.81	1553.89	2066.35
df	164	164	164	164
CFI	0.95	0.948	0.94	0.953
TLI	0.94	0.940	0.93	0.946
RMSEA	0.050	0.049	0.035	0.041
SRMR	0.033	0.033	0.064	0.106
Depressed Affect				
3 Blues	0.58	0.58	0.83	0.87
6 Depress	0.72	0.72	0.88	0.91
9 Failure	0.49	0.49	0.84	0.88
10 Fearful	0.38	0.38	0.75	0.79
14 Lonely	0.59	0.59	0.85	0.89
17 Cry	0.37	0.37	0.78	0.84
18 Sad	0.64	0.64	0.89	0.92
Somatic				
1 Bothered	0.45	0.45	0.72	0.75
2 Appetite	0.39	0.39	0.62	0.65
5 Mind	0.56	0.56	0.69	0.70
7 Effort	0.23	0.23	0.27	0.26
11 Sleep	0.49	0.49	0.64	0.65
13 Talk	0.45	0.45	0.65	0.68
20 Getgoing	0.53	0.53	0.77	0.80
Positive Affect				
4 Good	0.63	0.63	0.67	0.70
8 Hopeful	0.52	0.52	0.55	0.59
12 Happy	0.74	0.74	0.85	0.87
16 Enjoy	0.81	0.81	0.86	0.89
Interpersonal				
15 Unfriendly	0.47	0.47	0.72	0.75
19 Dislike	0.71	0.71	0.93	0.96
Factor correlations				
Dep Aff./Som	0.90	0.90	0.95	0.96
Dep Aff/Pos Aff	0.68	0.68	0.80	0.84
Dep Aff/Inter	0.74	0.74	0.83	0.85
Som/Pos Aff	0.60	0.60	0.72	0.74
Som/Inter	0.66	0.66	0.78	0.79
Pos Aff/Inter	0.51	0.51	0.65	0.68

The key findings from the comparisons are as follows:

The ML estimates of model fit, factor loadings and factor correlations produced from LISREL and M*plus* are virtually identical.

Model fit estimates, factor loadings and factor correlations from WLS estimation in LISREL and M*plus* are close but not quite identical. There are some minor differences in the fit statistics but factor loadings and correlations are all higher in M*plus*.

WLS estimates (whether from LISREL or M*plus*) appear more satisfactory than those using ML. The fit statistics indicate a better fitting model and factor loadings and correlations are consistently and substantially higher than those shown for ML.

Overall the results from the comparison of the ML and WLS estimation techniques implemented in LISREL and M*plus* suggest that the factor loadings and factor correlations derived from previous CES-D analyses using ML might be downwardly biased.

Summary

In this chapter the results from a series of CFAs and higher order factor analyses of the CES-D using the M*plus* software package have been presented. The key findings can be summarised as follows:

CFA showed that the traditional four factor CES-D model provided a better fit to the data than a one, two, three or five factor CES-D model.

Higher order factor analyses indicated that a general factor of 'depressive symptomatology' accounted for around one half of the variation in item scores. Around 8 per cent of variance in total is accounted for by the specific (Depressed Affect, Somatic, Positive Affect & Interpersonal) factors.

Using the nested factor model, error variation is estimated to be in the order of 30 per cent to 40 per cent. This indicates that a large proportion of the variance in CES-D scores is unaccounted for in the models hypothesised.

A series of normality tests indicated that the data did not satisfy the assumption of MVN required by ML estimation techniques. Using criteria proposed by the developer of LISREL the less restrictive assumption of bivariate normality required for a WLS approach could be supported.

On the basis of a comparison of ML and categorical variable estimation techniques it was found that the latter produced better fitting models and more accurate parameter estimates.

Overall, the higher order CFA results provide good support for the unidimensionality of the CES-D. This finding has important implications for the measurement model to be used as the basis of the measurement invariance analyses tests which are presented in the next chapter of this thesis.

8

SEM Measurement Invariance Analyses

The chapter presents the results from a series of multiple group confirmatory factor analyses (MG-CFA) which are performed to test CES-D measurement invariance across gender and year level. In order to assist the interpretation of the results, this chapter begins with a general introduction to measurement invariance testing with SEM and the various types of measurement invariance are outlined. Next, the technical aspects of defining and identifying a measurement model and the use of fit statistics for testing measurement invariance models are addressed. Using results derived from the IRT and SEM factor analyses presented earlier, as well as theoretical considerations, a one factor CES-D model is proposed as the measurement model and the form of this model is detailed.

The main sections of the present chapter comprise the results from the measurement invariance tests. Beginning with gender, seven types of measurement invariance tests are performed. The rationale for each type of measurement invariance test is described in detail. Following this, the impact of any lack of gender measurement invariance on latent means and on raw scores is examined and compared with estimates derived from the IRT analyses. Finally, a series of analyses is performed to test for measurement invariance across year levels. Exactly the same series of measurement invariance tests are performed across year level as were performed across gender but to avoid redundancy these results are presented in a slightly more abbreviated form.

General framework for testing measurement invariance using SEM

Earlier in this report, the statistical technique of MG-CFA for testing measurement invariance in psychological scales was introduced. The basic idea of a MG-CFA is the estimation of the same measurement model in two or more groups and then the

testing of the equality of estimates of particular parameters in the different groups. Because some types of parameters may show equality while other types of parameters may not, measurement invariance is said to exist at different levels. In this section these different levels of measurement invariance are explained and the analytical framework for the analyses presented.

Invariance in a MG-CFA can be conceptualised at the level of the measurement model and at the level of the structural model. What is known as the 'measurement model' refers to the links between the latent variables in a model and their observed measures while the 'structural model' is taken to reflect the links between the latent variables themselves (Byrne et al., 1989). In testing for measurement invariance it is the parameters associated with the measurement model that are of most interest, in particular item factor loadings and intercepts. These parameters are analogous to the item discrimination and item difficulty parameters (Ferrando, 1996) of the standard 2PL IRT model outlined in Chapter 6

The relevance of these SEM parameters (factor loadings and intercepts) can be seen from the standard MG-CFA that has been outlined in the literature in many places. For example, Steenkamp and Baumgartner (1998, p.79) define a standard MG-CFA as follows:

$$x^g = \tau^g + \Lambda^g \xi^g + \delta^g$$

where x^g is a p × 1 vector of observed variables (in group *g*)

τ^g is a p × 1 vector of item intercepts

Λ^g is a p × m matrix of factor loadings

ξ^g is an m × 1 vector of latent variables

δ^g is a p × 1 vector of errors of measurement

This model assumes *p* items, *m* latent variables, *g* for group membership and the same factor structure for each group. The equation shows that, provided the groups show the same factor structure (configural invariance), observed scores are a function of underlying factor scores but that observed scores may not be comparable across groups because of different intercepts (τ) and scale metrics or factor loadings (λ). The third parameter in the equation, the error term (δ) provides information on the differential precision of items across groups but it is not considered necessary that items be equally reliable for valid mean comparisons across groups (see Little, 1997 for discussion).

The standard MG-CFA defined above does not take into account the unique ability of M*plus* to incorporate a threshold structure into a model for categorical variables. For categorical variables the intercept term is replaced and is more correctly modelled by thresholds. The standard MG-CFA therefore is modified and the term τ^g is replaced by the term τ^g_{c-1} which now represents a p × 1 vector of item thresholds with *c* categories.

Table 39 presents a summary of the main types of measurement invariance. Each of these types of invariance are tested in the present study with exception of factor covariances. Factor covariances are not tested because a one factor CES-D model is specified as the measurement model. Further details clarifying each type of

measurement invariance are provided in the results section for each hypothesis tested. Prior to presenting these results however a measurement model needs to be proposed, a process for assessing model fit decided, and issues related to reference indicator selection and model identification dealt with.

Defining a CES-D measurement model and assessing fit

To test the different levels of measurement invariance outlined in Table 39 a CES-D factor model must be specified. This raises the somewhat vexing question as to which factor model to use. Earlier, it was shown that a general 'Depression' factor accounts for around one half of the variation in CES-D item scores with very little variation accounted for by the specific factors of Depressed Affect, Somatic, Positive Affect and Interpersonal.

Table 39 Types of measurement invariance

H#	Hypothesis Name	Symbolic Statement	Conceptual Meaning
Measurement model level			
0	Invariant covariance	$\sum^{g} = \sum^{g}$	An omnibus test of the equality of covariance matrices across groups
1	Configural invariance	$\Lambda^{g}_{(\,form\,)} = \Lambda^{g}_{(\,form\,)}$	A test of whether a model with the same pattern of fixed and free factor loadings for each group fits the data well in all groups
2	Metric invariance	$\Lambda^{g} = \Lambda^{g}$	A test of whether factor loadings across groups are equal
3	Scalar invariance	$\tau^{g}_{c-1} = \tau^{g}_{c-1}$	A test of whether the thresholds of items are equal across groups
4	Invariant uniquenesses	$\Theta^{g} = \Theta^{g}$	A test of whether the amount of measurement error is equal across groups
Structural model level			
5	Invariant factor variance	$\Phi^{g}_{j} = \Phi^{g}_{j}$	A test of whether factor variances are equal across groups
6	Invariant factor covariances	$\Phi^{g}_{jj} = \Phi^{g}_{jj}$	A test of whether factor covariances (where applicable) are equal across groups
7	Equal factor means	$\kappa^{g} = \kappa^{g}$	A test of whether the latent means are equal across groups

This finding provides empirical support for using a one factor model as the CES-D measurement model. In addition, the one factor model (see Table 27) was found to provide acceptable fit to the data for both boys and girls. Conceptually it can be noted that this model is implicitly assumed by most users of the CES-D who sum the CES-D to form a total score as recommended by Radloff (1977) and consistent with the view that CES-D individual scores represent a single trait of depressive

symptomatology. For these reasons the tests of measurement invariance are based on the CES-D one factor model.

For M*plus* categorical variable models the range of fit statistics produced is similar, but not identical, for single and multiple group analyses. Four fit statistics for multiple group categorical models are presented in the output: the chi-square test of model fit, the CFI, the TLI and the RMSEA. The SRMR, which is shown for single group categorical models, is not calculated for multiple group models.

Fit statistics in a MG-CFA are used to evaluate whether across group (gender or year level) restrictions in multiple group analyses are necessary for valid latent mean comparisons. Little (1997) describes two main approaches for this task. The first, and more common strategy, is based on a statistical rationale while the second approach is based around a modelling rationale.

The statistical rationale involves conducting equivalence tests using the chi-square statistic. In this method the chi-square difference between two models, one with restrictions (e.g. equal factor loadings) and one without, is tested with their degrees of freedom equal to the difference in their degrees of freedom. If the test is significant this provides evidence that the measurement parameters tested are different across the groups.

A draw back of the statistical approach is that the chi-square is a very sensitive measure of model fit and (particularly in a large sample) may incorrectly lead to the rejection of models with restrictions on the basis of trivial differences. This difficulty might be particularly troublesome when comparing the fit of identical models across groups that differ in sample size (Marsh & Byrne, 1993).

The modelling rationale uses the other measures of fit, in this study RMSEA, CFI and TLI, which are less influenced by sample size and have been shown to be useful in fit evaluation (Bentler, 1990; McDonald & Marsh, 1990). The advantage of this approach is that various sources of mis-fit (random or systematic) can be taken to be substantively trivial provided either overall model fit remains acceptable or the relative change to the fit statistic is minor. A draw back of the modelling approach is that precise criteria for using fit statistics for this purpose have not been established.

In the only simulation study for continuous variable models of this issue, Cheung and Rensvold (1999) conclude that a change in CFI of -0.1 or less indicates that a invariance hypothesis should not be rejected, a change between greater than -0.1 and -0.2 indicates a suspicious difference and a change greater than –0.2 indicates a definite difference between models. Vandenberg and Lance (2000) recommend further study before these guidelines are accepted and it is not clear whether they would be appropriate to the categorical variable models estimated in the present study.

There are further complexities inherent in the modelling approach to testing invariance hypotheses with SEM fit statistics. Marsh and Byrne (1993) for example explain how it is possible for the TLI fit statistic actually to increase (show a better fitting model) with the introduction of invariance constraints. For example, a model with factor loadings constrained to be equal across groups might appear to fit better than a model in which factor loadings are allowed to vary across groups. This can occur because the introduction of the constraints leads to a lower chi-square to degrees of freedom ratio.

In the present study a blend of both statistical testing and modelling rationales is used. A large number of equivalence tests using the chi-square statistic are performed to test differences between models. Where multiple testing within a model is

performed, significance levels are treated by the Bonferroni correction (see Bollen, 1989, p. 369 for discussion). In addition, the alternative fit statistics are examined for changes to overall model fit. Individual tests for the practical significance of any lack of measurement invariance in terms of impact of latent means are also performed.

Identification issues

The identification of measurement invariance models raises both methodological and substantive issues. Both issues are complex and doubly so for categorical variable models with threshold structures. From a methodological perspective it is not clear what constraints are necessary for identification and at a practical level different constraints are required in different software packages. As Millsap (2001, ¶ 3) notes:

The issue of thresholds in multigroup ordinal CFA is poorly dealt with in the literature. I have yet to find a fully complete discussion, with proofs of necessary and sufficient conditions for identification on these parameters. LISREL's approach is to fix thresholds (all of them) across groups to common values. This is not a necessary condition for identification in polytomous cases. Mplus uses a different identification process, and leads to different estimates. One must introduce SOME constraints on either the thresholds, or on the factor means/intercepts. LISREL chooses the former, but goes farther than is strictly necessary. For example, if you have a five-point item, you need not require all four thresholds to be invariant across groups to achieve identification.

The key to the identification issue lies in correctly specifying the configural model. This model is, in a sense, the most important model because all additional measurement invariance tests build upon it (Vandenberg & Lance, 2000). What is required is a model with the fewest constraints possible and in a form which will allow the effects of additional constraints of equal factor loadings and thresholds to be examined. According to Millsap (personal communication) the minimum constraints needed for identification in the case of a one factor CES-D configural invariance model are as follows:

(a) fixing one item (known as the marker item) to have the same loading in both groups;

(b) for each item, fixing one threshold to invariance;

(c) for the marker item fixing a second threshold to invariance;

(d) setting the factor mean in the reference group (Boys) to zero and free in the other group (Girls); and

(e) setting the scaling factors to one in the reference group and free in the other group.

Essentially this approach is followed in constructing the configural invariance model with the additional constraint that the factor variance is set to one in the reference group (Boys) and free to be estimated in the other group (Girls). This additional constraint is introduced to allow factor scores to be expressed in a *Z* score metric.

The substantive identification issue concerns the selection of the factor loadings and thresholds to be set to invariance. These choices are substantively important because if the factor loadings or thresholds chosen to be fixed (usually to unity) and equal across groups are in reality not invariant then this may cause other factor loadings or thresholds to appear not to be invariant when they are. Ideally, during test

development items would be included which are known to be invariant along with the new items but for intact, existing tests, a decision is required.

In the analyses Item 6 (*Depress*) is used as the marker item and, in all models estimated, the factor loading for this item is set to unity across groups (gender and year level). This item was chosen as the reference indicator for a number of reasons. The first reason is based on the clear centrality of the dysphoria symptom to depression (Suh & Gallo, 1997). Second IRT analyses in the present study (see Chapter 6) showed that this item is very effective for both boys and girls and that this item exhibits minimum DIF across gender. These results are consistent with previous IRT CES-D analyses in also showing a strong relation between Item 6 (Depress) to the latent construct. Finally it can be noted this choice is consistent with that of Breithaupt and Zumbo (2002) who appear to be the only previous SEM CES-D researchers to have carefully considered this issue.

The choice of which threshold for each item to set to be equal across groups is less clear cut. Unfortunately because of the binary scoring system used by Breithaupt and Zumbo (2002) this problem was not faced these researchers and so little guidance is available in the literature. From the earlier descriptive and IRT analyses of the present study it did appear that the first threshold might be the most biased (see Table 21). This was because the key difference between boys and girls appears less to be in the severity rating of a symptom but rather because more girls than boys acknowledge the presence of a symptom. The difference, and most likely source of invariance, therefore will be most pronounced in the first threshold.

Following this logic the third threshold for each item (with one exception), representing the change from Option 2 to Option 3 are chosen as the marker thresholds. The one exception is for Item 7 (*Effort*). The IRT analyses for this item (see Table 25) showed that for equivalent levels of depressive symptomatology boys are more likely to endorse Option 3 than girls. The gender differences for Option 1 and Option 2 on the other hand are very minor and therefore for Item 7 (*Effort*) the second threshold is chosen as the marker threshold.

In the following sections, using a one factor CES-D measurement model and with the identification constraints outlined above, each of the types of measurement invariance shown in Table 39 are tested. The first series of tests are performed across gender and are explained in some detail. Following this, an examination of the impact of any lack of measurement invariance on latent means is performed. Finally, the same series of measurement invariance tests are performed across year levels (Year 8 to Year 10) and to avoid redundancy these are described in slightly less detail.

Gender invariant covariance (Hypothesis 0: $\sum^{g} = \sum^{g}$)

The first hypothesis tested is whether boys and girls have equivalent population covariance matrices for the observed variables (Jöreskog, 1971; Schaie & Hertzog, 1985). Strictly speaking, in the present study, polychoric correlation matrixes are tested, but for consistency with the literature the term ‘covariance matrices’ is used. If covariance matrices do not differ across groups then measurement invariance is established and further tests of measurement equivalence are not required (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). On the other hand, if the groups’ covariance matrices do differ then additional tests to establish the cause of the lack of equivalence can be performed (Schaie & Hertzog, 1985).

In order to test the equality of covariance matrices across groups, a M*plus* multiple group analysis is performed. The model specified for this analysis, M0, does not comprise any factors but simply a series of 'correlated errors'. That is, the error term of Item 1 is correlated with the error term of Item 2, 3, 4 … 20, the error term of Item 2 is correlated with the error term of Item 3, 4, 5 … 20 and so on. This specification in M*plus*, recommended by Muthén (2001a), estimates the covariance matrix in the Theta parameter matrix. The M*plus* syntax used to estimate this model is provided in Appendix C (Program 3.4).

The fit statistics for M0 are: χ^2 = 925.78, df = 40, $p < 0.001$, CFI = 0.981, TLI = 0.817, RMSEA = 0.081. The significant chi-square value for this model, the relatively poor TLI statistic (well fitting models should produce a TLI statistic ą 0.90) and marginal RMSEA statistic (ideally ż 0.05) indicate that the hypothesis of equal covariance matrices across boys and girls can be rejected. Further investigation into the source of the lack of fit is therefore warranted.

Gender configural invariance (Hypothesis 1: $\Lambda^{g}{}_{(\,form\,)}$ = $\Lambda^{g}{}_{(\,form\,)}$)

The next step is to establish configural invariance by testing whether the proposed single factor measurement model (see Figure 7) applies to the data in each group. This model, the configural gender invariance model (M1), is important because it will test whether boys and girls are employing the same conceptual frame of reference to the construct (depressive symptomatology) hypothesised to underlie the observed variables (CES-D items). If boys and girls are not employing the same frame of reference, then further tests of measurement invariance are not justified and it will not be possible to test for differences in the structural parameters (e.g. latent means) of the model.

According to Steenkamp and Baumgartner (1998) configural invariance is supported if the model fits the data well across groups and factor loadings are significantly and substantially different from zero. If configural invariance is supported then further tests, nested within the test of configural invariance can proceed. In this sense, additional measurement invariance tests build upon the configural model (Vandenberg & Lance, 2000). The configural invariance model (M1) is specified (see Appendix C, Program 3.5 for syntax) as follows:

> The factor loadings for the marker item (*Depress* (6)) are constrained to be equal for boys and girls.
>
> The factor loadings for the remaining 19 items are set free to be estimated separately for boys and girls.
>
> The third threshold for each item (with one exception) is constrained to be equal for boys and girls.
>
> For Item 7 (*Effort*) the second threshold (as opposed to the third) is constrained to be equal for boys and girls.
>
> For the marker item (*Depress* (6)) the second threshold is constrained to be equal for boys and girls.
>
> The remaining thresholds are set free to be estimated separately for boys and girls.

The factor mean for boys is set to zero and the factor mean for girls is set free to be estimated.

The factor variance for boys is set to one and set free to be estimated for girls.

The scale factors for boys are set to one and set free for girls.

The fit of the configural gender invariance model (M1) is satisfactory. The chi-square is significant ($\chi^2 = 3540.32$, df = 340, $p < 0.001$) but the other indices indicate relatively good fit (CFI = 0.931, TLI = 0.922, RMSEA = 0.053). The factor loadings are all significant and for both boys and girls are above 0.50 with the exception of Item 7 (*Effort*) which is 0.14 (Boys) and 0.17 (Girls). Taken together the acceptable fit statistics and high factor loadings indicate that a one factor CES-D model provides for configural invariance across gender.

The latent mean for the girl sample is 0.311 with a standard error of 0.09. The variance of the latent mean in the girl sample is 1.12 with a standard error of 0.19. The form of M1, the pattern of fixed and free parameters, the factor loadings and thresholds are shown in Table 40.

Table 40 Factor loading and thresholds from the gender configural invariance model (M1)

		Boys				Girls			
		Factor Loading	Threshold 1	2	3	Factor Loading	Threshold 1	2	3
1	Bothered	0.73	0.66	1.34	*1.74	0.66	0.46	1.18	*1.74
2	Appetite	0.55	0.84	1.48	*2.02	0.73	0.52	1.34	*2.02
3	Blues	0.88	0.73	1.24	*1.72	0.91	0.67	1.23	*1.72
4	Good	0.65	-0.10	0.60	*1.13	0.69	-0.11	0.57	*1.13
5	Mind	0.65	-0.23	0.68	*1.39	0.68	-0.08	0.76	*1.39
6	Depress (M)	*0.91	0.42	1.03	*1.49	*0.91	0.42	1.03	*1.49
7	Effort	0.14	-0.26	*0.38	1.11	0.17	0.01	*0.38	0.71
8	Hopeful	0.52	-0.47	0.43	*1.08	0.51	-0.36	0.43	*1.08
9	Failure	0.90	0.85	1.30	*1.77	0.88	0.93	1.39	*1.77
10	Fearful	0.79	0.81	1.40	*2.03	0.79	0.91	1.54	*2.03
11	Sleep	0.59	0.35	1.01	*1.58	0.64	0.25	1.01	*1.58
12	Happy	0.85	-0.01	0.86	*1.47	0.77	0.27	0.89	*1.47
13	Talk	0.64	0.31	1.11	*1.84	0.66	0.39	1.20	*1.84
14	Lonely	0.90	0.62	1.20	*1.67	0.87	0.65	1.22	*1.67
15	Unfriendly	0.79	0.44	1.26	*1.87	0.66	0.67	1.36	*1.87
16	Enjoy	0.87	0.02	0.79	*1.33	0.81	0.21	0.84	*1.33
17	Cry	0.75	1.27	1.73	*2.14	0.90	0.90	1.57	*2.14
18	Sad	0.93	0.60	1.22	*1.70	0.89	0.48	1.16	*1.70
19	Dislike	0.91	0.45	1.23	*1.70	0.91	0.49	1.27	*1.70
20	Getgoing	0.76	0.30	1.11	*1.66	0.77	0.36	1.13	*1.66

* denotes fixed parameter; M denotes marker item

Gender metric invariance (Hypothesis 2: $\Lambda^g = \Lambda^g$)

The next hypothesis examines whether factor loadings are the same across groups. This is a stronger test of invariance than performed in M1 because it requires not only that the factor pattern be equivalent across groups, but also that there be equality of scaling units across groups. This level of invariance has received the most consideration in the literature (see Schaire et al., 1998 for a summary) and is sometimes termed 'weak factorial invariance'. As Vandenberg and Lance (2000, p. 37) explain:

> Factor loadings are the regression slopes relating the X_{jk} to their corresponding variables, ξ (Bollen. 1989), and thus represent the expected change in the observed score on the item per unit change on the latent variable. Thus the test of the null hypothesis that $\Lambda_x^g = \Lambda_x^g$ is a test of equality of scaling units across groups (Jöreskog, 1969; Schmitt, 1982; Vandenberg & Self, 1993).

To test whether boys and girls are responding to items in the same way, the matrix of factor loadings is constrained to be invariant across gender. In all other respects this model, termed the 'full metric invariance model' (M2) is identical to the model estimated to test configural invariance (M1). The full metric invariance model (M2: $\chi^2 = 3655.04$, df = 359, $p < 0.001$) showed a deterioration in fit compared with the configural invariance model.

The change in the chi-square statistic is significant ($\Delta\chi^2 = 114.72$, df = 19, $p < 0.001$) and one of the alternative fit indices shows a deterioration in fit (M1:CFI = 0.931, M2:CFI = 0.928, ΔCFI = -0.003). On the other hand the TLI fit index indicates improved fit (M1:TLI = 0.922, M2: TLI = 0.924, ΔTLI = +0.002) as did the RMSEA statistic (M1:RMSEA = 0.053, M2: RMSEA = 0.052, ΔRMSEA = -0.001). On the basis of the relatively large chi-square statistic it is concluded that the hypothesis of full metric invariance is not supported.

Although the hypothesis of full metric invariance is not supported it is still possible to validly test structural parameters (e.g. latent means) of a model (Byrne et al., 1989) This is achieved under what is termed as 'partial invariance'. Partial metric invariance refers to when some factor loadings are different across groups while others are invariant. According to Byrne et al. (1989) as long as one item other than the marker item shows invariance then further tests of invariance and substantive analyses (comparison of latent means) can be meaningful. A similar logic can be applied at the level of scalar invariance where hypotheses of either full and partial scalar invariance can be tested.

In order to investigate whether partial metric invariance can be supported across gender, a further 19 models are estimated setting each factor loading free individually. This procedure creates the problem of testing multiple hypotheses, and so using the Bonferroni correction, the probability level is set to 0.0026 (0.05/19), which for one degree of freedom corresponds to a critical chi-square value of 9.55. Using this critical chi-square value as a guide to statistical significance, four items: *Bothered* (1), *Appetite* (2), *Unfriendly* (15) and *Cry* (17) exhibit a lack of metric invariance. These results are presented in Table 41 with the chi-square values greater than the critical value (9.55) shown in bold.

Table 41 Change in chi-square value from setting factor loadings free

		χ^2	$\Delta\chi^2$	ΔCFI	ΔTLI	ΔRMSEA
1	Bothered	3641.19	**13.85**	0.001	0.000	0.000
2	Appetite	3628.45	**26.59**	0.001	0.001	0.000
3	Blues	3652.52	2.52	0.000	0.000	0.000
4	Good	3647.06	7.98	0.001	0.000	0.000
5	Mind	3654.00	1.04	0.000	0.000	0.000
6	Depress (M)					
7	Effort	3653.70	1.34	0.000	0.000	0.000
8	Hopeful	3654.99	0.05	0.000	0.000	0.000
9	Failure	3653.22	1.82	0.000	0.000	0.000
10	Fearful	3654.54	0.50	0.000	0.000	0.000
11	Sleep	3654.47	0.57	0.000	0.000	0.000
12	Happy	3649.20	5.84	0.001	0.000	0.000
13	Talk	3654.92	0.12	0.000	0.000	0.000
14	Lonely	3655.04	0.00	0.000	0.000	0.000
15	Unfriendly	3629.95	**25.09**	0.001	0.001	0.000
16	Enjoy	3652.73	2.31	0.000	0.000	0.000
17	Cry	3633.71	**21.33**	0.001	0.000	0.000
18	Sad	3649.66	5.38	0.001	0.000	0.000
19	Dislike	3651.07	3.97	0.000	0.000	0.000
20	Getgoing	3654.96	0.08	0.000	0.000	0.000

M denotes marker item

To test whether partial metric invariance could be established a model, termed the partial metric invariance model (M3), is estimated with the factor loadings of the items: *Bothered* (1), *Appetite* (2), *Unfriendly* (15) and *Cry* (17) set free (see Appendix C, Program 3.6 for syntax). All other factor loadings are constrained to be equal across boys and girls. The fit statistics for this model are acceptable ($\chi^2 = 3571.69$, df = 355, $p < 0.001$, CFI = 0.930, TLI = 0.925, RMSEA = 0.052).

The change in the chi-square statistic from the configural model (M1) in which all factor loadings were allowed to be free although statistically significant ($\Delta\chi^2 = 31.37$, df = 15, $p = 0.008$) is relatively small. In addition, compared with M1, the CFI (Δ = -0.001) shows only modest deterioration in fit while the TLI (Δ = +0.003) and RMSEA (Δ = -0.001) show modest improvement. On this basis it is concluded that the hypothesis of partial metric invariance is supported.

Gender scalar invariance (Hypothesis 3: $\tau^g_{c-1} = \tau^g_{c-1}$)

Hypothesis 3 tests whether the thresholds of items are invariant across boys and girls. By convention it is only appropriate to perform scalar tests on items demonstrating metric invariance (see Bollen, 1989, p. 366). The reason for this is when factor loadings (slopes) differ it is highly unlikely that intercept invariance is possible (Millsap, 1995). Consequently, the four items failing to show metric invariance, namely: *Bothered* (1), *Appetite* (2), *Unfriendly* (15) and *Cry* (17) are not tested for scalar invariance. An initial scalar model (M4) is estimated as follows.

> Factor loadings for the marker item, Item 6 (*Depress*), are constrained to be equal for boys and girls.
>
> Factor loadings for the items: *Bothered* (1), *Appetite* (2), *Unfriendly* (15) and *Cry* (17) are set free and the factor loadings for the remaining 15 items are held constant across groups.
>
> Thresholds are fixed for the purposes of identification as in the earlier analyses. This is the third threshold of each item is constrained, except for Item 7 (*Effort*) where the second threshold is used and the second threshold of the marker item is constrained.
>
> The thresholds (not used for identification) for the four items failing metric invariance are set free. The remaining thresholds for the metrically invariant items are constrained to be equal across groups.
>
> The factor mean for boys is set to zero and the factor mean for girls is set to be free.
>
> The factor variance for boys is set to one and free for girls.
>
> The scale factors for boys are set to one and set free for girls.

In summary, the key difference between the initial scalar model (M4) and the partial metric model (M3) is that the thresholds of the 15 items showing metric invariance are constrained to be equal across boys and girls.

The fit statistics for M4 are acceptable (χ^2 = 3992.60, df = 386, $p < 0.001$, CFI = 0.922, TLI = 0.923, RMSEA = 0.053) but the change in chi-square statistic from the partial metric model (M3) in which all thresholds (not required for identification purposes) were allowed to be free is large ($\Delta\chi^2$ = 420.91, df = 31, $p < 0.001$). In addition, the alternative fit indices show a significant deterioration in fit (ΔCFI = -0.008, ΔTLI = -0.002, ΔRMSEA = +0.001). On this basis it is concluded that the hypothesis of full scalar invariance is not supported.

In order to investigate the source of the deterioration of fit from constraining the thresholds of the invariant items, a further 31 models (15 items by 2 thresholds and the first threshold of the marker item) are estimated freeing each threshold individually. This procedure creates the problem of testing multiple hypotheses and the alpha level was treated by the Bonferroni correction. In this instance 31 threshold invariance chi-square difference tests are performed and so the critical probability level is set to 0.0016 (0.05/31), which for one degree of freedom, corresponds to a critical chi-square value of 9.97. The results from the series of analyses freeing each threshold individually are presented in Table 42.

Quite clearly the first threshold for six items: *Blues* (3), *Good* (4), *Effort* (7), *Sleep* (11), *Happy* (12) and *Sad* (18) and the third threshold for Item 7 (*Effort*) are not invariant across gender. A model (M5) relaxing these seven constraints (with the

remainder fixed across gender) produced a marked improvement to fit (χ^2 = 3670.30, df = 379, $p < 0.001$, CFI = 0.929, TLI = 0.928, RMSEA = 0.051) compared to M4 in which the thresholds of all 15 items showing metric invariance were constrained to be equal across boys and girls.

Table 42 Change in chi-square value from constraining thresholds (Gender)

		Threshold 1 $\Delta\chi^2$	Threshold 2 $\Delta\chi^2$	Threshold 3 $\Delta\chi^2$
1	Bothered			
2	Appetite			
3	Blues	**18.83**	0.14	
4	Good	**19.23**	3.57	
5	Mind	3.83	1.45	
6	Depress (M)	4.71		
7	Effort	**37.53**		**23.79**
8	Hopeful	1.58	6.19	
9	Failure	1.60	3.14	
10	Fearful	4.49	5.53	
11	Sleep	**44.69**	0.02	
12	Happy	**82.59**	0.08	
13	Talk	1.06	5.12	
14	Lonely	0.18	0.12	
15	Unfriendly			
16	Enjoy	2.03	0.04	
17	Cry			
18	Sad	**38.16**	2.12	
19	Dislike	0.07	0.07	
20	Getgoing	0.53	0.08	

The change in chi-square statistic from M5 compared with the partial metric invariance model (M3) remains statistically significant ($\Delta\chi^2$ = 98.61, df = 24, $p < 0.01$) although not large relative to the degrees of freedom. The deterioration in the CFI fit statistic is small (ΔCFI = -0.001) while other fit statistics showed improved fit (ΔTLI = +0.003, ΔRMSEA = -0.001). Given that no other threshold parameter stands out on the basis of being able to produce a highly significant improvement to fit, this model (M5) is accepted as the final scalar invariance model.

The final scalar invariance model (M5) forms the basis for testing the structural parameters (factor variances and factor means) of the measurement model as well item residual variances. M5 incorporates the findings that four items: *Bothered* (1), *Appetite* (2), *Unfriendly* (15) and *Cry* (17) fail to show metric invariance across gender and the first threshold for six items: *Blues* (3), *Good* (4), *Effort* (7), *Sleep* (11), *Happy* (12) and *Sad* (18) and the third threshold for Item 7 (*Effort*) fail to show scalar invariance across gender (see Appendix C, Program 3.7 for syntax). The form of M5,

the pattern of fixed and free parameters, factor loading estimates and thresholds are shown in Table 43.

Table 43 Item factor loadings and thresholds from the final scalar model (M5: Gender)

		Boys				Girls			
		Factor Loading	Threshold 1	2	3	Factor Loading	Threshold 1	2	3
1	Bothered	*0.73	*0.65	*1.33	1.73	*0.64	*0.39	*1.14	1.73
2	Appetite	*0.56	*0.83	*1.47	2.00	*0.71	*0.45	*1.30	2.00
3	Blues	0.88	*0.71	1.19	1.68	0.88	*0.60	1.19	1.68
4	Good	0.66	*-0.11	0.55	1.10	0.66	*-0.20	0.55	1.10
5	Mind	0.66	-0.21	0.69	1.38	0.66	-0.21	0.69	1.38
6	Depress (M)	0.91	0.37	0.99	1.47	0.91	0.37	0.99	1.47
7	Effort	0.16	*-0.27	0.36	*1.09	0.16	*-0.01	0.36	*0.69
8	Hopeful	0.52	-0.46	0.41	1.09	0.52	-0.46	0.41	1.09
9	Failure	0.89	0.86	1.33	1.76	0.89	0.86	1.33	1.76
10	Fearful	0.79	0.83	1.44	2.01	0.79	0.83	1.44	2.01
11	Sleep	0.60	*0.34	0.98	1.55	0.60	*0.18	0.98	1.55
12	Happy	0.84	*0.00	0.88	1.52	0.84	*0.18	0.88	1.52
13	Talk	0.64	0.31	1.12	1.81	0.64	0.31	1.12	1.81
14	Lonely	0.90	0.60	1.20	1.67	0.90	0.60	1.20	1.67
15	Unfriendly	*0.79	*0.44	*1.26	1.86	*0.65	*0.62	*1.34	1.86
16	Enjoy	0.87	0.06	0.81	1.37	0.87	0.06	0.81	1.37
17	Cry	*0.75	*1.27	*1.73	2.15	*0.90	*0.85	*1.54	2.15
18	Sad	0.93	*0.57	1.18	1.72	0.93	*0.43	1.18	1.72
19	Dislike	0.91	0.43	1.23	1.69	0.91	0.43	1.23	1.69
20	Getgoing	0.76	0.29	1.10	1.65	0.76	0.29	1.10	1.65

* denotes free parameter; M denotes marker item

Table 43 shows that the factor loadings for Item 2 (*Appetite*) (Boys: 0.56; Girls: 0.71) and Item 17 *(Cry*) (Boys: 0.75; Girls: 0.90), are higher for girls than boys. Factor loadings for Item 1 (*Bothered*) (Boys: 0.73; Girls: 0.64) and Item 15 (*Unfriendly*) (Boys: 0.79; Girls: 0.65) on the other hand are higher for boys than girls. For these four items which fail to show metric invariance their thresholds are not constrained to be equal across gender. The meaning of threshold differences for items failing to show metric invariance is unclear.

Table 43 also shows that the first threshold for six other items: *Blues* (3), *Good* (4), *Effort* (7), *Sleep* (11), *Happy* (12) and *Sad* (18) and the third threshold for Item 7 (*Effort*) are allowed to vary across gender. These items demonstrated metric invariance (equal factor loadings) but differences in their thresholds. With respect to

the first threshold differences, most are larger for boys than girls: *Blues* (Boys: 0.71; Girls: 0.60)*, Good* (Boys: -0.11; Girls: -0.20)*, Sleep* (Boys: 0.34; Girls: 0.18) and *Sad* (Boys: 0.57; Girls: 0.43).

Larger thresholds from items showing metric invariance indicate that that particular response option is less likely to be endorsed by members of one group compared with the other. Differences in the first threshold (Option 0 – Option 1) reflect group differences in acknowledging the presence of a symptom. The results therefore indicate that girls are more likely than boys to note the presence of the symptoms: *Blues* (3), *Good* (4), *Sleep* (11) and *Sad* (18).

Two of the first thresholds are larger for girls compared with boys: *Effort* (Boys: -0.27; Girls: -0.01) and *Happy* (Boys: 0.00; Girls: 0.18). This indicates that boys are more likely than girls to acknowledge the presence of these symptoms. With respect to the third threshold for Item 7 (*Effort*) (Boys: 1.09; Girls: 0.69) the larger value for boys indicates that they are less likely to report experiencing this symptom 'most or all of the time' than are girls.

The final scalar invariance model (M5) is used to test whether item error variances, factor variances and latent means are equal across boys and girls. Model 5 it will be recalled is estimated with item residual variances free in both groups, the factor variance in the boys' group set at one and free in the girls' group and finally the factor mean is set to zero in the boys' group and free in the girls' group.

Gender invariant uniquenesses (Hypothesis 4: $\Theta^g = \Theta^g$)

A model (M6) identical to M5 is estimated but with the added constraint that item residual variances are equal for boys and girls (see Appendix C, Program 3.8 for syntax). This is achieved by setting the item scale factors to one in both groups and the factor variances in both groups to one. Factor variances in both groups are set to one because it is only appropriate to test invariant uniqueness if factor variances are invariant across groups (Cole & Maxwell, 1985).

Factor variances are set to unity on the basis that this assumption (which is tested in the next section) is met. In addition, only the residual variances for items demonstrating metric invariance are tested (i.e. the items *Bothered* (1), *Appetite* (2), *Unfriendly* (15) & *Cry* (17) are not tested). This is because it is only when items are metrically invariant and factor variances are invariant can items be equally reliable across group (Raju et al., 2002; Steenkamp & Baumgartner, 1998).

The fit statistics for M6 are (χ^2 = 4114.13, df = 396, $p < 0.001$, CFI = 0.919, TLI = 0.923, RMSEA = 0.053). The change in chi-square statistic from a model (to be presented next) in which factor variances are also required to be equal (M7) is significant ($\Delta\chi^2$ = 405.07, df = 16, $p < 0.001$) and all three alternative fit indices show a deterioration in fit (ΔCFI = -0.010, ΔTLI = -0.005, ΔRMSEA = +0.002). On this basis it is concluded that the hypothesis of equal item residual variances is not supported.

To investigate the source of the deterioration of fit from constraining item residual variances a further 16 models, for the 16 metrically invariance items, are estimated freeing each item residual variance individually. Using the Bonferroni correction the critical probability level is set to 0.0031 (0.05/16) which for one degree of freedom corresponds to a critical chi-square value of 8.75.

Table 44 shows the results from the series of models estimated to test whether item residual variances are equal across boys and girls. Given that only metrically invariant items are examined, only 16 tests are performed. Six items: *Good* (4), *Depress* (6), *Effort* (7), *Hopeful* (8), *Happy* (12) and *Getgoing* (20) produce a statistically significant deterioration in model fit when item residual variances are held equal across boys and girls. For all six of these items when the residual variances are not constrained, boys show higher values, indicating poorer reliability, than girls.

Table 44 Change in chi-square value from constraining item residual variances (Gender)

	Item residual variances free		Constrained	$\Delta\chi^2$
	Boys	Girls		from constraining
1 Bothered	0.46	0.38		
2 Appetite	0.67	0.86		
3 Blues	0.20	0.24	0.22	2.86
4 Good	0.55	0.34	0.49	**53.05**
5 Mind	0.54	0.48	0.53	4.98
6 Depress (M)	0.15	0.11	0.13	**12.73**
7 Effort	0.97	0.19	0.94	**213.93**
8 Hopeful	0.71	0.50	0.68	**56.02**
9 Failure	0.18	0.16	0.17	0.37
10 Fearful	0.35	0.36	0.36	0.06
11 Sleep	0.62	0.59	0.61	0.78
12 Happy	0.27	0.17	0.23	**49.91**
13 Talk	0.57	0.49	0.55	6.44
14 Lonely	0.17	0.17	0.17	0.00
15 Unfriendly	0.37	0.34		
16 Enjoy	0.22	0.18	0.21	6.98
17 Cry	0.43	0.35		
18 Sad	0.13	0.12	0.12	0.13
19 Dislike	0.15	0.18	0.16	2.97
20 Getgoing	0.40	0.27	0.35	**32.28**

Gender invariant factor variances (Hypothesis 5: $\Phi_j^g = \Phi_j^g$)

A model (M7) identical to M5 is estimated but with the constraint that factor variances are equal for boys and girls (see Appendix C, Program 3.9 for syntax). The fit statistics for M7 are: ($\chi^2 = 3709.06$, df = 380, $p < 0.001$, CFI = 0.928, TLI = 0.928, RMSEA = 0.051) and the change in chi-square statistic from M5 in which factor

variances are not required to be equal is small but statistically significant ($\Delta\chi^2$ = 38.76, df = 1, $p < 0.001$). Only one of the alternative fit indices shows a deterioration in fit (ΔCFI = -0.001) while the other fit statistics remain unchanged. On the basis of the relatively minor deterioration in model fit it is concluded that the hypothesis of equal factor variances could not be rejected.

Gender invariant factor covariances (Hypothesis 6: $\Phi_{jj}^{g} = \Phi_{jj}^{g}$)

An invariance test of factor covariances is not applicable because only one factor is specified in the measurement model.

Gender equal factor means (Hypothesis 7: $\kappa^{g} = \kappa^{g}$)

A model (M8) identical to M5 is estimated but with the added constraint that the factor means are equal for boys and girls (see Appendix C, Program 3.10 for syntax). The fit statistics for M8 are: (χ^2 = 3723.78, df = 380, $p < 0.001$, CFI = 0.927, TLI = 0.927, RMSEA = 0.051) and the change in chi-square statistic from M5 in which latent means are not required to be equal is significant ($\Delta\chi^2$ = 53.48, df = 1, $p <$ 0.001). Two of the alternative fit indices also show a deterioration in fit (ΔCFI = -0.002, ΔTLI = -0.001). On this basis it is concluded that the hypothesis of equal latent means is rejected.

Summary of gender measurement models

In this section a brief review of the gender invariance tests performed to this point is provided. The fit statistics for the gender measurement models (M0-M8) are summarised in Table 45.

Table 45 Gender model fit statistics

	χ^2	df	CLI	TLI	RMSEA
M0 Invariant covariance	925.78	40	0.981	0.817	0.081
M1 Configural invariance	3540.32	340	0.931	0.922	0.053
M2 Full metric invariance	3655.04	359	0.928	0.924	0.052
M3 Partial metric invariance	3571.69	355	0.930	0.925	0.052
M4 Initial scalar invariance	3992.60	386	0.922	0.923	0.053
M5 Final partial scalar invariance	3670.30	379	0.929	0.928	0.051
M6 M5 with constrained item variances	4114.13	396	0.919	0.923	0.053
M7 M5 with constrained factor variances	3709.06	380	0.928	0.928	0.051
M8 M5 with constrained means	3723.78	380	0.927	0.927	0.051

The first model (M0) examined the equivalence of boys' and girls' covariance matrices. The null hypothesis, that these covariance matrices are equal, was rejected indicating that further investigation into the source of the differences in covariance matrices was warranted.

The second model (M1) sought to establish configural invariance for the CES-D. This level of invariance tests whether boys and girls employ the same conceptual frame of reference to the construct (depressive symptomatology) hypothesised to underlie CES-D items. Using the one factor CES-D measurement model, configural invariance was supported by a well fitting model across groups.

The third model (M2) tested whether factor loadings (metric invariance) were the same across boys and girls. The null hypothesis that factor loadings were equal across boys and girls was rejected and four items: *Bothered* (1), *Appetite* (2), *Unfriendly* (15) and *Cry* (17) were found vary across gender. A model (M3) with the factor loadings of these items allowed to vary between boys and girls fitted the data nearly as well as the configural invariance model (M1). On this basis it was concluded that partial gender metric invariance had been demonstrated.

Next, the thresholds (scalar invariance) of the 16 items showing equal factor loadings were constrained to be equal across groups. M4, the initial scalar invariance model, showed a significant deterioration in fit compared with M3 in which thresholds were not constrained to be equal. Following a series of individual tests, it was concluded that seven thresholds were primarily responsible for the loss of fit. A final scalar invariance model (M5) was estimated allowing these seven thresholds to vary across groups.

With M5 as the final measurement model, the equivalence of item residual variances (M6), factor variances (M7) and factor means (M8) across groups were tested. The results from these analyses showed that several items did not have equal residual variances, the hypothesis of equal factor variances could not be rejected and finally the hypothesis that boys and girls have equal factor means was able to be rejected.

The impact of the lack of gender measurement invariance

In this section a series of models are estimated to examine the impact on latent means from the failure of the CES-D to exhibit full gender measurement invariance. The results are presented in Table 46.

The first model (M9) assumes full metric and scalar invariance by constraining factor loadings and thresholds to be equal for boys and girls. The latent mean estimate produced for this model for girls is 0.225 with a standard error of 0.024.

Next, in model M10 the factor loadings for the four items failing to exhibit metric invariance are set free. This has the effect of increasing the girl latent mean to 0.270. M11 sets the factor loadings and thresholds for the four items failing metric invariance free and in addition, sets the seven thresholds failing to show scalar invariance free. This model corresponds to M5 (the final partial scalar invariance model) and produces a much lower girl latent mean of 0.188. The difference in latent means between a model assuming full metric and scalar invariance and a model which allows the factor loadings and thresholds of variant items to vary therefore is a reduction in the latent mean for girls of 0.037.

Table 46 Impact of CES-D gender measurement models on latent means

		χ^2	df	Latent mean Estimate	se	Latent mean Δ from M9
M9	Full metric and full scalar invariance	4590.61	398	0.225	0.024	
M10	Partial metric and full scalar invariance	4336.20	394	0.270	0.024	+0.045
M11	Partial metric and partial scalar invariance	3670.30	379	0.188	0.025	-0.037
M12	M9 with *Bothered* factor loading free	4590.36	397	0.224	0.024	-0.001
M13	M12 with *Bothered* thresholds free	4503.50	395	0.209	0.024	-0.016
M14	M9 with *Appetite* factor loading free	4543.50	397	0.239	0.024	+0.014
M15	M14 with *Appetite* thresholds free	4426.69	395	0.204	0.024	-0.021
M16	M9 with *Unfriendly* factor loading free	4494.22	397	0.249	0.025	+0.024
M17	M16 with *Unfriendly* thresholds free	4430.95	395	0.250	0.025	+0.025
M18	M9 with *Cry* factor loading free	4497.39	397	0.230	0.024	+0.005
M19	M18 with *Cry* thresholds free	4413.09	395	0.217	0.024	-0.008
M20	M9 with first threshold of *Blues* free	4581.64	397	0.223	0.024	-0.002
M21	M9 with first threshold of *Good* free	4567.31	397	0.208	0.025	-0.017
M22	M9 with first threshold of *Effort* free	4558.72	397	0.220	0.024	+0.005
M23	M9 with first threshold of *Sleep* free	4557.93	397	0.211	0.024	-0.014
M24	M9 with first threshold of *Happy* free	4490.56	397	0.253	0.024	+0.028
M25	M9 with first threshold of *Sad* free	4574.57	397	0.218	0.024	-0.007
M26	M9 with third threshold of *Effort* free	4562.93	397	0.233	0.024	+0.008

The rather small difference in latent means might have occurred because the impact at the item level although quite large cancels itself out, with some items favouring girls and others favouring boys. Alternatively, the small difference might have occurred because the impact of the lack of invariance at the item level is trivial. To test these competing explanations a series of further models is estimated to examine the impact on latent means from freeing factor loadings and thresholds individually.

In M12 and M13 the factor loadings and thresholds of the Item 1 (*Bothered*) are allowed to vary across groups. The impact on the latent mean from allowing factor loadings to be estimated separately for boys and girls is negligible (0.001) but the impact from freeing the thresholds is a larger reduction of 0.016. Is this change in latent means in the direction expected?

Earlier, descriptive analyses (see Table 14) showed that the mean value for Item 1 (*Bothered*) was higher for girls (0.56) than it was for boys (0.33). IRT analyses (see Figure 3) also showed that, for most of the sample, for this item, girls exhibited slightly higher item scores compared to boys for equal levels of depressive symptomatology. These results suggest that when this DIF is controlled, as it is by allowing factor loadings and thresholds to vary across groups, then the latent mean for girls, relative to boys should decrease. Reassuringly (given the complexity of these analyses), this is exactly what does happen.

For Item 2 (*Appetite*) (M14 & M15) and Item 17 (*Cry*) (M18 & M19) a similar pattern of impact on latent means is observed. Both these items serve to increase total scores for girls relative to boys and when this is controlled, the girl latent mean value is reduced. With respect to Item 15 (*Unfriendly*), this item increases scores for boys. Descriptive analyses indicated a higher mean value for this item for boys (0.45) compared to girls (0.42) and IRT analyses showed significantly higher item scores for boys compared to girls at equivalent levels of depressive symptomatology. As expected therefore when this DIF is taken into account in the model (M17) the latent mean value for girls increases by 0.025.

In the final series of models (M20-M26) item thresholds which are not invariant are allowed to vary across boys and girls individually. With three exceptions allowing these thresholds to vary decreases the latent mean for girls. The exceptions are the first thresholds for Item 7 (*Effort*) and Item 12 (*Happy*) and the third threshold for Item 7 (*Effort*). When these thresholds are allowed to vary across boys and girls the latent mean for girls increases. These results could also have been expected on the basis of the earlier IRT analyses.

In summary, the overall effect of the lack of measurement invariance in CES-D is to increase total scores for girls. The magnitude of this increase, however, is fairly small. It will be remembered that in the models estimated, the variance of the latent factor for boys is set to one. This scaling allows the factor scores to be expressed in a *Z* score metric. A model that does not account for the lack of gender invariance produces a girl latent mean value of 0.225 or just over one fifth of a standard deviation higher score than boys. The so-called 'best' model incorporating varying factor loadings and thresholds produces a girl mean value of 0.188 or just under one fifth of a standard deviation higher score than boys.

Translating this impact estimate across to a raw mean total score change is problematic but approximately the observed gender difference in total scores in the present data set is probably around 20% ((100/188)*225) larger than it would otherwise be if full measurement invariance applied. The raw difference in total scores was 2.87 (Boys mean = 10.80; Girls mean = 13.67) and a reduction of 20% to this difference equates to 0.574 of a CES-D point. This estimate of the impact of the lack of gender measurement (one half of a CES-D point) on total CES-D scores is roughly in the same order of magnitude as the estimated impact calculated from the previous IRT analyses.

Year level invariance analyses

In this section the results of measurement invariance analyses across year level are presented. The key question addressed in these analyses is whether the CES-D shows measurement invariance across students aged (on average) between 13 years (Year 8), 14 years (Year 9) and 15 years (Year 10). An identical series of analyses are

undertaken to those used to test gender invariance. To avoid redundancy a less detailed explanation of these year level analyses is provided.

The first model (M0Y) tests whether the CES-D shows year level invariant covariance (whether $\sum^g = \sum^g$). This model provides a very good fit to the data: (χ^2 = 259.60, df = 80, $p < 0.001$, CFI = 0.996, TLI = 0.973, RMSEA = 0.032). The significant chi-square statistic indicates that the null hypothesis that the three groups' covariance matrices are equal should be rejected but the alternative fit indices indicate excellent overall fit. With some justification it might be concluded that year level measurement invariance has been demonstrated for the CES-D. Clearly a judgement is called for but for the sake of completeness further tests of invariance are performed. It might be expected however, that few, if any items will show signs of a lack of measurement invariance.

Next configural invariance ($\Lambda^g{}_{(form)} = \Lambda^g{}_{(form)}$) is tested. A model (M1Y) is estimated using the one factor CES-D model allowing factor loadings and threshold to vary across groups. To identify this model the factor loadings for Item 6 (*Depress*) and the third threshold of each item are constrained to be equal across year levels. In addition, the second threshold for the marker item is constrained to be equal across groups. This model (M1Y) provides quite a good fit to the data: (χ^2 = 3777.63, df = 510, $p < 0.001$, CFI = 0.932, TLI = 0.924, RMSEA = 0.053) and factor loadings, with the exception Item 7 (*Effort*) are all above 0.50. These results indicate that students (data from boys and girls are combined) are employing a similar frame of reference to the construct of depressive symptomatology across year levels.

Using the configural model as the base model, factor loadings ($\Lambda^g = \Lambda^g$) are constrained to be equal across year levels. The fit (χ^2 = 3898.17, df = 548, $p < 0.001$, CFI = 0.930, TLI = 0.927, RMSEA = 0.052) of the year level metric invariance model (M2Y) shows a deterioration in fit compared with M1Y. The change in the chi-square is significant ($\Delta\chi^2$ = 120.54, df = 38, $p < 0.001$) and the CFI decreases by 0.002. The TLI (Δ = +0.003) and RMSEA (Δ = -0.001) fit indices on the other hand indicate a marginal improvement to model fit. Again it is a matter of judgement whether the hypothesis of full metric invariance is supported or not, but for completeness further tests of individual factor loadings are carried out.

Table 47 shows the results from 38 tests of individual factor loadings. Using the Bonferroni correction the critical probability level is set to 0.0013 (0.05/38), which for one degree of freedom corresponds to a critical chi-square value of 10.34. Using this critical value, Item 14 (*Lonely*), in Year 10 shows a lack of metric invariance. The factor loading for this item when loadings are constrained across year levels is 0.91. The factor loading for this item in the Year 10 group when it was allowed to be free is 0.78 indicating that this item is less salient to this older age group.

To test whether partial metric invariance could be established across year levels a model (partial metric invariance model: M3Y) is estimated with the factor loading of Item 14 (*Lonely*) in the Year 10 group set free. All other factor loadings are constrained to be equal across groups. The fit statistics for this model are acceptable (χ^2 = 3877.34, df = 547, $p < 0.001$, CFI = 0.931, TLI = 0.928, RMSEA = 0.052). The change in chi-square statistic from the configural model (M1Y) in which all factor loadings are allowed to be free although statistically significant ($\Delta\chi^2$ = 99.71, df = 37, $p < 0.01$) is relatively small and in addition, the TLI statistic and RMSEA value indicate slight improvement to model fit. On this basis it is concluded that the hypothesis of partial metric invariance is supported.

Year level scalar invariance ($\tau^{g}_{c-1} = \tau^{g}_{c-1}$) is tested by constraining the thresholds of the metrically invariant items to be equal across year level. The fit statistics for this model (M4Y) are acceptable (χ^2 = 4227.69, df = 623, $p < 0.001$, CFI = 0.925, TLI = 0.931, RMSEA = 0.051) but the change in chi-square statistic from the partial metric model (M3Y) in which thresholds are allowed to be free is large ($\Delta\chi^2$ = 350.35, df = 76, $p < 0.001$). In addition, one of the alternative fit indices showed a deterioration in fit (ΔCFI = -0.006). On this basis it is concluded that the hypothesis of scalar invariance is not supported.

In order to investigate the source of the deterioration of fit from constraining the thresholds of the invariant items, a further 76 models are estimated freeing each threshold individually. The alpha level is treated by the Bonferroni correction. With 76 threshold invariance chi-square difference tests the critical probability level is set to 0.00066 (0.05/76) which for one degree of freedom corresponds to a critical chi-square value of 11.63. The results are presented in Table 48.

Table 47 Change in chi-square value from setting factor loadings free (Year level)

	Year 9		Year 10	
	χ^2	$\Delta\chi^2$	χ^2	$\Delta\chi^2$
Constrained. Model	3898.17			
1 Bothered	3889.87	8.30	3897.47	0.70
2 Appetite	3896.60	1.57	3898.00	0.17
3 Blues	3896.21	1.96	3888.47	9.70
4 Good	3896.18	1.99	3898.17	0.00
5 Mind	3898.14	0.03	3897.85	0.32
6 Depress (M)				
7 Effort	3895.80	2.37	3890.96	7.21
8 Hopeful	3896.66	1.51	3890.00	8.17
9 Failure	3896.91	1.26	3895.36	2.81
10 Fearful	3898.13	0.04	3898.10	0.07
11 Sleep	3890.32	7.85	3898.06	0.11
12 Happy	3896.16	2.01	3891.13	7.04
13 Talk	3898.00	0.17	3893.90	4.27
14 Lonely	3894.78	3.39	3877.34	**20.83**
15 Unfriendly	3898.15	0.02	3897.66	0.51
16 Enjoy	3896.99	1.18	3897.99	0.18
17 Cry	3897.95	0.22	3896.80	1.37
18 Sad	3898.10	0.07	3897.47	0.70
19 Dislike	3897.16	1.01	3898.12	0.05
20 Getgoing	3895.70	2.47	3897.82	0.35

The results in Table 48 show that all the first and second thresholds between Year 8 and Year 9 are invariant. Between Year 8 and Year 10 the first thresholds are not invariant for the items: *Bothered* (1), *Effort* (7) and *Unfriendly* (15) and the second threshold is not invariant for Item 20 (*Getgoing*). Also between Year 8 and Year 10 the first threshold for Item 5 (*Mind*) showed a trend towards significance with a large chi square value (9.64). This value was markedly higher than the other non-significant chi square values but less than the critical value of 11.63. Relaxing only the four constraints (with the remainder fixed across year level) for the items: *Bothered* (1), *Effort* (7), *Unfriendly* (15) and *Getgoing* (20) produces a marked improvement to the model: (χ^2 = 4140.85, df = 619, $p < 0.001$, CFI = 0.927, TLI = 0.932, RMSEA = 0.050).

Table 48 Change in chi-square value from constraining thresholds (Year level)

		Year 9				Year 10			
		Threshold 1		Threshold 2		Threshold 1		Threshold 2	
		χ^2	$\Delta\chi^2$	χ^2	$\Delta\chi^2$	χ^2	$\Delta\chi^2$	χ^2	$\Delta\chi^2$
1	Bothered	4227.69	0.00	4227.16	0.53	4214.37	**13.32**	4221.75	5.94
2	Appetite	4225.89	1.80	4227.61	0.08	4226.48	1.21	4227.56	0.13
3	Blues	4227.34	0.35	4227.39	0.30	4227.67	0.02	4227.63	0.06
4	Good	4227.61	0.08	4227.34	0.35	4225.62	2.07	4225.39	2.30
5	Mind	4225.91	1.78	4227.65	0.04	4218.05	9.64	4224.68	3.01
6	Depress (M)	4227.61	0.08			4227.59	0.10		
7	Effort	4226.82	0.87	4225.97	1.72	4197.92	**29.77**	4221.16	6.53
8	Hopeful	4225.01	2.68	4227.33	0.36	4226.25	1.44	4226.15	1.54
9	Failure	4223.55	4.14	4222.64	5.05	4224.43	3.26	4226.48	1.21
10	Fearful	4227.68	0.01	4226.84	0.85	4227.68	0.01	4227.66	0.03
11	Sleep	4227.62	0.07	4227.57	0.12	4227.31	0.38	4227.13	0.56
12	Happy	4223.90	3.79	4227.32	0.37	4227.66	0.03	4227.41	0.28
13	Talk	4227.58	0.11	4227.48	0.21	4224.29	3.40	4220.56	7.13
14	Lonely	4227.67	0.02	4224.42	3.27				
15	Unfriendly	4222.32	5.37	4220.44	7.25	4199.24	**28.45**	4226.56	1.13
16	Enjoy	4227.05	0.64	4225.12	2.57	4226.34	1.35	4226.69	1.00
17	Cry	4226.91	0.78	4225.90	1.79	4223.38	4.31	4226.60	1.09
18	Sad	4225.03	2.66	4226.38	1.31	4227.62	0.07	4227.32	0.37
19	Dislike	4227.57	0.12	4227.69	0.00	4226.99	0.70	4223.46	4.23
20	Getgoing	4227.69	0.00	4218.44	9.25	4225.58	2.11	4214.30	**13.39**

M denotes marker item

The change in chi-square statistic compared with the partial metric invariance model (M3Y) is relatively small (relative to the degrees of freedom), although statistically significant ($\Delta\chi^2 = 263.51$, df = 72, $p < 0.01$). Two of the alternative fit statistics for model M5 show improvement over M3Y (ΔTLI = +0.004, ΔRMSEA = -0.002) but the other measure indicates a deterioration (ΔCFI = -0.003). On the basis that no other threshold parameter stands out as being able to produce a substantial improvement to fit, this model (M5Y) is accepted as the final scalar invariance model.

Factor loadings estimated from M5Y are presented in Table 49. The key point of interest in this table is that the factor loading for Item 14 (Lonely) is lower in Year 10 compared with Years 8 and 9 (Year 8 & 9: 0.92; Year 10: 0.72).

Table 49 Factor loadings from the final scalar model (M5Y)

	Year 8	Year 9	Year 10
1 Bothered	0.75	0.75	0.75
2 Appetite	0.63	0.63	0.63
3 Blues	0.87	0.87	0.87
4 Good	0.67	0.67	0.67
5 Mind	0.68	0.68	0.68
6 Depress (M)	0.93	0.93	0.93
7 Effort	0.22	0.22	0.22
8 Hopeful	0.55	0.55	0.55
9 Failure	0.89	0.89	0.89
10 Fearful	0.80	0.80	0.80
11 Sleep	0.63	0.63	0.63
12 Happy	0.85	0.85	0.85
13 Talk	0.66	0.66	0.66
14 Lonely	0.92	0.92	*0.79
15 Unfriendly	0.75	0.75	0.75
16 Enjoy	0.87	0.87	0.87
17 Cry	0.85	0.85	0.85
18 Sad	0.93	0.93	0.93
19 Dislike	0.90	0.90	0.90
20 Getgoing	0.79	0.79	0.79

* denotes free parameter; M denotes marker item

Thresholds estimated from M5Y are presented in Table 50. Table 50 shows that the Year 10 first thresholds for the items: *Bothered* (1), *Effort* (7), *Unfriendly* (15) and the second threshold for Item 20 (*Getgoing*) are allowed to vary across year levels. With respect to the first thresholds, for Item 1 (*Bothered*) (Year 8 & 9: 0.50; Year 10: 0.41) this symptom is more likely to be acknowledged in Year 10. For Item 7 (*Effort*) (Year 8 & 9: -0.20; Year 10: -0.06) and Item 15 (*Unfriendly*) (Year 8 & 9: 0.44; Year

10: 0.57) the reverse is true with these symptoms less likely to be acknowledged in Year 10. For Item 20 (*Getgoing*) the second threshold is more likely to be endorsed in Year 10 (Year 8 & 9: 1.02; Year 10: 0.92).

Table 50 Thresholds from the final scalar model (M5Y)

		Year 8			Year 9			Year 10		
	Thresholds	1	2	3	1	2	3	1	2	3
1	Bothered	0.50	1.15	1.62	0.50	1.15	1.62	*0.41	1.15	1.62
2	Appetite	0.53	1.11	1.61	0.53	1.11	1.61	0.53	1.11	1.61
3	Blues	0.54	1.00	1.42	0.54	1.00	1.42	0.54	1.00	1.42
4	Good	-0.17	0.47	0.98	-0.17	0.47	0.98	-0.17	0.47	0.98
5	Mind	-0.23	0.59	1.23	-0.23	0.59	1.23	-0.23	0.59	1.23
6	Depress (M)	0.31	0.87	1.30	0.31	0.87	1.30	0.31	0.87	1.30
7	Effort	-0.20	0.44	1.04	-0.20	0.44	1.04	*-0.06	0.44	1.04
8	Hopeful	-0.47	0.37	1.02	-0.47	0.37	1.02	-0.47	0.37	1.02
9	Failure	0.73	1.14	1.51	0.73	1.14	1.51	0.73	1.14	1.51
10	Fearful	0.71	1.25	1.75	0.71	1.25	1.75	0.71	1.25	1.75
11	Sleep	0.21	0.83	1.33	0.21	0.83	1.33	0.21	0.83	1.33
12	Happy	0.02	0.77	1.39	0.02	0.77	1.39	0.02	0.77	1.39
13	Talk	0.25	0.97	1.59	0.25	0.97	1.59	0.25	0.97	1.59
14	Lonely	0.52	1.03	1.43	0.52	1.03	1.43	*0.44	*0.94	1.43
15	Unfriendly	0.44	1.23	1.79	0.44	1.23	1.79	*0.57	1.23	1.79
16	Enjoy	0.01	0.70	1.19	0.01	0.70	1.19	0.01	0.70	1.19
17	Cry	0.87	1.36	1.79	0.87	1.36	1.79	0.87	1.36	1.79
18	Sad	0.42	1.01	1.49	0.42	1.01	1.49	0.42	1.01	1.49
19	Dislike	0.35	1.05	1.48	0.35	1.05	1.48	0.35	1.05	1.48
20	Getgoing	0.23	1.02	1.51	0.23	1.02	1.51	0.23	*0.92	1.51

* denotes free parameter; M denotes marker item

A model (M6Y) is estimated to test the hypothesis of year level invariant uniquenesses ($\Theta^g = \Theta^g$). This model is identical to M5Y but with the added constraint that item residual variances are equal across year levels. This was achieved by setting the item scale factors to one in the three groups and the factor variances to one. Only the residual variances for items demonstrating metric invariance are tested. This means the residual variance for Item 14 (*Lonely*) in Year 10 is not tested. The fit statistics for M6Y are: (χ^2 = 4439.54, df = 660, $p < 0.001$, CFI = 0.921, TLI = 0.932, RMSEA = 0.050).

The change in chi-square statistic from a model (M7Y: to be presented later) in which factor variances were also required to be equal is significant ($\Delta\chi^2$ = 294.85, df = 39, p

< 0.001) and two alternative fit indices show a deterioration in fit (ΔCFI = -0.006, ΔTLI = -0.001). It is concluded that the hypothesis of equal item residual variances is not supported.

To investigate the source of the deterioration of fit from constraining the item residual variances a further 39 models (20 Year 9 items and 19 Year 10 items) are estimated freeing each item residual variance individually. Using the Bonferroni correction the critical probability level is set to 0.0013 (0.05/39 = 0.0013) which for one degree of freedom corresponds to a critical chi-square value of 10.34.

Table 51 shows the results from the series of models estimated to examine whether item residual variances are equal across year levels. The residual variances for nine items appear not to be equal across Year 8 and Year 9 and the residual variances for nine items appear not to be equal across Year 8 and Year 10. In all these 18 cases the residual variance is higher (indicating poorer reliability) for Year 8 than it is for Year 9 or Year 10. This suggests that the reliability for nearly one half of CES-D items improves across Year 8 to Year 10.

A model (M7Y) identical to M5Y is estimated but with the constraint that factor variances ($\Phi_j^g = \Phi_j^g$) are equal across year levels. The fit statistics for M7Y are: (χ^2 = 4144.69, df = 621, p < 0.001, CFI = 0.927, TLI = 0.933, RMSEA = 0.050) and the change in the chi-square statistic from M5Y in which factor variances were not required to be equal is not statistically significant ($\Delta\chi^2$ = 3.84, df = 2, p = 0.147). On this basis it is concluded that the hypothesis of equal factor variances across year levels is supported.

A model (M8Y) identical to M5Y is estimated but with the added constraint that the year level factor means ($\kappa^g = \kappa^g$) are equal. The fit statistics for M8Y are: (χ^2 = 4142.76, df = 621, p < 0.001, CFI = 0.927, TLI = 0.933, RMSEA = 0.050) and the change in chi-square statistic from M5Y in which latent means were not required to be equal is not significant ($\Delta\chi^2$ = 1.95, df = 2, p < 0.377). On this basis it is concluded that the null hypothesis of equal latent means could not be rejected.

The finding that latent mean estimates are equal across year levels is to be expected on the basis of simple descriptive statistics. Using the dataset with boys and girls combined the raw mean total CES-D scores across year levels are as follows: Year 8: 12.50 (SD = 10.19); Year 9: 11.91 (SD = 9.82); Year 10: 11.81 (SD = 9.87). The latent mean estimates (from M5Y) with Year 8 as the reference group are Year 8: 0.00; Year 9: 0.022; Year 10: 0.037. Expressed in the Z score metric these differences are in the order of less than one twentieth of a standard deviation.

A series of models are estimated to examine the impact on latent means from the failure of the CES-D to exhibit full measurement invariance across year levels. These results are shown in Table 52. The first model (M9Y) assumes full metric and scalar invariance by constraining all factor loadings and thresholds to be equal across year levels. The latent mean estimates produced from this model are Year 9 (0.026) and Year 10 (0.041).

Next, in model M10Y, the factor loading for Item 14 (*Lonely*) in Year 10 which failed to exhibit metric invariance is set free. This has the effect of decreasing the latent mean in Year 9 (0.021) and Year 10 (0.036). M11Y sets the factor loading and thresholds for Item 14 (*Lonely*) free in Year 10 and in addition, sets the four thresholds failing to show scalar invariance free. This model corresponds to M5Y (the final partial scalar year level invariance model) and produces a Year 9 latent mean of 0.023 and a Year 10 latent mean estimate of 0.039.

Table 51 Change in chi-square value from constraining item residual variances (Year level)

		Item residual variances free		Constrained	$\Delta\chi^2$
		Year 8	Year 9		from constraining
1	Bothered	0.45	0.39	0.43	3.45
2	Appetite	0.61	0.52	0.59	4.11
3	Blues	0.25	0.19	0.22	5.84
4	Good	0.56	0.38	0.50	**34.99**
5	Mind	0.55	0.47	0.53	5.37
6	Depress (M)	0.14	0.12	0.13	1.07
7	Effort	0.96	0.76	0.95	**11.96**
8	Hopeful	0.70	0.57	0.68	**14.57**
9	Failure	0.21	0.11	0.16	**16.23**
10	Fearful	0.36	0.30	0.34	3.14
11	Sleep	0.61	0.46	0.57	**16.82**
12	Happy	0.29	0.22	0.26	**13.17**
13	Talk	0.58	0.45	0.54	**13.94**
14	Lonely	0.16	0.17	0.16	0.16
15	Unfriendly	0.44	0.35	0.41	9.03
16	Enjoy	0.25	0.18	0.21	**13.91**
17	Cry	0.28	0.30	0.29	0.42
18	Sad	0.15	0.09	0.12	**19.82**
19	Dislike	0.19	0.14	0.17	7.95
20	Getgoing	0.39	0.35	0.37	1.96

		Item residual variances free		Constrained	$\Delta\chi2$
		Year 8	Year 10		from constraining
1	Bothered	0.45	0.37	0.42	6.55
2	Appetite	0.61	0.49	0.58	8.18
3	Blues	0.25	0.14	0.18	22.90
4	Good	0.56	0.36	0.50	40.81
5	Mind	0.55	0.41	0.51	19.58
6	Depress (M)	0.14	0.11	0.12	2.79
7	Effort	0.96	0.54	0.94	58.45
8	Hopeful	0.70	0.56	0.68	15.5
9	Failure	0.21	0.13	0.17	9.78
10	Fearful	0.36	0.27	0.33	8.20
11	Sleep	0.61	0.45	0.57	19.89
12	Happy	0.29	0.19	0.25	31.63
13	Talk	0.58	0.39	0.53	32.1
14	Lonely	0.16	0.14		
15	Unfriendly	0.44	0.40	0.43	1.95
16	Enjoy	0.25	0.17	0.21	17.76
17	Cry	0.28	0.25	0.27	1.15
18	Sad	0.15	0.10	0.13	11.85
19	Dislike	0.19	0.16	0.18	2.38
20	Getgoing	0.39	0.26	0.34	23.89

Table 52 Impact of CES-D year level measurement models on latent means

	χ^2	df	Latent mean	
			Year 9	Year 10
M9Y Full metric and full scalar invariance	4255.49	626	0.026	0.041
M10Y Partial metric and full scalar invariance	4243.85	625	0.021	0.036
M11Y Partial metric and partial scalar invariance	4140.85	619	0.023	0.039

The difference in latent means between a model assuming full metric and scalar invariance and a model which allows variant factor loadings and thresholds therefore is a reduction in the latent mean estimate of the order of 0.002. The magnitude of this difference in latent means is small enough to be considered trivial. A summary of the year level measurement models presented in this section is provided in Table 53.

Table 53 Year level model fit statistics

	χ^2	df	CLI	TLI	RMSEA
M0Y Invariant covariance	259.60	80	0.996	0.973	0.032
M1Y Configural invariance	3777.63	510	0.932	0.924	0.053
M2Y Full metric invariance	3898.17	548	0.930	0.927	0.052
M3Y Partial metric invariance	3877.34	547	0.931	0.928	0.052
M4Y Initial scalar invariance	4227.69	623	0.925	0.931	0.051
M5Y Final partial scalar invariance	4140.85	619	0.927	0.932	0.050
M6Y M5Y with constrained item variances	4439.54	660	0.921	0.932	0.050
M7Y M5Y with constrained factor variances	4144.69	621	0.927	0.933	0.050
M8Y M5Y with constrained means	4142.76	621	0.927	0.933	0.050

Summary

In this chapter SEM techniques are used to examine CES-D measurement invariance across gender and year level groups. The key findings presented in this chapter can be summarised as follows.

Gender configural invariance is supported using a one factor CES-D measurement model suggesting that boys and girls employ the same conceptual frame of reference to the CES-D.

Full metric invariance is not supported with the factor loadings for: *Bothered* (1), *Appetite* (2), *Unfriendly* (15) and *Cry* (17) varying across gender. In addition, the first threshold for six items: *Blues* (3), *Good* (4), *Effort* (7), *Sleep* (11), *Happy* (12) and *Sad* (18) and the third threshold for Item 7 (*Effort*) failed to demonstrate scalar invariance across gender.

A null hypothesis of equal gender factor variances is unable to be rejected. The residual variances for six items: *Good* (4), *Depress* (6), *Effort* (7), *Hopeful* (8), *Happy* (12) and *Getgoing* (20) were found to vary across gender. For all items, residual variances were higher (indicating poorer reliability) for boys than for girls.

The overall impact of the lack of measurement invariance on gender latent means is relatively minor. The latent mean estimate of depressive symptomatology for girls remains significantly higher compared to boys even when the lack of measurement invariance is taken into account.

It is estimated that when expressed as a difference in raw total observed scores, the lack of CES-D gender invariance adds approximately one half of a CES-D point to girls' scores.

Across year levels the factor loading for Item 14 (*Lonely*) in Year 10, is found not to be invariant. In addition, between Year 8 and Year 10, the first threshold is not invariant for the items: *Bothered* (1), *Effort* (7) and *Unfriendly* (15) and the second threshold is not invariant for Item 20 (*Getgoing).*

Overall the impact on year level latent mean estimates from the lack of measurement invariance in factor loadings and thresholds is very minor.

9

HLM Analyses of CES-D School Effects

This chapter presents the results from a series of analyses which examine possible school effects on student levels of depressive symptomatology. These analyses seek to establish the extent to which schools taking part in the EDED program vary with respect to student CES-D scores and whether this variation increases between Year 8 and Year 10 of high school, consistent with a school effect on student levels of depressive symptomatology.

In the first section descriptive analyses at the school level are performed to examine student CES-D scores across schools. These are carried out using both CES-D total scores as a continuous measure of depressive symptomatology and also by creating a dichotomous variable (CES-D22) coded zero for scores 21 or less and coded one for scores 22 or greater (see Method section for further details). When aggregated to the school level this variable indicates the proportion of high CES-D scoring students in each school.

In the second section a series of nine HLM linear models are estimated to examine the variation in school average CES-D scores by year level and gender. In these analyses the CES-D total score is treated as a continuous measure. In the third section these same analyses are repeated but this time with the dichotomous variable CES-D22. For these analyses nine HLM logistic models are estimated and because of difficulties in decomposing the variance in HLM logistic models the analyses are replicated with HLM linear models.

In the final section of this chapter a series of analyses are undertaken to examine the possibility of differences between the public and private schools. These are carried out using CES-D total scores as a continuous measure of depressive symptomatology and by creating a dichotomous variable (School type) coded one for public schools and two for private schools. Initially six HLM linear models are estimated for each year level and school type separately. Following this the variable 'School type' is used as a Level-2 predictor in three HLM linear models for each year level. This analysis is repeated, controlling for the possibility that the gender proportion between

public and private sectors might vary, by the addition to the models estimated of a Level-1 predictor (Gender).

Descriptive analyses of school CES-D differences

A total of 26 schools participated in the EDED program and Table 54 shows the number of students taking part at each school by year level and gender. Because not every school participated in the program for the full three years or was coeducational several data cells in this table are missing. Specifically, School 7 closed in the third wave of data collection (Year 10), School 17 is a girls only school and did not take part in the year of the third wave of data collection (Year 10), Schools 19 and 20 were boys only and School 26 is a girls only school.

Table 54 Number of students by school, year level and gender

School	Year 8			Year 9			Year 10		
	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total
1	86	85	171	88	86	174	82	90	172
2	101	97	198	101	96	197	104	92	196
3	44	24	68	48	25	73	39	27	66
4	72	70	142	68	66	134	75	62	137
5	20	20	40	13	18	31	43	29	72
6	15	15	30	10	11	21	10	10	20
7	33	27	60	21	22	43	-	-	-
8	54	55	109	47	58	105	53	48	101
9	32	20	52	28	23	51	27	25	52
10	105	88	193	110	77	187	86	75	161
11	61	55	116	63	55	118	61	61	122
12	29	20	49	27	24	51	21	12	33
13	99	76	175	102	82	184	97	70	167
14	66	46	112	62	32	94	69	38	107
15	41	40	81	46	38	84	41	48	89
16	21	12	33	19	12	31	22	14	36
17*	-	59	59	-	54	54	-	-	-
18*	23	25	48	23	28	51	23	17	40
19*	87	-	87	95	-	95	100	-	100
20*	60	-	60	67	-	67	84	-	84
21*	56	42	98	56	47	103	52	46	98
22*	25	17	42	21	26	47	21	24	45
23*	27	20	47	30	26	56	21	29	50
24*	59	38	97	69	45	114	49	47	96
25*	52	56	108	38	43	81	43	54	97
26*	-	31	31	-	29	29	-	17	17
Total	1268	1038	2306	1252	1023	2275	1223	935	2158

Very little school level data were collected during the EDED program because only student level analyses were anticipated by the program designers. At the school level the 16 public and 10 private schools are able to be distinguished and in tables where individual school level data are provided, private schools are identified by an *. In 1996 most (70.7%) school students were enrolled in government schools with the remainder enrolled in Catholic schools (19.6%) or Independent schools (9.7%) (Ministerial Council on Education, Employment, Training and Youth Affairs, 1998). The present sample comprised 70.7 per cent government school students and 29.3 per cent Independent schools.

Table 55 School CES-D means by school, year level and gender

School	Year 8			Year 9			Year 10		
	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total
1	11.97	15.80	13.87	8.85	13.02	10.91	10.04	13.32	11.76
2	11.55	11.92	11.73	11.11	11.74	11.42	8.92	13.61	11.12
3	12.09	10.13	11.40	12.15	15.48	13.29	10.44	13.59	11.73
4	11.65	11.86	11.75	9.90	12.03	10.95	9.36	11.65	10.39
5	12.05	9.85	10.95	13.15	14.94	14.19	11.56	18.93	14.53
6	11.13	12.07	11.60	15.10	19.55	17.43	9.80	13.20	11.50
7	13.15	11.89	12.58	10.33	11.32	10.84	-	-	-
8	9.57	13.05	11.33	9.02	13.90	11.71	12.17	13.88	12.98
9	14.28	15.85	14.88	10.50	19.35	14.49	10.74	17.92	14.19
10	11.51	14.83	13.03	11.28	12.25	11.68	9.97	11.68	10.76
11	14.11	16.00	15.01	11.62	13.20	12.36	9.80	10.61	10.20
12	13.31	11.70	12.65	11.63	14.33	12.90	8.14	11.42	9.33
13	11.36	14.53	12.74	9.93	10.65	10.25	9.18	13.00	10.78
14	11.15	20.11	14.83	10.02	22.22	14.17	10.38	20.55	13.99
15	10.44	12.65	11.53	13.76	16.03	14.79	10.37	15.10	12.92
16	7.86	11.58	9.21	10.21	14.42	11.84	11.82	15.93	13.42
17*	-	11.68	11.68	-	10.69	10.69	-	-	-
18*	10.70	13.92	12.38	8.30	14.32	11.61	10.26	18.82	13.90
19*	11.54	-	11.54	9.87	-	9.87	9.34	-	9.34
20*	10.98	-	10.98	10.58	-	10.58	8.99	-	8.99
21*	8.39	12.14	10.00	8.98	14.11	11.32	11.29	13.33	12.24
22*	13.40	16.24	14.55	9.81	13.38	11.79	11.19	12.04	11.64
23*	9.52	14.25	11.53	10.17	13.04	11.50	9.90	14.17	12.38
24*	11.93	14.08	12.77	15.39	12.82	14.38	14.47	14.09	14.28
25*	11.13	15.23	13.26	9.39	12.95	11.28	11.26	16.11	13.96
26*	-	12.26	12.26	-	14.07	14.07	-	13.76	13.76
Total	11.47	13.75	12.50	10.73	13.35	11.91	10.18	13.94	11.81

An approximate estimate of school size can be derived by taking into account that on average 85 per cent of students at each school took part in the EDED program and that the majority of the high schools in the sample have enrolments across five year levels (Year 8 to Year 12). On this basis School 1 with 171 students at Year 8 would be likely to have around a 1000 students. A small school such as School 6 with 30 students at Year 8 would comprise approximately 176 students.

Student CES-D means and standard deviations for each school by year level and gender are shown in Table 55 and Table 56. Considerable variability among mean values across schools is evident in each of the three year levels. School average scores for Year 8 ranged from 9.21 (Year 8: School 16) to 14.88 (Year 8: School 9). Of note are the relatively high standard deviations for each school.

Table 56 Standard deviations of school CES-D means by school, year level and gender

School	Year 8			Year 9			Year 10		
	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total
1	8.48	12.45	10.78	7.44	11.00	9.58	8.83	10.94	10.09
2	8.61	10.53	9.58	9.64	10.61	10.11	6.39	10.27	8.73
3	10.00	6.04	8.81	9.56	12.45	10.67	7.93	12.09	9.88
4	7.73	10.19	9.00	9.85	10.64	10.27	7.67	10.61	9.16
5	8.91	6.98	7.98	9.45	13.49	11.82	8.91	11.75	10.70
6	4.37	5.19	4.74	12.36	15.10	13.71	6.03	10.40	8.46
7	9.25	8.51	8.87	8.93	6.76	7.82	-	-	-
8	7.18	12.18	10.12	6.18	10.83	9.33	11.74	10.09	10.97
9	12.26	12.23	12.15	6.65	15.51	12.23	8.22	12.69	11.11
10	9.36	11.17	10.33	9.57	9.66	9.59	9.41	9.82	9.61
11	9.88	10.42	10.14	9.90	8.85	9.42	9.16	8.67	8.89
12	9.16	10.58	9.69	8.25	9.19	8.72	9.13	9.60	9.29
13	9.18	11.26	10.23	8.28	7.65	7.99	7.85	11.15	9.54
14	8.57	13.51	11.69	6.19	14.61	11.41	8.94	13.53	11.79
15	7.17	12.41	10.10	9.96	11.40	10.63	7.43	11.58	10.12
16	5.04	11.37	7.98	6.39	12.19	9.13	13.11	11.08	12.36
17*	-	11.15	11.15	-	11.12	11.12	-	-	-
18*	7.18	13.23	10.78	6.99	12.23	10.55	6.32	13.25	10.63
19*	10.99	-	10.99	8.24	-	8.24	6.87	-	6.87
20*	6.94	-	6.94	8.27	-	8.27	6.87	-	6.87
21*	7.57	9.77	8.73	7.91	9.76	9.12	10.46	11.30	10.85
22*	9.49	12.49	10.75	7.47	9.92	9.00	9.23	8.40	8.71
23*	7.57	14.26	11.05	7.97	10.69	9.35	6.42	12.29	10.37
24*	9.50	11.24	10.21	11.79	10.03	11.16	9.85	10.94	10.34
25*	9.43	14.44	12.40	6.53	9.30	8.26	8.87	11.38	10.57
26*	-	11.55	11.55	-	9.56	9.56	-	9.93	9.93
Total	8.89	11.47	10.19	8.81	10.76	9.82	8.52	11.04	9.87

Table 57 Coefficient of variation by school, year level and gender

School	Year 8			Year 9			Year 10		
	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total
1	.71	.79	.78	.84	.84	.88	.88	.82	.86
2	.75	.88	.82	.87	.90	.89	.72	.75	.79
3	.83	.60	.77	.79	.80	.80	.76	.89	.84
4	.66	.86	.77	1.00	.88	.94	.82	.91	.88
5	.74	.71	.73	.72	.90	.83	.77	.62	.74
6	.39	.43	.41	.82	.77	.79	.62	.79	.74
7	.70	.72	.71	.86	.60	.72	-	-	-
8	.75	.93	.89	.68	.78	.80	.96	.73	.84
9	.86	.77	.82	.63	.80	.84	.77	.71	.78
10	.81	.75	.79	.85	.79	.82	.94	.84	.89
11	.70	.65	.68	.85	.67	.76	.93	.82	.87
12	.69	.90	.77	.71	.64	.68	1.12	.84	1.00
13	.81	.78	.80	.83	.72	.78	.86	.86	.88
14	.77	.67	.79	.62	.66	.81	.86	.66	.84
15	.69	.98	.88	.72	.71	.72	.72	.77	.78
16	.64	.98	.87	.63	.85	.77	1.11	.70	.92
17*	-	.96	.96	-	1.04	1.04	-	-	-
18*	.67	.95	.87	.84	.85	.91	.62	.70	.76
19*	.95	-	.95	.83	-	.83	.74	-	.74
20*	.63	-	.63	.78	-	.78	.76	-	.76
21*	.90	.80	.87	.88	.69	.81	.93	.85	.89
22*	.71	.77	.74	.76	.74	.76	.82	.70	.75
23*	.80	1.00	.96	.78	.82	.81	.65	.87	.84
24*	.80	.80	.80	.77	.78	.78	.68	.78	.72
25*	.85	.95	.94	.70	.72	.73	.79	.71	.76
26*	-	.94	.94	-	.68	.68	-	.72	.72
Total	.77	.83	.82	.82	.81	.82	.84	.79	.84

The ratio of standard deviation to mean, known as the coefficient of variation (Snedecor & Cochran, 1967) for the majority of schools across the three year levels is around 0.80. These are shown in Table 57. These high coefficients of variation indicate a high level of variability within schools with respect to student average CES-D scores. No clear pattern with respect to gender is evident – in some schools the coefficient of variation is larger for boys (e.g. Year 8: School 3) while in other schools it is larger for girls (e.g. Year 8: School 16).

Figure 14 (Boys & Girls combined), Figure 15 (Boys only) and Figure 16 (Girls only) plot mean CES-D scores (shown in Table 55) across the three year levels for each of the 26 schools. A horizontal line denoting the overall school mean across the three year levels (estimated from a HLM analysis) is also shown in the figures. From Figure 14 it can be seen that for the majority of schools no clear pattern of change is evident. There were some exceptions: for example School 11 shows decreasing average student CES-D scores across the three year levels consistent with a positive school effect on student mental health while School 16 shows increasing average student CES-D scores across the three year levels consistent with a negative school effect. Overall, however, most schools show either very little change (e.g. School 14) or inconsistent change (e.g. School 6). This same pattern of results is repeated in the figures for boys (Figure 15) and girls (Figure 16)

The proportion of students scoring 22 CES-D points or above is shown in Table 58 by school, year level and gender. Consistent with the data presented earlier using CES-D mean scores considerable variability is evident between schools with respect to the proportion of high scoring students. In terms of school proportions with both boys and girls combined these range from 0 per cent (Year 8: School 6) to 28.6 per cent (Year 9: School 6). Generally high school mean CES-D levels are reflected in high levels of students scoring equal to or above the cut-point of 22.

Figure 17 (Boys & Girls combined), Figure 18 (Boys only) and Figure 19 (Girls only) show visually the data presented in Table 58. A horizontal line denoting the overall school average proportion of high scoring cases across the three year levels (estimated from a HLM analysis) is also shown in the figures. Considerable variability is evident across year levels within schools. For example, School 16 went from 3 per cent of students scoring at high CES-D levels in Year 8 to 25 per cent in Year 10. Again for the majority of schools, no clear pattern of change is evident.

The basic pattern of results is repeated in the figures for boys (Figure 18) and girls (Figure 19). Visual comparison of the figures for boys and girls suggests greater variation for girls in school proportions of high scoring students both across year levels within schools and between schools generally

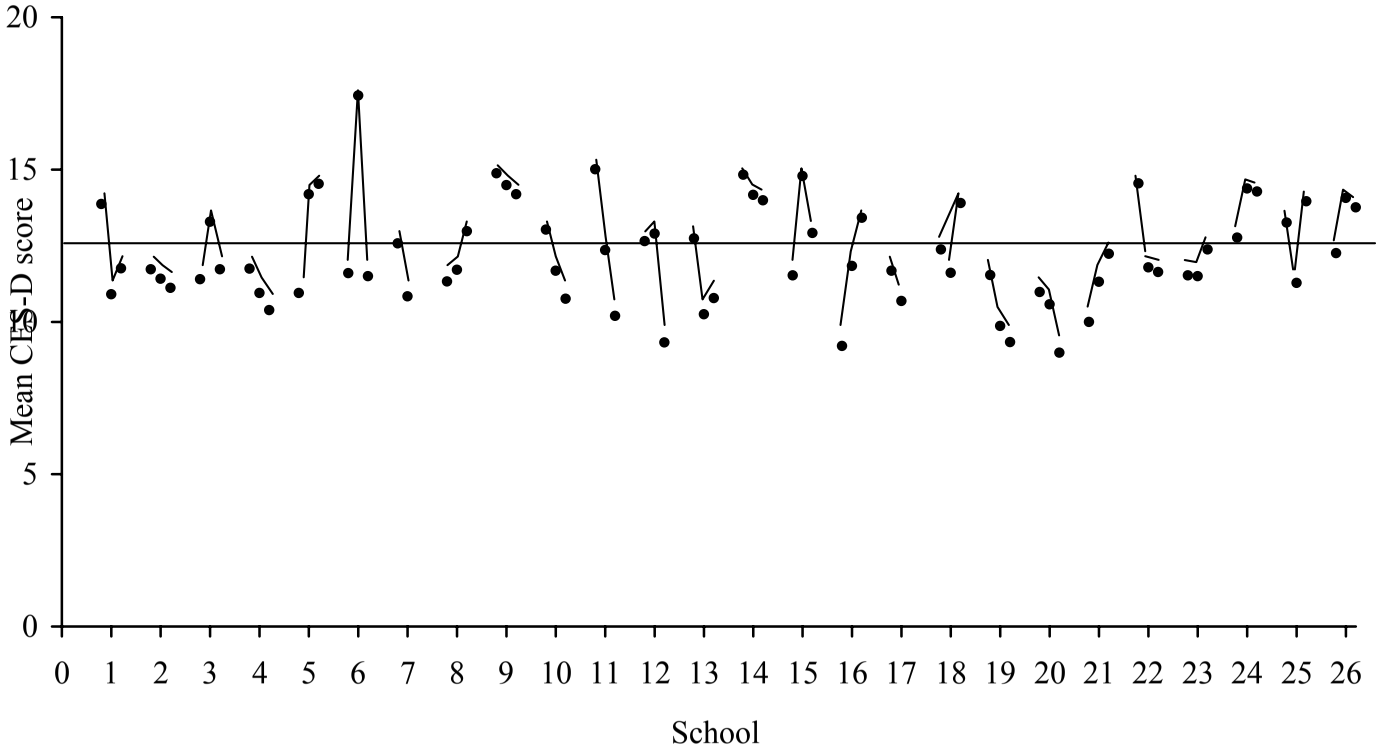


Figure 14 Individual school mean CES-D scores at Year 8, 9 and 10

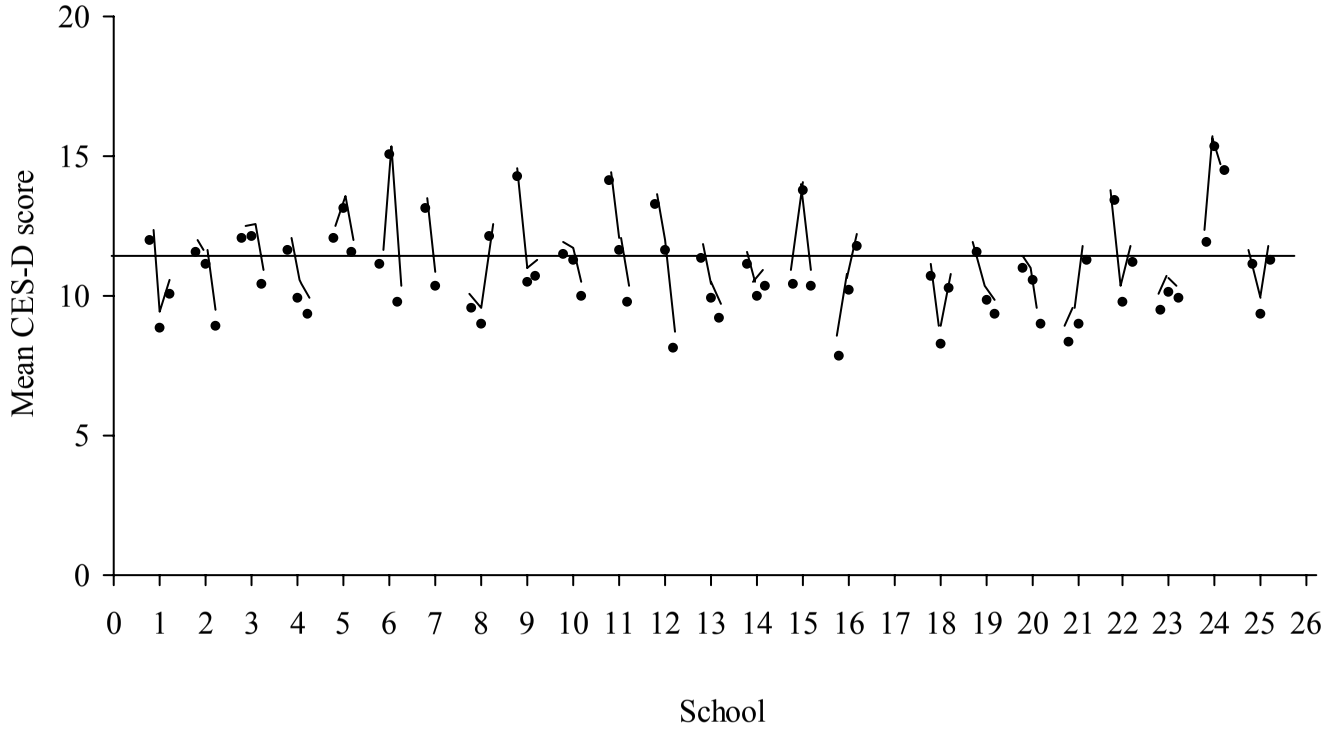


Figure 15 Individual school mean CES-D scores at Year 8, 9 and 10 (Boys)

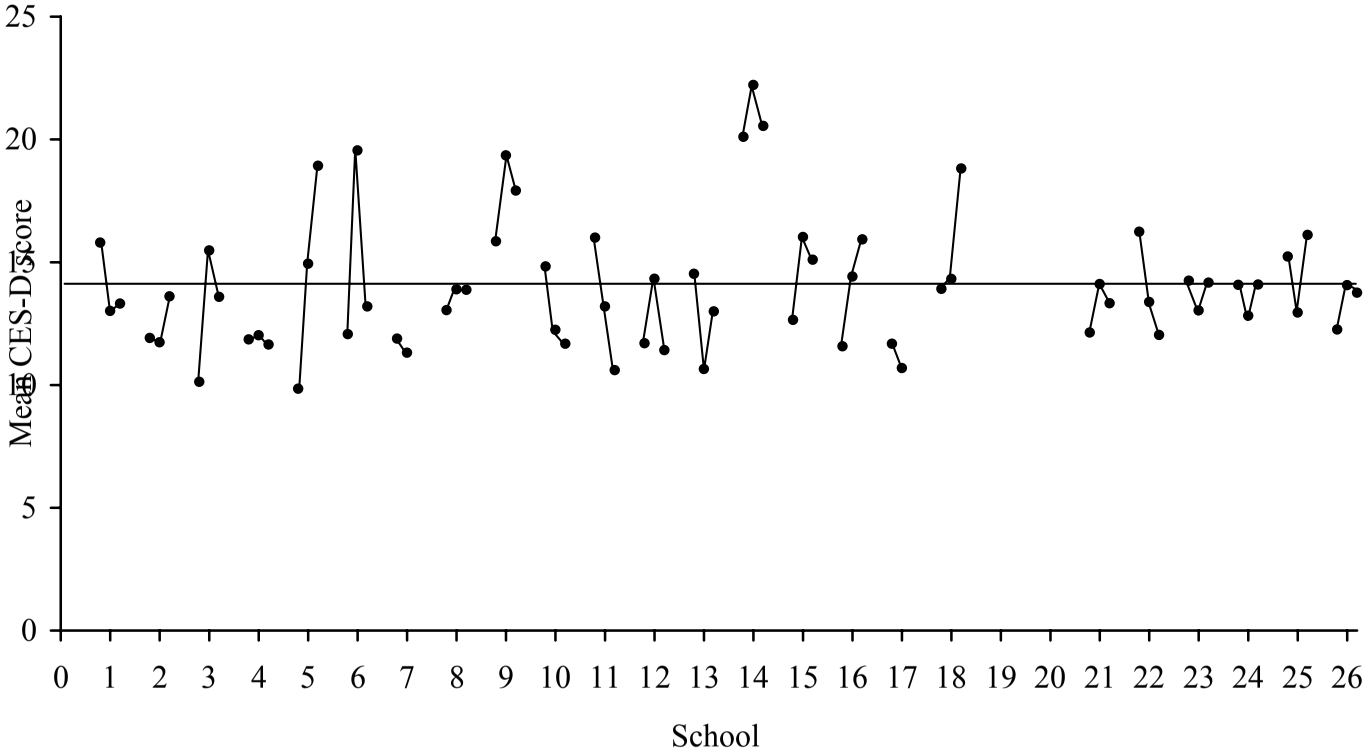


Figure 16 Individual school mean CES-D scores at Year 8, 9 and 10 (Girls)

Table 58 Percentage of high scoring CES-D cases by school, year level and gender

School	Year 8			Year 9			Year 10		
	Boys	Girls	Total	Boys	Girls	Total	Boys	Girls	Total
1	8.1	27.1	17.5	8.0	18.6	13.2	12.2	16.7	14.5
2	11.9	16.5	14.1	12.9	13.5	13.2	5.8	19.6	12.2
3	15.9	8.3	13.2	14.6	20.0	16.4	12.8	18.5	15.2
4	11.1	15.7	13.4	5.9	16.7	11.2	9.3	17.7	13.1
5	15.0	10.0	12.5	23.1	22.2	22.6	16.3	31.0	22.2
6	0.0	0.0	0.0	20.0	36.4	28.6	10.0	20.0	15.0
7	15.2	11.1	13.3	9.5	9.1	9.3	-	-	-
8	5.6	20.0	12.8	4.3	22.4	14.3	15.1	20.8	17.8
9	21.9	30.0	25.0	7.1	30.4	17.6	11.1	40.0	25.0
10	12.4	25.0	18.1	12.7	10.4	11.8	10.5	13.3	11.8
11	23.0	23.6	23.3	14.3	18.2	16.1	9.8	13.1	11.5
12	20.7	15.0	18.4	14.8	25.0	19.6	4.8	16.7	9.1
13	11.1	27.6	18.3	6.9	9.8	8.2	10.3	18.6	13.8
14	9.1	37.0	20.5	4.8	43.8	18.1	10.1	44.7	22.4
15	9.8	15.0	12.3	19.6	36.8	27.4	9.8	29.2	20.2
16	0.0	8.3	3.0	5.3	16.7	9.7	18.2	35.7	25.0
17*	-	13.6	13.6	-	18.5	18.5	-	-	-
18*	13.0	28.0	20.8	4.3	25.0	15.7	4.3	29.4	15.0
19*	11.5	-	11.5	7.4	-	7.4	6.0	-	6.0
20*	11.7	-	11.7	10.4	-	10.4	8.3	-	8.3
21*	3.6	14.3	8.2	5.4	21.3	12.6	15.4	19.6	17.3
22*	24.0	23.5	23.8	4.8	15.4	10.6	19.0	12.5	15.6
23*	11.1	20.0	14.9	10.0	11.5	10.7	9.5	17.2	14.0
24*	10.2	21.1	14.4	27.5	13.3	21.9	24.5	25.5	25.0
25*	7.7	19.6	13.9	7.9	20.9	14.8	18.6	27.8	23.7
26*	-	12.9	12.9	-	24.1	24.1	-	11.8	11.8
Total	11.6	20.1	15.4	10.6	18.9	14.3	11.1	21.9	15.6

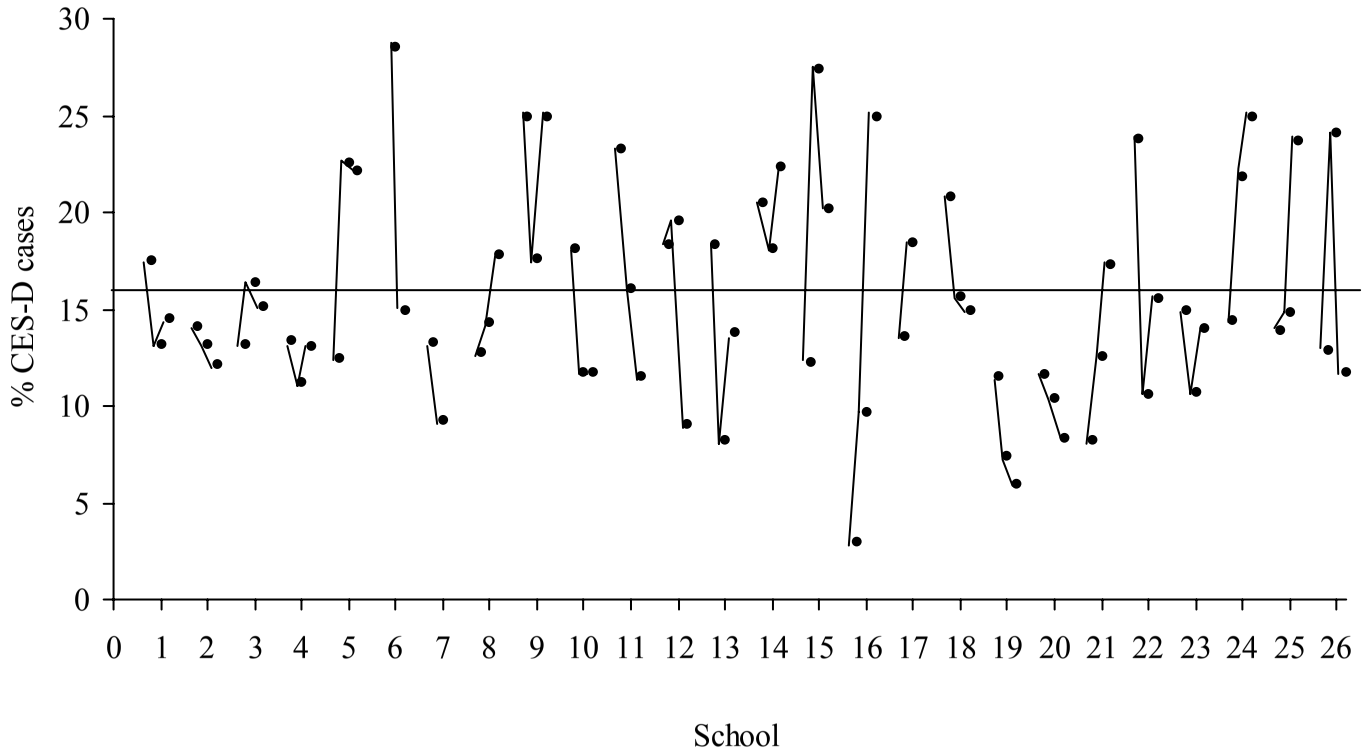


Figure 17 Percentage of high scoring CES-D cases: Year 8, 9 and 10

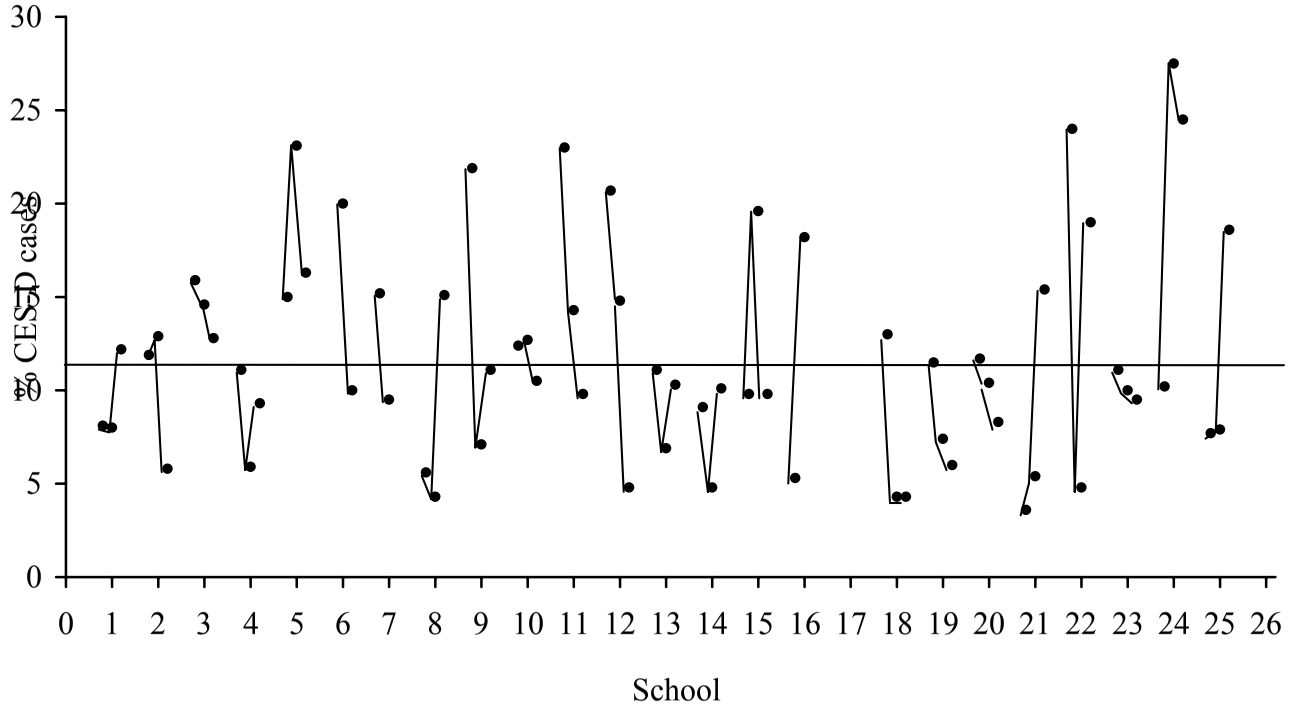


Figure 18 Percentage of high scoring CES-D cases: Year 8, 9 and 10 (Boys)

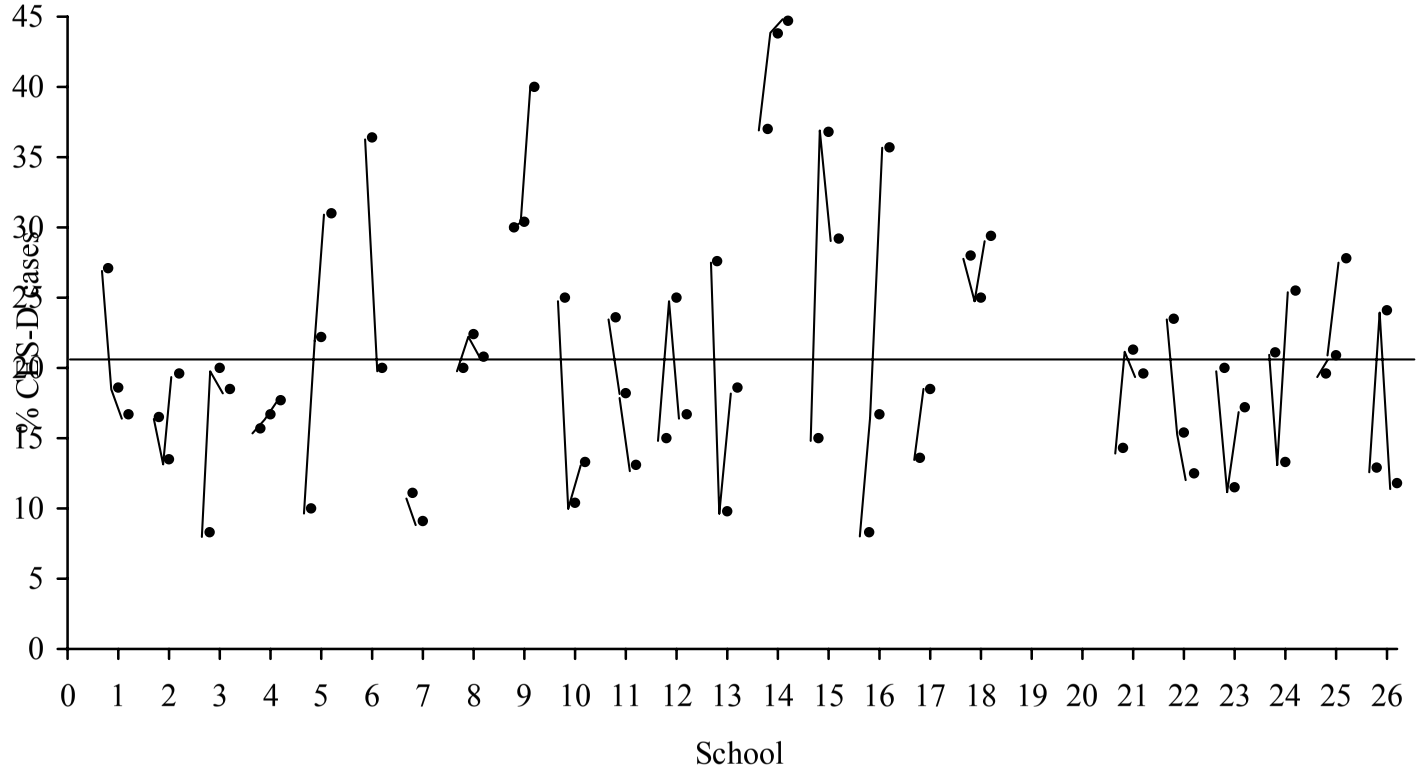


Figure 19 Percentage of high scoring CES-D cases: Year 8, 9 and 10 (Girls)

HLM analyses of school effects on student CES-D scores (continuous)

For the analysis of possible school effects on student levels of depressive symptomatology, multilevel modelling techniques specifically developed for hierarchical or clustered data are used. For each year level cohort, a two level fully unconditional hierarchical model is estimated for boys and girls separately and then combined. These models are referred to as HLM null models because predictors are not specified at either the student or school level. They are the simplest possible hierarchical modes and are equivalent to a one-way ANOVA with random effects.

The results from these models provide the foundation for examining the amount of variability associated with the two levels: between students within schools (Level-1) and between schools (Level-2). The multilevel models estimated in this section are shown using HLM equation notation.

Two-level HLM linear null model for CES-D scores (continuous)

Level-1 (Student) $Y_{ij} = \beta_{oj} + r_{ij}$

where

Y_{ij} is the CES-D score of a student.

β_{oj} is the intercept or mean CES-D score for all students in schools.

r_{ij} is the random student effect, normally distributed with a mean of zero and a variance of σ^2.

The indices *i*, and *j* denote students and schools respectively where there are :

i = 1, 2, …, *n* students within schools *j*.

j = 1, 2, …, J schools.

Level-2 (School) $\beta_{oj} = \gamma_{oo} + \mu_{oj}$

where

γ_{oo} is the grand mean.

μ_{oj} is the random effect associated with each school set at a mean of zero and a variance of $\tau_{00.}$

The EDED comprises data collected from students once in first year (Year 8), once in second year (Year 9) and finally once in third year (Year 10) of high school. In this study schools are used as their own controls and no attempt is made to adjust statistically for student background factors. School variation from data provided by the Year 8 students is assumed to reflect differences in the characteristics of students at intake. At Year 9 and Year 10, school variation will reflect both differences between student background characteristics and between schools. The difference between the Year 8 school variation and the variation at Year 9 and 10 is taken to reflect the school effect on student depressive symptomatology.

The HLM models estimated in the present study therefore treat the data as if it were collected from three separate cross-sectional studies and the sample sizes correspond to the total number of students at each time period (Year 8: 2306; Year 9: 2275; Year 10: 2158).

Table 59 Mean CES-D score: HLM linear models by year level and gender

	Year 8			Year 9			Year 10		
	Boys	Girls	Boys + Girls	Boys	Girls	Boys + Girls	Boys	Girls	Boys + Girls
Grand Mean	11.47	13.75	12.50	10.73	13.35	11.91	10.18	13.94	11.81
Std. Error	0.24	0.48	0.21	0.25	0.34	0.21	0.24	0.36	0.21
Level-2 variance (τ_{00})	0.148	2.200	0.778	1.524	3.697	1.185	0.492	3.420	1.735
Level-1 variance (σ^2)	78.829	129.556	103.180	76.159	112.559	95.314	72.187	119.002	97.602
Intraclass corr.coefficient	0.0019	0.0167	0.0075	0.0196	0.0318	0.0123	0.0068	0.0279	0.0178
χ^2	28.08	38.25	40.45	45.70	54.01	52.63	27.96	44.71	58.10
Probability	0.21	0.02	0.03	<0.01	<0.01	<0.01	0.18	<0.01	<0.01
Reliability	0.09	0.39	0.37	0.47	0.54	0.48	0.49	0.51	0.57

Nine HLM two level linear models are estimated with students at Level-1 nested within schools at Level-2 for each gender and year level. The results from these analyses are shown in Table 59.

Taking Year 8 (Boys & Girls combined) by way of example the results are interpreted as follows. The grand mean estimate for CES-D scores is 12.50 with a standard error of 0.21 indicating a 95% Confidence Interval of 12.09 – 12.91. The estimated variance of the true school means around the grand mean is 0.778 and at the student level it is 103.180. These estimates show that most variability in CES-D scores is at the student level.

The intraclass correlation coefficient represents the proportion of variance in CES-D scores between schools. The formula given by Bryk and Raudenbush (1992, p. 95) is as follows:

$$\frac{Variance\,explained(level1)}{Variance\,explained(hierarchical)} \approx \frac{\tau_{00}}{\tau_{00}+\sigma^2} = \rho$$

where

τ_{00} is the level 2 variance – covariance component.

σ^2 is the level 1 variance.

ρ is the intraclass correlation.

Using the data for Year 8 (Boys & Girls combined) in Table 59 by way of example, the intraclass correlation coefficient is calculated as 0.0075 (0.778/(0.778+103.180)) or 0.75%. This indicates that at Year 8 less than 1 per cent of variance in CES-D scores is at the school level and the vast majority is at the student level.

The hypothesis that all schools have the same mean is rejected ($\chi^2 = 40.45$, $p = 0.03$) providing evidence of statistically significant variability among school mean CES-D scores.

The overall reliability of school sample means as estimates of true school means is given (Bryk & Raudenbush, 1992, p. 63) as:

$$\frac{Parameter\,variance}{Parameter\,variance + error\,variance} = \frac{\hat{\tau}_{00}}{\left(\hat{\tau}_{00} + \left(\hat{\sigma}^2 / \pi_j\right)\right)} = \hat{\lambda}_j$$

where

$\hat{\tau}_{00}$ is the parameter variance.

$\hat{\sigma}^2$ is the error variance.

π_j is the school sample size.

$\hat{\lambda}_j$ is the reliability.

Using the data for Year 8 (Boys & Girls combined) in Table 59 the reliability figure of 0.37 indicates that sample means are only moderately reliable as indicators of true school means.

From Table 59 it can be seen that the intraclass correlation coefficients for the three year levels (with both girls and boys combined) are as follows: Year 8: 0.0075; Year 9: 0.0123; Year 10: 0.0178. The intraclass correlation coefficient at Year 8 indicates the variability between high schools at intake resulting from differences between students. The increased intraclass correlation coefficient found across the Year 9 and Year 10 cohorts is consistent with a school effect on CES-D scores.

An overall school effect can be approximated by subtracting the Year 10 Level 2 variance estimate from the Year 8 Level 2 variance estimate to give an estimated school effect of 0.957 – just under 1 per cent.

The intraclass correlation coefficients calculated at the three year levels represent the theoretically maximal amount of the total variability in CES-D scores that is explainable by **all** school factors. In other words, all possible school level variables such as size of school, school environment or educational leadership, in total in the present study, cannot account for any more than 2 per cent and more likely only 1 per cent of variance in levels of student CES-D scores.

Also shown in Table 59 are the results from HLM linear analyses for each of the three year levels by boys and girls separately. Consistent with the earlier results for both boys and girls the majority of variability in CES-D scores is at the student level with the intraclass correlation coefficient at most (Girls: Year 9) 3 per cent. For boys the hypothesis that all schools have the same mean is not able to be rejected at Year 8 and Year 10 but for girls significantly different school means are found at each of the year levels.

The intraclass correlation coefficient across year levels does not increase in a linear fashion. For both boys and girls there is a marked increase between Year 8 and 9 but then a decrease from Year 9 to 10. The reliabilities of school sample means as estimates of true school means for both boys and girls are moderate with the notable exception of boys at Year 8 where the reliability is only 0.09.

HLM analyses of school effects on student CES-D scores (dichotomous)

In order to examine the possibility that school differences might be more evident for high levels of depressive symptomatology nine additional HLM analyses are performed using the binary dependent variable CES-D22. This variable is coded 1 for high scorers (22 or above on the CES-D) or 0. For these analyses HLM models for binary outcomes (analogous to logistic regression) are estimated.

A very readable account of HLM logistic models is provided by Rumberger (1995) who analysed the factors associated with dropping out (a binary outcome) from high school. These models are shown below using HLM equation notation.

Two-level HLM logistic null model for CES-D scores (dichotomous)

Level-1 (Student)

$$\text{Prob}(Y_{ij} = 1|\beta_j) = \phi_{ij}$$

where

Y_{ij} is the probability of a student scoring 22 CES-D points or higher.

β_j is the intercept of mean proportion of students scoring 22 CES-D points of higher in schools.

ϕ_{ij} is the Bernoulli variance.

The indices *i*, and *j* denote students and schools respectively where there are :

$i = 1, 2, \ldots, n$ students within schools *j*.

$j = 1, 2, \ldots,$ J schools.

Level-2 (School) $\beta_{oj} = \gamma_{oo} + \mu_{oj}$

where

β_{oj} is the intercept or mean CES-D score for all students in schools.

γ_{oo} is the grand mean.

μ_{oj} is the random effect associated with each school set at a mean of zero and a variance of τ_{00}.

This model is interpreted in a similar fashion to the model shown earlier with the main difference being that a logit link function is required to constrain the predicted outcome values to lie within an interval of zero and one. Because of the non-linear link function, output from these HLM models contain what are termed 'unit specific' and 'population average' estimates. In the present study because no predictors are specified in the models the differences between these estimates are trivial and the population average estimates are reported. Table 60 shows the results from nine two level multilevel logistic models.

Taking Year 8 (both boys and girls combined) by way of example the results are interpreted as follows. The grand mean estimate for the proportion of CES-D cases is 15.29. The estimated variance of the true school means around the grand mean is 0.0278 and using the value of ((pi*pi)/3) or 3.2898 for the fixed Level-1 estimate (Snijders & Bosker, 1999) the intraclass correlation coefficient representing the proportion of variance in CES-D scores between schools is calculated as 0.0084 (0.0278/(0.0278+3.2898)) or 0.84%.

The hypothesis that all schools have the same proportion of high scoring cases is not rejected ($\chi^2 = 36.32$, $p = 0.07$). This indicates that the difference between schools in terms of their proportion of high scoring cases is not statistically significant. The overall reliability of school proportions of cases as estimates of true school proportions is 0.23 indicating that the sample estimates are not very reliable.

From Table 60 it can be seen that the intraclass correlation coefficient for the three year levels are as follows: Year 8 – 0.0084 (0.84%); Year 9 – 0.0198 (1.98%); Year 10 – 0.0240 (2.40%). Assuming that the intraclass correlation coefficient at Year 8 indicates the variation between high schools at intake resulting from differences between students, the increased intraclass correlation coefficient found across the Year 9 and Year 10 cohorts is consistent with a school effect on the proportion of high scoring CES-D cases. Also shown in Table 60 are the results from six separate HLM logistic analyses for each of the three year levels by boys and girls separately.

Table 60 Proportion of CES-D cases: HLM logistic models by year level and gender

	Year 8			Year 9			Year 10		
	Boys	Girls	Boys + Girls	Boys	Girls	Boys + Girls	Boys	Girls	Boys + Girls
Grand Mean	11.64	19.75	15.29	10.64	19.78	14.78	11.31	22.03	15.94
Std. Error	1.11	1.11	1.07	1.14	1.12	1.08	1.11	1.12	1.09
Level-2 variance (τ_{00})	0.0322	0.0620	0.0278	0.1643	0.1154	0.0663	0.0411	0.0909	0.0809
Level-1 variance (σ^2)	3.2898	3.2898	3.2898	3.2898	3.2898	3.2898	3.2898	3.2898	3.2898
Intraclass corr. coefficient	0.0097	0.0185	0.0084	0.0476	0.0339	0.0198	0.0123	0.0269	0.0240
χ^2	30.53	33.03	36.32	41.07	38.99	42.48	24.79	33.30	43.12
Probability	0.14	0.08	0.07	0.01	0.02	0.02	0.31	0.04	<0.01
Reliability	0.14	0.28	0.23	0.41	0.40	0.39	0.17	0.37	0.45

For both boys and girls the intraclass correlation coefficient increases between Year 8 and 9 but then decreases from Year 9 to 10. This finding is consistent with the trend for both boys and girls shown in the earlier analyses with the CES-D when analysed as a continuous variable. The reliabilities of school sample estimates for both boys and girls are at best moderate and in particular the estimate for boys at Year 8 (0.14) and Year 10 (0.17) would probably better be described as poor.

The interpretation of the intraclass correlation coefficient derived from a HLM logistic model is not clear cut because this value is a variance at the latent level which does not translate across to the variance of a observed dichotomous variable (Snijders, 1999). Although these variance estimates do not translate across to observed dependent variables it has been argued that there is some value in examining the changes to these estimates across different models (Feiveson, 2001).

Clearly the issue of the correct interpretation of an intraclass correlation coefficient from HLM binary model is complex and beyond the scope of the present research to resolve. In response to the difficulty, a further set of HLM models are estimated, this time using the CES-D22 binary outcome variable in an HLM linear model. This is analogous to using a ordinary regression analysis with an binary outcome variable instead of the generally more appropriate logistic procedure.

Table 61 shows the results from this series of repeated analyses using HLM linear models. Interestingly the pattern of results is very similar to those obtained from the HLM logistic models. First, the intraclass correlation coefficients for the three year levels increases in a manner consistent with a school effect on the proportion of high scoring CES-D cases. Second, for both boys and girls there is an increase in intraclass correlation coefficients between Year 8 and 9 but then a decrease from Year 9 to 10.

The magnitude of the intraclass correlation coefficients from the logistic and linear models are clearly not the same but proportionally the increase (and decrease) across year levels is similar. For example the results from the HLM logistic model showed the intraclass correlation coefficient increasing by a factor of 2.36 from Year 8: 0.0084 to Year 9: 0.0198. Consistent with this, the results from the HLM linear model show the intraclass correlation coefficient increasing by a factor of 2.50 from Year 8: 0.0036 to Year 9: 0.0090.

HLM analyses of school type on student CES-D scores (continuous)

In this section differences between the public and private schools in the EDED sample are examined. Table 62 shows the results from six HLM null models (see earlier HLM linear null model equation). These models are estimated separately for each of the three year levels (Year 8, 9, & 10) and by each of the two school types (public and private). Table 62 shows that the difference in means between public and private schools is not large at intake (Year 8) or either Year 9 or Year 10. No consistent change in intraclass correlation coefficients is evident for public schools with intraclass correlation coefficients increasing between Year 8 (0.0097) and Year 9 (0.0143) but then decreasing in Year 10 (0.0118). For private schools a more consistent pattern is evident with increasing intraclass correlation coefficients across Year 8 (0.0022), Year 9 (0.0116) and Year 10 (0.0331).

In order to test whether public and private school mean levels of depressive symptomatology are different a series of HLM models with school type as a Level-2 predictor are calculated. These models are shown below using HLM equation notation.

Two-level HLM model for test of school type (public – private) on CES-D scores (continuous)

Level-1 (Student) $Y_{ij} = \beta_{oj} + r_{ij}$

where

Y_{ij} is the CES-D score of a student.

β_{oj} is the intercept or mean CES-D score for all students in schools.

r_{ij} is the random student effect, normally distributed with a mean of zero and a variance of σ^2.

The indices *i*, and *j* denote students and schools respectively where there are :

i = 1, 2, …, *n* students within schools *j*.

j = 1, 2, …, J schools.

Level-2 (School) $\beta_{oj} = \gamma_{oo} + \gamma_{o1}$(School type) + μ_{oj}

where

γ_{oo} is the grand mean.

γ_{o1}(School type) is the Level-2 predictor (School type).

μ_{oj} is the random effect associated with each school set at a mean of zero and a variance of $\tau_{00.}$

In addition, HLM linear models are estimated with school type as a Level-2 predictor and gender as a Level-1 predictor. Adding gender as a Level-1 predictor controls for the possibility that the proportion of boys and girls might vary between school types or across year levels. These models are shown below using HLM equation notation.

Two-level HLM model for test of school type (public – private) on CES-D scores (continuous): controlling for gender

Level-1 (Student) $Y_{ij} = \beta_{oj} + \beta_{1j}$ (Gender) + r_{ij}

where

Y_{ij} is the CES-D score of a student.

β_{oj} is the intercept or mean CES-D score for all students in schools.

β_{1j} (Gender) is the Level-1 predictor (Gender).

r_{ij} is the random student effect, normally distributed with a mean of zero and a variance of σ^2.

Table 61 Proportion of CES-D cases: HLM logistic models by year level and gender

	Year 8			Year 9			Year 10		
	Boys	Girls	Boys + Girls	Boys	Girls	Boys + Girls	Boys	Girls	Boys + Girls
Grand Mean	11.64	19.74	15.28	10.64	19.83	14.80	11.32	22.07	15.94
Std. Error	1.00	1.50	0.87	1.29	1.69	1.01	1.02	1.79	1.14
Level-2 variance (τ_{00})	0.00036	0.00164	0.00047	0.00186	0.00311	0.00110	0.00043	0.00296	0.00145
Level-1 variance (σ^2)	0.10224	0.15943	0.13017	0.09325	0.15041	0.12180	0.09851	0.16573	0.13019
Intraclass corr. coefficient	0.0035	0.0102	0.0036	0.0196	0.0203	0.0090	0.0043	0.0175	0.0110
χ^2	31.34	33.38	36.08	45.42	42.77	44.90	25.96	36.07	44.12
Probability	0.12	0.08	0.07	<0.01	<0.01	<0.01	0.25	0.02	<0.01
Reliability	0.15	0.29	0.23	0.47	0.43	0.41	0.18	0.40	0.46

Table 62 Mean CES-D score: HLM null linear models by year level and school type

	Year 8		Year 9		Year 10	
	Public	Private	Public	Private	Public	Private
Grand Mean	12.63	12.01	12.35	11.65	11.88	12.15
Std. Error	0.36	0.41	0.39	0.46	0.38	0.67
Level-2 variance (τ_{00})	0.980	0.247	1.409	1.060	1.171	3.074
Level-1 variance (σ^2)	99.982	110.934	97.453	90.311	98.401	89.714
Intraclass corr. Coefficient	0.0097	0.0022	0.0143	0.0116	0.0118	0.0331
χ^2	29.44	9.01	37.01	16.26	29.95	28.59
Probability	0.01	0.44	<0.01	0.06	0.008	<0.01
Reliability	0.46	0.13	0.53	0.43	0.51	0.67

The indices *i*, and *j* denote students and schools respectively where there are :

i = 1, 2, …, *n* students within schools *j*.

j = 1, 2, …, J schools.

Level-2 (School) $\beta_{oj} = \gamma_{oo} + \gamma_{o1}(\text{School type}) + \mu_{oj}$

where

γ_{oo} is the grand mean.

γ_{o1}(School type) is the Level-2 predictor (School type).

μ_{oj} is the random effect associated with each school set at a mean of zero and a variance of τ_{00}.

The results from these models are shown in Table 63. For all intents and purposes the analyses with and without gender produced similar results and for brevity only the results from the analyses with gender included are discussed. At each year level the difference in school type means between public and private schools is not statistically significant (Year 8: $t = -1.14$, $p = 0.26$; Year 9: $t = -0.39$, $p = 0.69$; Year 10: $t = 1.12$, $p = 0.26$). Gender on the other hand is highly significant at each year level with *t* ratios around five.

Table 63 Mean CES-D score: HLM linear models test of school type by year level

	Year 8	Year 9	Year10
Controlling for School type			
Grand Mean	12.41	12.07	11.99
Std. Error	0.26	0.30	0.35
School type			
Coefficient	-0.62	-0.68	0.21
Std. Error	0.53	0.61	0.79
T ratio	-1.17	-1.13	0.26
Probability	0.26	0.27	0.80
Controlling for School type and gender			
Grand Mean	12.44	11.89	11.86
Std. Error	0.27	0.30	0.29
Gender			
Coefficient	2.26	2.62	3.80
Std. Error	0.48	0.57	0.49
T ratio	4.69	4.58	7.79
Probability	<0.01	<0.01	<0.01
School type			
Coefficient	-0.61	-0.25	0.64
Std. Error	0.54	0.63	0.57
T ratio	-1.14	-0.39	1.12
Probability	0.26	0.69	0.26

Summary

The main findings from the results presented in this chapter can be summarised as follows:

The variability in student CES-D scores is much greater within schools than it is between schools.

At intake to high school the CES-D intraclass correlation coefficient is less than 1 per cent and although this increases to Year 9 and Year 10 consistent with a school effect the size of this effect is very small.

When possible school effects on student CES-D scores are analysed for boys and girls separately both show an increase in intraclass correlation coefficients between Year 8 and Year 9 but then a decrease from Year 9 to Year 10.

A series of HLM logistic analyses examine possible school effects on the proportion of high CES-D scoring students and produce essentially the same results to those outlined above.

Differences between public and private high schools with respect to student mean levels of depressive symptomatology are not statistically significant at either intake (Year 8) or at Year 9 or Year 10.

School sample mean CES-D scores are only moderately reliable as indicators of true school means and sample estimates of each schools' proportion of high scoring students (using a cut-point CES-D score) are not very reliable.

10

Discussion

In this chapter, the results from the present study are discussed following the sequence in which the chapters with results are presented. The findings from basic analyses using descriptive statistics are examined first, followed by a detailed review of the general psychometric properties of the CES-D derived from the IRT analyses. In the process of developing the most appropriate measurement model to be used in the SEM invariance analyses, a considerable number of CFAs are performed. These results provide new information about the factor structure of the CFA and have important implications for future CES-D measurement invariance studies. Finally, an integrated summary at the item level of the IRT and SEM gender and year level DIF analyses is presented. In the following concluding chapter to this report the research findings are summarised, several limitations are acknowledged and the methodological and substantive implications from the study are explored.

Descriptive statistics

Although the CES-D has been widely used in the United States, very little Australian research has been published for large community samples of young adolescents. This study found an average total CES-D total score across boys and girls and year levels of 12.08 (SD = 9.97). This overall mean total CES-D score is lower than that obtained from studies of high school students in different parts of the United States: South Carolina: mean = 15.6 (Garrison et al., 1991b), Connecticut: mean = 16.7 (Tolor & Murphy, 1985), Oregon: mean = 17.0 (Roberts et al., 1991), rural Southern communities: mean = 17.16 (Doerfler et al., 1988), and Boston: mean = 14.98 (Gore et al., 1992).

The sample of students (aged between 13 to 15 years of age) in this study is younger than samples in many American studies and this may possibly account for the lower levels of depressive symptomatology. Importantly, and consistent with other studies using adolescent samples, the average total CES-D score is higher for girls (mean = 13.67, SD = 11.10) compared with boys (mean 10.80, SD = 8.76). A gender effect size is calculated to be 0.26 and this is similar to the gender effect sizes which are

calculated (see Table 1) for other CES-D adolescent samples: 0.24 (Doerfler et al., 1988); 0.25 (Gjerde et al., 1988); 0.23 (Roberts et al., 1991) and 0.23 (Sheeber et al., 1997).

Simple descriptive statistics examining gender differences at the item level show that the mean value of all CES-D items (with two exceptions) are higher for girls than for boys. The two exceptions are for Item 7 (*Effort*) and Item 15 (*Unfriendly*). For these items, mean values are less for girls than boys. The girl mean values for Item 17 (*Cry*) and Item 2 (*Appetite*) are nearly double the corresponding values for boys. The mean value for Item 18 (Sad) is ninth highest for girls but only fourteenth highest for boys suggesting that this item is more salient to girls. In addition, analyses at the response option level found that girls have a greater propensity to acknowledge the presence of the following symptoms: *Bothered* (1), *Blues* (3), *Good* (4), *Depress* (6), *Sleep* (11), *Lonely* (14), *Sad* (18) and *Dislike* (19).

Broadly, these descriptive results are in line with the findings from the two previous in depth studies of CES-D item gender differences. Between them, Clark et al. (1981) and Roberts et al. (1990a) using descriptive statistics identified gender discrepancies for the items: *Bothered* (1), *Appetite* (2), *Effort* (7), *Failure* (9), *Unfriendly* (15), *Cry* (17) and *Sad* (18). In the present study most of these items (but not Item 9 (*Failure*)) are also identified as showing possible gender DIF on the basis of large differences in mean values, differences in item mean rank orders or differences in response option distributions. These preliminary results at the descriptive level suggest that the gender differences observed in the present dataset are typical of CES-D gender differences found in other samples.

In the present study overall levels of depressive symptomatology decrease slightly across Year 8 to Year 10 (ages 13 to 15) but the gender effect size actually increases: Year 8: 0.20; Year 9: 0.24; Year 10: 0.34. Several key studies have found that levels of depressive symptomatology increase around the ages of 13 to 15 years primarily as a result of adolescent girls reporting higher levels of depressive symptomatology (Ge et al., 1994; Petersen, Sarigiani & Kennedy, 1991; WichstrØm, 1999). On the other hand exceptions to this finding are common with, for example, Seiffe-Krenke and Stemmler (2002) showing no increase in levels of depressive symptoms (as measured by the *Child Behavior Checklist*) across the ages of 14 to 17 years and a decrease in the gender effect size over time.

One possible explanation for the decrease in overall levels of depressive symptomatology between Years 8 to 10 is a possible confounding effect in the analyses from using year level (or grade) as a proxy for age. This seems unlikely given that grade level has been shown to be more closely associated with internalising symptoms than age (Prescott, 1998). Subtle changes in the sample across year levels may have occurred with higher rates of attrition in the EDED program among students with higher levels of depressive symptomatology. Because of difficulties in matching students across year levels, this possibility cannot be ruled out but it is also true that the number of students participating in each wave of the survey remained fairly stable. It is also possible that students' scores were reduced in response to repeated exposure to the questionnaire as a screening instrument.

General psychometric properties of the CES-D

Previous analyses of the CES-D in adolescent samples using traditional statistical techniques have reported that CES-D total scores are positively skewed due to the

endorsement of a large number of Option 0 item responses. The present analyses confirm these findings with total score skewness values across gender and year level in the order of 1.30 to 1.60 (see Table 10) and values of skewness for many items (see Table 37) greater than 1.50. The internal consistency of the CES-D as measured by Cronbach's coefficient alpha (Boys: 0.87; Girls: 0.92) in the present sample is similar to the reported alphas obtained from samples of North American high school students (e.g. Roberts et al., 1990a: Boys: 0.88; Girls: 0.91).

For the most part, item to total score correlations are acceptable (> 0.40) but with the notable exceptions of Item 2 (*Appetite*) (Boys: 0.36; Girls: 0.46), Item 4 (*Good*) (Boys: 0.38; Girls: 0.51), Item 7 (*Effort*) (Boys: 0.10; Girls: 0.27) and Item 8 (*Hopeful*) (Boys: 0.28; Girls: 0.36). Consistent with Clark et al. (1981) the majority of items (with one exception), show higher item-scale correlations for girls compared with boys. The exception is Item 15 (*Unfriendly*) which shows a higher item-scale correlation for boys (0.49) compared with girls (0.43).

The non-parametric IRT analyses performed in this study provide a great deal of new information about the psychometric properties of the CES-D when used with young adolescent samples. This is because previous studies of the psychometric properties of the CES-D in adolescent samples have used traditional statistical techniques. Traditional techniques are limited because they produce statistics which represent an average across levels of individual variation and they do not take into account the fact that scale performance may vary across different levels of the target trait (in this case 'depressive symptomatology'). New information about the CES-D from the IRT analyses is provided at both the item and at the scale level. The results at the item level are discussed first.

Using non-parametric IRT techniques the relationship between individual item responses and the construct defined by the item response function (depressive symptomatology) is examined for each item. The resulting option characteristic curves (OCCs) and item characteristic curves (ICCs) show that for many items, the ranges in which options are endorsed can be easily identified and OCCs change rapidly with changes in levels of depressive symptomatology. In IRT terms, these items can be said to be effective and the ICCs for these items exhibit a recognisable S-shaped trace line. On this basis the items: *Depress* (6), *Happy* (12), *Lonely* (14), *Enjoy* (16) and *Sad* (18) are classified as very effective items. A group of items: *Blues* (3), *Good* (4), *Mind* (5), *Talk* (13), *Dislike* (19) and *Getgoing* (20) are found to be somewhat satisfactory but fail to discriminate sharply across a narrow band of depressive symptomatology.

Two types of problems are evident with the remaining items. The first problem is that seven items are dominated by the response option zero. These items are: *Bothered* (1), *Appetite* (2), *Failure* (9), *Fearful* (10), *Sleep* (11), *Unfriendly* (15) and *Cry* (17). For these items, across low to moderate levels of depressive symptomatology the probability that Option 0 is endorsed is greater than 50 per cent. It is of note that three of these items: *Bothered* (1), *Appetite* (2) and *Sleep* (11) relate to somatic symptoms. This suggests that somatic complaints might be particularly important for adolescents with high levels of depressive symptomatology and reflect a more serious form of depressive symptomatology. This empirical finding lends support to the notion which is emphasised in the ICD-10 and to a lesser extent in the DSM-IV classification systems that somatic depressive symptoms have special clinical significance.

The second type of problem is that two items: Item 7 (*Effort*) and Item 8 (*Hopeless*) discriminate across only low levels of depressive symptomatology and have a pronounced lack of discrimination in the moderate to high range of depressive

symptomatology. These two items are consistently identified as showing poor psychometric properties across the descriptive, IRT and SEM analyses. Both items suffer low item to total score correlations and the shape of the ICCs for these items are distinctly different and poorer from all other items. SEM analyses indicate (using a nested second order factor model) that factor loadings to a general factor of depressive symptomatology for these items are less than 0.50. These low factor loadings are in contrast to the factor loadings for the majority of items around 0.70 or higher (see Table 34).

This study is not the first to identify problems with Item 7 (*Effort*). Numerous studies have reported either low item to total score correlations or poor factor loadings for this item in adolescent (see Dick et al., 1994) and adult (see Beals et al., 1991; Cheung & Bagley, 1998; Clark et al., 1981; Liang et al., 1989; Orme et al., 1986; Wong, 2000) samples. In older adult groups the psychometric properties for this item appear better (see Callahan & Wolinsky, 1994; McCallum et al., 1995) but a more systematic examination of the literature is required to confirm this impression. On the basis that low energy or fatigue is explicitly recognised in both the DSM-IV and ICD-10 classification systems this item would appear to have good face validity. The fact that Item 7 (*Effort*) therefore consistently shows poor psychometric properties is perplexing. Any further development of the CES-D could consider whether this item adequately captures the fatigue associated with high levels of depressive symptomatology.

Only one study appears to have reported a problem for Item 8 (*Hopeful*). Callahan and Wolinsky (1994) found that many elderly patients with multiple chronic illnesses have difficulty in responding to this item. These authors argue (quite plausibly) that a lack of hope for the future in elderly patients might simply be a realistic appraisal rather than an indication about their level of depressive symptomatology. In the present sample, students do not have difficulty responding to this item and rates of missing data are not noticeably higher than for other items. This suggests that different factors are responsible for the poor performance of this item in the present community sample of healthy adolescents.

There has been considerable research interest in the relationship between pessimism or hopelessness about the future, depressive symptomatology and youth suicide. The role of hopelessness in depression among adolescents is not clearly understood (see Hankin, Abramson & Siler, 2001 for discussion) and although a general correlation is expected it seems unlikely that hopelessness is related in a simple linear fashion to depressive symptomatology (Young et al., 1996). The present IRT results show that levels of hopelessness are sensitive to changes in depressive symptomatology but only at very low levels of depressive symptomatology. This means for example that individuals with total CES-D scores of around 20 to 40 are likely to have approximately similar Item 8 (*Hopeless*) scores.

Speculatively the present results for Item 8 (*Hopeless*) are consistent with the view that cognitive factors such as the so called ‘personal fable’ (e.g. believing that accidents will not happen to them) in adolescence act as a buffer in the relationship between depression and hopelessness. It should be noted that this interpretation is based around a measure of hopelessness which in effect comprises a single CES-D item. In addition, previous CES-D item analyses have not identified any peculiarities with Item 8 (*Hopeless*). Nonetheless the finding that both Item 7 (*Effort*) and Item 8 (*Hopeless*) appear to share such similar and distinctive ICCs is interesting and worthy of further research.

At the scale level the IRT results clearly show that the reliability and the amount of information provided by the CES-D varies considerably across different levels of depressive symptomatology. At median levels of depressive symptomatology (total scores of around 10) the reliability of the CES-D and the relative amount of information provided by the CES-D is high. The reliability and the amount of information provided by the CES-D decreases rapidly between scores of 10 to 20 and for scores of between 20 and 30 the reliability and information provided is at its lowest.

This pattern of results is at odds with the only other test information function (TIF) of the CES-D provided by Santor and Ramsay (1998). In a sample of American college students (mean age 20 years) these authors showed that the CES-D provided relatively more information in the moderate to severe range of depressive symptomatology compared with the amount of information provided at low levels of depression. The present results suggest that the opposite may be true in young adolescent samples. This implies that the CES-D is more suited to estimating mean levels of depression in community adolescent population groups compared with discriminating between individuals experiencing moderate to high levels of depressive symptomatology.

The reason for this can be seen at the item level from an examination of the steepness of ICCs across CES-D total scores of 20 (approximately the 75th percentile) and 40 (approximately the 90th percentile). Seven items show very effective discrimination in this range. These items are: *Bothered* (3), *Depress* (6), *Failure* (9), *Lonely* (14), *Sad* (18), *Dislike* (19), *Getgoing* (20). These items show more than one and a half CES-D points difference between total scores of 20 to 40. On the other hand, several items show very poor discrimination across this range of depressive symptomatology including: *Effort* (7), *Hopeful* (8) and *Unfriendly* (15) with the remaining items being only modestly effective.

The present TIF finding has important implications for the use of the CES-D as a screening tool in school based mental health intervention programs. For the optimal performance of a screening or diagnostic tool, maximum information should be provided around the diagnostic or screening cut-points (Cooke, Michie, Hart & Hare. 1999). Screening cut-points for the CES-D when used in adolescent samples are typically set in the moderate range of depressive symptomatology. The results from the present study indicate that this is not where the CES-D provides its maximum information. Maximum information is provided around median (low) levels of depressive symptomatology indicating that the CES-D is best suited to estimating group mean levels of depression in normal adolescent populations.

Factor structure of the CES-D

CFAs of the CES-D in the present study replicate and extend previous factor studies carried out in samples of North American adults. A number of competing factor models are tested including two (Cheung & Bagley, 1998) and three (Beals et al., 1991; Ying, 1998) factor models that combine the Depressed Affect and Somatic factors and a five factor model based on the traditional four factor model along with a fifth so-called 'Selfworth' factor. The traditional four factor model is found to provide the best fit to the data. These results suggest that the traditional four factor model although originally derived from data collected from adults is also valid for young adolescents.

The finding that a four factor model (with oblique or correlated factors) provides a good account of the pattern of correlations between CES-D items should not be interpreted as indicating that the CES-D comprises four different subscales. The CES-D was designed to provide a single total score measure of depressive symptomatology and it was not intended by its developer that the CES-D would specifically assess four different facets (Depressed Affect, Positive Affect, Somatic and Interpersonal) of depression (Helmes & Nielson, 1998). In this respect the key issue is the extent to which the CES-D exhibits unidimensionality.

Second–order CFA is a widely used technique for examining the extent to which a psychological scale exhibits unidimensionality. Previous second–order CFAs of the CES-D with adult samples (viz. Hertzog et al., 1990; McCallum et al., 1995) have provided evidence for the existence of a well defined general factor that accounts for a large proportion of total common CES-D item variance. The present study extends these analyses through the use of a nested factor model. This model proposed by Gustafsson (1992) is similar in form to the Schmid and Leiman (1957) second-order parameterisation performed by McCallum et al. (1995) but has the added benefit of allowing the variation not accounted for by the common factors (residual error) to be estimated.

Consistent with previous second-order factor studies the present results reveal the presence of a general well defined single dominant factor (depressive symptomatology) that accounts for around one half of the total CES-D item variance. The three specific (or subsidiary) factors: Somatic (2%) Positive Affect (4%) and Interpersonal (2%) in total account for less than 10 per cent of total variance. Overall, the existence of a strong single dominant general factor provides good support for the unidimensionality of the CES-D in a young adolescent sample. The results from the second-order factor analyses therefore provide empirical justification for the use of the total CES-D score as a measure of depressive symptomatology in younger adolescent populations.

The existence of a general well defined single dominant factor (depressive symptomatology) that accounts for most CES-D item variance has important implications for the specification of the model required for measurement invariance testing. Separate analyses of the nested factor model by gender reveal that the finding of a general well defined single dominant factor (depressive symptomatology) that accounts for most CES-D item variance is valid for both boys and girls. This result therefore provides initial justification for proposing that a one factor model be specified as the measurement model for the further SEM invariance tests.

In the present study, in light of the nested factor results, a one factor measurement model forms the basis for the SEM gender and year level invariance tests. The first measurement invariance test (configural invariance) performed with the one factor measurement model examines whether this proposed model applies to the data in each of the groups (Boys & Girls; Year 8, Year 9 & Year 10). The results from these tests show that the single factor measurement model fits the data well across groups and on this basis it is concluded that a one factor CES-D model provides for configural invariance across gender and year level groups.

Establishing configural invariance for the CES-D using the one factor model is important because it indicates that boys and girls and students across year levels (Year 8 to Year 10) are employing the same conceptual frame of reference to the construct (depressive symptomatology) that is hypothesised to underlie CES-D items. Previous gender and age SEM analyses have not addressed the issue of configural invariance for the CES-D and these studies proceed to perform tests of metric

invariance (equal factor loadings) on the assumption that configural invariance is met. This study provides the first concrete evidence of gender and year level configural invariance for the CES-D.

Given that this study shows that a one factor CES-D model provides configural invariance across gender and year levels what recommendations can be made about the most appropriate factor model to be used in any future measurement invariance studies. It has been long known (see Cronbach, 1951) that if a scale comprises a general factor as well as several smaller group factors then it is likely that the general factor will account for most of the scale score variance. Consistent with this view, this study and previous studies using second-order factor models (e.g. McCallum et al., 1995) have confirmed that a general well defined single dominant factor (depressive symptomatology) accounts for most CES-D item variance. On this basis there is a clear argument for an initial preference towards a one factor model in future SEM measurement invariance studies.

It can be noted that most previous SEM measurement invariance studies have specified a multidimensional (three or four factors) measurement model for the CES-D (but see Breithaupt & Zumbo, 2002 for a notable exception). A better strategy for future CES-D measurement invariance studies might be to begin by performing separate second-order factor analyses to test for the unidimensionality of the CES-D in each comparison group. If, as is likely, these analyses show that a common trait runs through CES-D items in each group then this finding provides initial support for the use of a one factor measurement model. The general strategy being recommended here can be illustrated by reference to a hypothetical study of CES-D measurement invariance across different cultural groups.

Assume that the hypothetical cross cultural investigation has as its main focus the question of whether Asian people make more dysphoric responses to the Positive Affect items compared to people from Western cultures. If the results from the present study and previous CES-D second-order factor studies are generalisable then a primary factor of depressive symptomatology will account for most item variance in both Western and Asian groups. It is also likely that a single factor measurement model will fit the data well across groups. Having established configural invariance with the one factor model, more stringent metric and scalar invariance tests can then be made.

If Asian people give more dysphoric responses to the Positive Affect items for equivalent levels of depressive symptomatology this will be evident with lower factor loadings or thresholds. Lower factor loadings indicate that these items are less strongly correlated with depressive symptomatology and the lower thresholds will show that Asian people are more likely to acknowledge the presence of a lack of positive affect (remembering the items are reverse coded) at lower levels of depressive symptomatology. The interpretation from the one factor measurement model therefore is quite straightforward but if a four factor measurement model had been specified problems abound.

Using a four factor CES-D model, for example, the positive affect items load to the Positive Affect factor and we are left trying to interpret the meaning of differences on Positive Affect item measurement parameters for equivalent levels of positive affect. Previous researchers have avoided this problem by focussing on differences in factor correlations but this overlooks the further complication that if groups vary in their factor variances then factor correlations are also expected to vary (Liang et al., 1989; Widaman & Reise, 1997). Therefore, on both conceptual and empirical grounds good arguments can be mounted for at least an initial preference for a one factor measurement model in future SEM CES-D measurement invariance studies.

This is the first study to provide an estimate of CES-D residual variance. Residual variance is the variance not accounted for by the latent variables in a model. For boys, 42 per cent and for girls 35 per cent of item variance is unable to be explained by the factors in the hypothesised CES-D model. This finding detracts from the otherwise positive view of the CES-D and the relatively high level of variability attributable to error is clearly of concern. Further studies are required to establish whether this rather high level is unique to the CES-D or would be evident in other self-report depression scales. It is also possible (although unlikely) that the result is peculiar to the present sample and might not be evident in other young adolescent or adult samples.

Consistent with the view that the CES-D emphasises the measurement of depressed affect, preliminary modelling indicated that the factor loadings to the specific factor of Depressed Affect (with the influence from the general factor partialled out) were zero. This finding implies the general factor is approximately equivalent to the construct measured by the Depressed Affect items. In contrast, around 4 per cent of variation remains with the Positive Affect specific factor indicating that these items are not purely antonyms to the depressed affect items. This explains why simplified factor solutions to the CES-D which combine the positive and negative affect items, loaded to one factor, as proposed by some researchers (e.g. Callahan & Wolinsky, 1994; Cheung & Bagley, 1998), provide poorer fit compared with models that maintain the distinction between these two groups of items.

Gender measurement invariance

In this section the results from the descriptive, IRT and SEM gender analyses are integrated and discussed. Most items showing DIF serve to increase total scores for girls. These items are: *Bothered* (1), *Appetite* (2), *Sleep* (11), *Cry* (17) and *Sad* (18). The SEM analyses (but not the IRT analyses) also identify Item 3 (Blues) and Item 4 (Good) as increasing scores for girls. Two items serve to increase scores for boys. These items, Item 7 (Effort) and Item 12 (Happy), are identified from both SEM and IRT analyses. In addition, Item 15 (Unfriendly) is identified in the SEM analysis (but not the IRT) as increasing scores for boys. These results indicate that many items in the CES-D show gender DIF in the younger adolescent age group.

First, these results are discussed in the context of the findings from the previous three SEM analyses and the one IRT analysis of CES-D gender measurement invariance. The SEM studies (Beals et al., 1991; Roberts et al., 1990a and Stommel et al., 1993) performed tests of metric invariance (equivalence of factor loadings) across gender. The IRT analysis (Gelin & Zumbo, 2003) used a mixture of ordinal regression and IRT methods to identify gender DIF. Between them these research groups detected six items: *Appetite* (2), *Effort* (7), *Hopeful* (8), *Happy* (12), *Talk* (13) and *Cry* (17) as failing to demonstrate metric invariance. The results from the present study confirm these findings for four of these items but not for Item 8 (Hopeful) or Item 13 (Talk).

Table 64 shows in summary format items exhibiting DIF across the Descriptive, IRT and SEM gender analyses. A large proportion of items show DIF in the current young adolescent population and overall the results are consistent across the different types of analyses. Seven items are identified as showing DIF through the IRT analyses and these same items (along with three additional items) are also identified from the SEM analyses. Importantly, and bolstering confidence in the results, the direction of the DIF (increasing scores for boys or increasing scores for girls) is consistent between the two sets of analyses.

Most items showing DIF serve to increase total scores for girls. These items are: *Bothered* (1), *Appetite* (2), *Sleep* (11), *Cry* (17) and *Sad* (18). The SEM analyses (but not the IRT analyses) also identify Item 3 (*Blues*) and Item 4 (*Good*) as increasing scores for girls. Two items serve to increase scores for boys. These items, Item 7 (*Effort*) and Item 12 (*Happy*), are identified from both SEM and IRT analyses. In addition, Item 15 (*Unfriendly*) is identified in the SEM analysis (but not the IRT) as increasing scores for boys. These results indicate that many items in the CES-D show gender DIF in the younger adolescent age group.

First, these results are discussed in the context of the findings from the previous three SEM analyses and the one IRT analysis of CES-D gender measurement invariance. The SEM studies (Beals et al., 1991; Roberts et al., 1990a and Stommel et al., 1993) performed tests of metric invariance (equivalence of factor loadings) across gender. The IRT analysis (Gelin & Zumbo, 2003) used a mixture of ordinal regression and IRT methods to identify gender DIF. Between them these research groups detected six items: *Appetite* (2), *Effort* (7), *Hopeful* (8), *Happy* (12), *Talk* (13) and *Cry* (17) as failing to demonstrate metric invariance. The results from the present study confirm these findings for four of these items but not for Item 8 (*Hopeful*) or Item 13 (*Talk*).

Table 64 Summary of gender invariance analyses

	Descriptive	IRT	SEM
1 Bothered	-	**G**	**G**
2 Appetite	**G**	**G**	**G**
3 Blues	-	-	**G**
4 Good	-	-	**G**
5 Mind	-	-	-
6 Depress	-	-	-
7 Effort	**B**	**B**	**B**
8 Hopeful	-	-	-
9 Failure	-	-	-
10 Fearful	-	-	
11 Sleep	-	**G**	**G**
12 Happy	-	**B**	**B**
13 Talk	-	-	-
14 Lonely	-	-	-
15 Unfriendly	**B**	-	**B**
16 Enjoy	-	-	-
17 Cry	**G**	**G**	**G**
18 Sad	**G**	**G**	**G**
19 Dislike	-	-	-
20 Getgoing	-	-	-

B: serves to increase scores for boys; G: serves to increase scores for girls

Item 8 (*Hopeful*) was identified by Gelin and Zumbo (2003) (but not by the other researchers) following a mixed ordinal logistic regression and IRT analysis. This item was found to increase scores for males in a large community sample of elderly people. In the present data the raw gender means for this item are very similar and both the IRT and SEM failed to reveal any problems with this item whatsoever. Importantly, the gender DIF in this study was only detected when the CES-D was scored in a binary fashion (zero indicating the symptom was not present, one indicating the symptom was present for at least some of the time) and no DIF for this item was identified when the CES-D was scored conventionally as it is in the present study.

In a similar fashion Item 13 (*Talk*) was identified by Stommel et al. (1993) (but not by the other researchers and the present study) as increasing scores for men in a sample of adult cancer patients. The present results reveal very little that would indicate a problem with Item 13 (*Talk*). Factor loadings (see Table 40) for Item 13 (*Talk*) are quite similar across gender (Boys: 0.64; Girls: 0.66) and when a model is estimated constraining thresholds to be equivalent this results in very little deterioration to model fit compared with a model in which thresholds are allowed to vary across gender.

From the IRT results a visual inspection of the ICC for Item 13 (*Talk*) (see Figure 3) shows that for moderate to high levels of depressive symptomatology boys do indeed show slightly higher item scores than girls. The DIF statistic of 0.017, however, indicates that overall, for the entire sample, the magnitude of this DIF is minor. The cause of the discrepancy between Stommel et al. (1993) and the present study therefore is unclear but the very different samples (adult cancer patients versus a community sample of adolescents) is one possible explanation.

The findings for the remaining four items: *Appetite* (2), *Effort* (7), *Happy* (12) and *Cry* (17) are consistent with the previous gender measurement invariance studies and in the present study show strong and unequivocal evidence of DIF from both the IRT and SEM analyses. The direction of the DIF is also consistent with the previous research with Item 2 (*Appetite*) and Item 17 (*Cry*) increasing scores for girls while Item 7 (*Effort*) and Item 12 (*Happy*) increases scores for boys. What then of the remaining items that are identified in the present study but which have not been found to exhibit DIF by previous researchers?

These newly discovered DIF items are: *Bothered* (1), *Blues* (3), *Good* (4), *Sleep* (11), *Unfriendly (15)* and *Sad* (18) and with the exceptions of Item 15 (*Unfriendly*) they serve to increase scores for girls. One obvious reason additional items were found to exhibit DIF in the present study is because a more comprehensive analysis of measurement invariance is undertaken. Previous studies tested only for metric invariance (equivalence of factor loadings) but in the present study scalar invariance (equivalence of thresholds) is also examined.

For four of the newly discovered items: *Blues* (3), *Good* (4), *Sleep* (11) and *Sad* (18) metric invariance is established but threshold differences are found. These threshold differences indicate that girls (for the same level of depressive symptomatology) are more likely to acknowledge the presence of the symptoms reported in the items: *Blues* (3), *Good* (4), *Sleep* (11) and *Sad* (18). Two of the newly discovered items: *Bothered* (1) and *Unfriendly (15)* fail to demonstrate metric invariance in the present study and theoretically these items could have also been identified in the previous studies.

With respect to Item 15 (*Unfriendly*) it can be noted that Stommel et al. (1993) dropped this item (because of its poor psychometric properties) and Beals et al.

(1991) used this item as a marker item (to the Interpersonal factor) in their analysis. Consequently metric invariance was not tested for Item 15 (*Unfriendly*) by these two groups of researchers. In addition, Beals et al. used Item 1 (*Bothered*) as the marker item to a combined Depressed Affect – Somatic factor and so this item was also not tested for metric invariance.

In the three previous SEM CES-D gender invariance studies the problem of which item to employ as the reference marker was not mentioned. The choice of marker item, however, is quite important because this item sets the standard by which a possible lack of invariance in the remaining items is judged. In Beals et al. (1991) the first item (lowest numbered) was arbitrarily chosen. A three factor measurement model was specified and so three marker items (one to each factor) needed to be chosen. The items chosen were: *Bothered* (1), *Good* (4) and *Unfriendly* (15). The present results suggest that these items were poor choices to serve as marker items. Too few details are provided by Roberts et al. (1990a) or Stommel et al. (1993) to identify the marker items used for their analyses.

Given the methodological difficulties inherent in previous gender SEM studies, arguably the present results have much to commend them. Prior to accepting this proposition, however, it is worth reviewing the differences across analyses that can be seen from the summary of results presented in Most items showing DIF serve to increase total scores for girls. These items are: *Bothered* (1), *Appetite* (2), *Sleep* (11), *Cry* (17) and *Sad* (18). The SEM analyses (but not the IRT analyses) also identify Item 3 (Blues) and Item 4 (Good) as increasing scores for girls. Two items serve to increase scores for boys. These items, Item 7 (Effort) and Item 12 (Happy), are identified from both SEM and IRT analyses. In addition, Item 15 (Unfriendly) is identified in the SEM analysis (but not the IRT) as increasing scores for boys. These results indicate that many items in the CES-D show gender DIF in the younger adolescent age group.

First, these results are discussed in the context of the findings from the previous three SEM analyses and the one IRT analysis of CES-D gender measurement invariance. The SEM studies (Beals et al., 1991; Roberts et al., 1990a and Stommel et al., 1993) performed tests of metric invariance (equivalence of factor loadings) across gender. The IRT analysis (Gelin & Zumbo, 2003) used a mixture of ordinal regression and IRT methods to identify gender DIF. Between them these research groups detected six items: *Appetite* (2), *Effort* (7), *Hopeful* (8), *Happy* (12), *Talk* (13) and *Cry* (17) as failing to demonstrate metric invariance. The results from the present study confirm these findings for four of these items but not for Item 8 (Hopeful) or Item 13 (Talk).

Table 64. Two types of discrepancies are evident. First, some items identified as showing DIF from the SEM analyses are not identified as showing DIF from the IRT analyses. Second, some DIF items identified from the complex statistical procedures (IRT or SEM) show little evidence of DIF when examined using simple descriptive statistics. Both of these types of discrepancies are discussed in turn below.

SEM analyses, but not the IRT analyses, identify three items as showing DIF. Two of these items: *Blues* (3) and *Good* (4) serve to increase scores for girls while the third item *Unfriendly* (15) increases scores for boys. The source of this discrepancy lies in the slightly different methods used in each approach to identify whether an item shows DIF. In the IRT analyses a DIF summary statistic is used to make this judgement while in the SEM analyses DIF is identified by the magnitude of differences in the parameter estimates of the measurement model (factor loading and thresholds).

The summary DIF statistic (produced from the TestGraf software package) is defined as a weighted square difference between ICCs across groups with the difference being weighted by the proportion in each group at each of the evaluation points. This statistic therefore captures and emphasises the potential impact DIF might have for group comparisons at the total score level. In the SEM analyses the statistical significance of differences in the parameter estimates of the measurement model are tested and the significance of these differences is not linearly related to their impact at the item score level.

The differences between the IRT and DIF approaches adopted in the present study for identifying DIF can be illustrated by considering the results for Item 4 (*Good*). The IRT ICC for this item (see Figure 3) indicates that girls at high (20 CES-D points or more) levels of depressive symptomatology report higher item scores than boys. The DIF summary statistic, however, is not large (0.051) because the difference in item scores is most pronounced only for a relatively small proportion of the sample. The SEM results on the other hand show that the first threshold for this item is lower for girls indicating that they are more likely than boys (for equivalent levels of depressive symptomatology) to report the presence of this symptom.

Both sets of analyses (IRT & SEM) for Item 4 (*Good*) are consistent therefore in showing that this item increases scores for girls. The IRT summary DIF statistic is small indicating that the impact on item scores across the whole sample is likely to be minor. Interestingly, in the SEM latent mean analysis of the impact on total scores of items showing DIF this is also found to be the case. When the first threshold of Item 4 (*Good*) is allowed to be free (controlling for the DIF) the effect of this is to reduce the latent mean estimate of depressive symptomatology for girls but only by a relatively small amount (see Table 46). In other words, although this threshold difference is statistically significant its impact on latent means is minor.

Of the 10 items identified from the IRT or SEM analyses as showing DIF, for one half of these, gender discrepancies are apparent from even simple descriptive statistics. For example, girl's mean values on Item 2 (*Appetite*) and Item 17 (*Cry*) are more than double the corresponding value for boys. Given that total scores for girls are not anywhere near double those of boys, the high mean value for girls on these items raises the suspicion of DIF. Not unexpectedly therefore strong evidence of DIF (serving to increase scores for girls) is found from the IRT and SEM analyses for these items.

Reassuringly the direction of the potential DIF shown from descriptive statistics is consistent with the IRT and SEM analyses. That is, the direction of the DIF (increasing scores for girls or increasing scores for boys) shown from the IRT or SEM analyses is what would be expected on the basis of the simple descriptive statistics for those items. It is also the case that five items: *Bothered* (1), *Blues* (3), *Good* (4), *Sleep* (11) and *Happy* (12) showing DIF from the more complex analyses are not detected on the basis of the descriptive analyses for those items. Retrospectively for some of these five items potential problems can be seen, but for several items nothing in the results from the descriptive statistics raises any suspicion of DIF.

For example, Item 12 (*Happy*) is identified from the IRT and SEM analyses as showing a high level of DIF – increasing scores for boys. The difference in mean values between boys and girls on this item is negligible (0.78 compared with 0.79) and their respective item mean ranks are identical (item means were the sixth highest for both boys and girls). The IRT analyses (see Figure 3) show that Item 12 (*Happy*) is a very effective item for both boys and girls but at around median CES-D total

scores (10) the expected item score for boys is about one third of a point higher than for girls. The DIF summary statistic for this item is relatively large (0.113) reflecting the fact that the DIF is quite pronounced for a large proportion of the sample.

Consistent with the IRT results for Item 12 (*Happy*), SEM analyses identified that the first threshold for this item is larger for girls (0.18) compared with boys (0.00). This means that boys are more likely to acknowledge the presence of this symptom (for equivalent levels of depressive symptomatology) compared with girls. From these findings it can be clearly seen that the DIF that is evident in Item 12 (*Happy*) would have been impossible to detect from the results of simple descriptive analyses. This illustrates the point that analyses that rely solely on manifest (or observed) variables are not diagnostic of bias or the lack of bias (Meredith & Millsap, 1992).

In summary, one half (10) of the CES-D items are found to exhibit DIF with most (seven) serving to increase scores for girls. A series of latent mean analyses were performed to examine the impact of this DIF at the total score level. These analyses show that overall the gender DIF in the CES-D serves to increase scores for girls but by only around one half of a CES-D point. The substantive implications of this finding are discussed in the next chapter but it is reassuring to recognise that the further examination of CES-D scores by gender are not invalidated by evidence of DIF to an extent that would seriously confound any findings that might emerge.

Year level measurement invariance

The finding from previous research that levels of depressive symptomatology appear to increase around the ages of 13 to 15 years raises the possibility that the meaning of depressive symptomatology might be changing for young people during early adolescence. The results from the present study suggest that this possibility is unlikely. Consistent with results from Roberts et al. (1990a) in a sample of older adolescents the present results suggest that the CES-D at the total score level exhibits a high level of measurement invariance across early adolescent age groups.

Across the IRT and SEM analyses six items are identified as showing a lack of measurement invariance across year levels: *Bothered* (1), *Mind* (5), *Effort* (7), *Lonely* (14), *Unfriendly* (15) and *Getgoing* (20). The magnitude of the DIF at the item level is relatively small and in the SEM latent analyses a model which allows for the DIF produces a latent mean estimate across year levels which is virtually identical to one in which the DIF is uncontrolled. These results indicate that CES-D mean differences across the early secondary high school year levels (Year 8 to Year 10) can be compared with some confidence.

The IRT year level analysis is performed separately for boys and girls. The results are consistent across gender with both Item 5 (*Mind*) and Item 7 (*Effort*) showing signs of DIF. For Item 7 (*Effort*) the interpretation of the DIF is clear cut, with item scores for boys and girls with low levels of depressive symptomatology decreasing between Year 8 and Year 9 and then decreasing again between Year 9 and Year 10. This finding is replicated in the SEM analyses. For Item 5 (*Mind*) the IRT DIF finding is more ambiguous. The ICCs across year level for this item (see Figure 4) shows that item scores for Year 8 and Year 10 are fairly similar across all levels of depressive symptomatology but for Year 9 students with moderate levels of depressive symptomatology they are slightly higher. DIF for this item is also evident from the SEM analyses with a chi-square test of threshold differences only marginally failing to meet the set critical chi-square value.

The SEM year level measurement invariance analyses are performed with the data for boys and girls combined. This is because of sample size restrictions. Both metric (equal factor loadings) and scalar (equal thresholds) measurement invariance is able to be established across Year 8 to Year 9. The Year 10 factor loading for Item 14 (*Lonely*) is significantly lower indicating that this item is less salient to this older age group. In addition, for the Year 10 group the first threshold for the items: *Bothered* (1), *Effort* (7) and *Unfriendly* (15) and the second threshold for Item 20 (*Getgoing*) are found to be significantly different from the Year 8 and Year 9 groups.

The results from the SEM year level analyses illustrate the concepts of 'DIF amplification' and 'DIF cancellation'. Roughly the same number of items serve to increase scores across year levels as those which serve to decrease scores across year levels. Consequently, when these items are combined the effect of the DIF at the scale level is minimal. In a SEM latent mean analysis, the differences between year level latent mean estimates produced from a model controlling for the five items showing DIF are virtually identical to the latent mean estimates produced from a model which did not control for the DIF. This shows that the existence of items with DIF in a test does not prove that total scores from the test are necessarily biased.

School effects on student depression

This study found statistically significant differences between school mean levels of student depressive symptomatology. Overall, however the school differences are less than expected if it were true that schools play a major role in shaping student mental health (Booth & Samdal, 1997) and are very small by comparison with the variation found between students. The absence of substantial differences between schools and the high student variation within schools is apparent both from the results of basic descriptive statistics and the more complex HLM analyses. In the present sample the size of the variation in average student CES-D depression scores between the high schools is small at intake (Year 8), and although this variation increases over Year 9 and Year 10, consistent with a school effect, the size of this is certainly not more than 2 per cent and more likely is around 1 per cent.

These findings suggest that the combined effect of intake differences between schools and the school effect for adolescent depressive symptomatology is very modest. The results therefore are consistent with the view that adolescent depressive symptomatology is largely driven by individual level psychological factors. This view is contrary to an expectation gathered from previous researchers that differences between school environments exert large impacts on student mental health. For example, Rutter, Giller and Hagell (1998, p. 331) in the context of discussing externalising behaviours state that:

Children spend a high proportion of their waking lives in schools. By their nature, schools constitute social organisations as well as educational establishments. There are major variations among schools in levels of both disruptive behaviour and delinquency, and these variations go well beyond what would be expected on the basis of differences at intake.

The present finding can also be contrasted with studies of school effects on academic achievement which have provided much larger estimates than in the current study for student depressive symptomatology. In a recent systematic meta analysis of Dutch, British and American school achievement studies, Bosker and Witziers (1995) estimated the true net (adjusted for student background characteristics) school effect

to be approximately six per cent. The considerably lower estimate for student depressive symptomatology found in the present study indicates that schools are very similar in their effects on student mood.

Obviously schools focus their attention on educational goals and their outcomes depend partly on how well these teaching programs are implemented in a given school context. In contrast, while nearly all schools will seek to provide protective and enriching environments for students, the prevention of student depression is not a primary goal. Nonetheless the result is contrary to an expectation that the variation in protective and risk factors between schools would produce a much larger school effect. These factors are not directly examined in the present study but it is clear that they are not producing large differential effects across schools which could be detected with a commonly used screening instrument for depressive symptomatology.

It is important to note that the measure of variation reported in the present study is based on the intraclass correlation coefficient. This statistic measures the extent of clustering in groups (see Keeves, 1997 for discussion) and is related to changes in mean values of a construct over time. In the present case the intraclass correlation coefficient reflects the degree to which students in each of the 26 schools are more like each other in terms of their level of depressive symptomatology than they are like the members of the other schools in the sample. Because students are not randomly allocated to schools, at intake (Year 8) some clustering was expected. As it turned out this was minimal and even though students shared many common experiences within each school, over three years the degree to which they became more like each other was also minimal.

11

Conclusions

This report began by arguing that adolescent depressive symptomatology is a major area of current psychological inquiry and the CES-D is a cornerstone for a large part of this research. It is widely believed that during adolescence overall levels of depressive symptomatology increase markedly from childhood and that female adolescents have higher levels of depressive symptomatology than male adolescents. Several theories have been developed to explain these gender and age differences and not an inconsiderable amount of research effort has been directed to garnering empirical support for them. Given the amount of research and theoretical effort directed towards explaining observed gender and age differences in CES-D scores it is vital that the psychological community has confidence in the measuring properties of this scale.

Crucially, however, there is very little evidence to support the notion that the CES-D measures the same thing across gender and early adolescence and it is possible that the observed CES-D total score difference across these groups is simply an artefact of the measuring process itself. The first general aim of the present study therefore is to investigate whether CES-D scores obtained from boys and girls across early adolescence can be validly compared (measurement invariance). The second general aim of the present study is to examine whether schools exert effects on student levels of depressive symptomatology independently of, or in addition to, individual level characteristics. This second general aim seeks at a broad level to test the notion that differences between school social environments are important for shaping student mental health.

In order to investigate possible CES-D gender, age and school effects the present study uses data collected from a large scale mental health screening program (Early Detection of Emotional Disorders: EDED) carried out in South Australia between 1994 and 1997. The present author played a key role during the data collection phase of EDED and gained an intimate knowledge of both the strengths and weakness of the EDED dataset. The number of students who participated in the EDED program is quite large and this enabled the application of relatively new statistical approaches that have not yet been applied to the CES-D.

The new statistical approaches (namely non-parametric IRT models, SEM for ordinal variables and HLM to examine school effects) became widely available during the late 1990s and have allowed the present researcher to undertake a more thorough test of the measurement properties of the CES-D across gender, age and school groups than has previously been possible. In this concluding chapter the research questions that motivated the present study (outlined in Chapter 3 of this report) are repeated and on the basis of the empirical results a brief summary answer to each question is provided.

Following the brief summary of the research findings, limitations to the present study are acknowledged. These limitations primarily relate to weaknesses in the EDED dataset and the manner in which the dataset is used in the analyses. Despite the limitations it is argued that the present study makes several important methodological contributions to the investigation of measurement invariance in self-report depression rating scales and school clustering of mental health variables which will be of benefit to future researchers. Finally, the substantive implications of the CES-D gender, age and schools are discussed in terms of the measurement of depressive symptomatology in young adolescents and for the types of mental health preventative programs which might be offered in schools.

Summary of Research Findings

1. **What overall levels of depressive symptomatology will be reported by Australian adolescents compared with their American counterparts?** Average total CES-D scores from the present community sample of young Australian adolescents are slightly lower than scores obtained from high school students in United States.
2. **Do girls show higher total CES-D scores than boys and do CES-D scores increase during early adolescence (Years 8 to 10: Ages 13 to 15 years)?** Average CES-D scores are higher for girls compared with boys. Contrary to expectations average CES-D scores decrease slightly across Year 8 to Year 10.
3. **Are gender and year level (age) differences at the total score level reflected in differences at the factor, item and response option level?** Descriptive statistical analyses reveal that the mean value for nearly every CES-D item is higher for girls than for boys. Several items (particularly those relating to Depressed Affect and Somatic symptoms) show a higher proportion of girls endorsing the presence of a symptom (Option 1) compared with boys.
4. **Are individual CES-D item scores equivalent across gender and year level at equal levels of depressive symptomatology?** Using non-parametric IRT models several items show relatively high levels of gender DIF. Most of these items serve to increase scores for girls but two items serve to increase scores for boys. Across year levels little evidence of DIF is apparent with one item showing an increase and one item showing a decrease across year levels.
5. **If item scores are not equivalent across gender and year level, what impact does this have on total scores?** The impact of gender DIF on total scores is estimated (from IRT models) to be around one quarter of a CES-D

point in the direction of artificially raising girls' scores. Across year levels the impact of the DIF is negligible.

6. **Are there gender or year level differences for CES-D items at the response option level, controlling for levels of depressive symptomatology?** For equal levels of depressive symptomatology the IRT analyses show that girls are more likely to endorse the presence (Option 1) of the following symptoms: *Bothered* (1), *Appetite* (2), *Sleep* (11), *Cry* (17), and *Sad* (18). Boys on the other hand are more likely to acknowledge the symptom of *Effort* (7) and report a higher level (more likely to endorse Option 3) of *Happy* (12 – reversed scored).
7. **What is the relative quality of the information provided by the CES-D across different levels of depressive symptomatology?** The quality of the information provided by the CES-D in the present sample is best at around scores of 10. Relatively poorer information is provided in the moderate (20 – 30) range of depressive symptomatology where screening cut-points are typically set.
8. **From the variety of factor models proposed for the CES-D which provides the best fit to the data?** The traditional four factor model provides a better fit to the data compared with a one, two, three or five factor model.
9. **Does the CES-D exhibit unidimensionality in an adolescent population?** The results from a nested factor model show that a general factor of depressive symptomatology accounts for around one half of the variance in item scores. Less than 10 per cent of variance is explained by the specific factors indicating good evidence for the unidimensionality of the CES-D in an adolescent population.
10. **To what extent might previous SEM analyses which have ignored the ordinal nature of the CES-D be in error?** The results from a comparison of ML and WLS estimation techniques suggests that better fitting models with higher factor loadings and correlations are obtained from the WLS approach which properly reflects the ordinal response format of CES-D items.
11. **Do boys and girls and students across year levels employ the same conceptual frame of reference to the construct hypothesised to underlie the CES-D (configural invariance)?** CES-D gender and year level configural invariance is supported on the basis of a one factor CES-D model fitting the data well and with high factor loadings across groups.
12. **Are the CES-D SEM measurement model parameters (factor loadings & thresholds) equivalent across gender and year level (metric & scalar invariance)?** Across gender, differences in factor loadings are found for four items: *Bothered* (1), *Appetite* (2), *Unfriendly* (15) and *Cry* (17) and in addition the first threshold for six items: *Blues* (3), *Good* (4), *Effort* (7), *Sleep* (11), *Happy* (12) and *Sad* (18) and the third threshold for Item 7 (*Effort*) fail to demonstrate scalar invariance. Across year levels the factor loading for Item 14 (*Lonely*) in Year 10, is found not to be invariant. In addition, between Year 8 and Year 10, the first threshold is not invariant for the items: *Bothered* (1), *Effort* (7) and *Unfriendly* (15) and the second threshold is not invariant for Item 20 (*Getgoing).*

13. **Are the CES-D SEM structural model parameters (factor variances and item residual variances) equivalent across gender and year level?** Factor variances appear equal across gender but the residual variances for several items are higher (indicating poorer reliability) for boys than for girls. A hypothesis of equal factor variances across year levels is supported and the reliability for nine items improves across Year 8 to Year 10.
14. **What is the impact of any lack of gender or age measurement invariance on CES-D total scores?** The results from a series of latent mean analyses show that girls exhibit higher levels of depressive symptomatology (the latent construct) than boys and this remains true even when the lack of gender measurement invariance is taken into account. Estimates from the SEM analyses indicate that the impact of gender CES-D DIF translates to around one half of a CES-D point at the total raw score level. Across year levels the CES-D exhibits a high level of measurement invariance and only very minor differences in factor loadings and thresholds are detected.
15. **What is the extent of clustering for school based CES-D data?** The variation in student CES-D scores is much greater within schools than it is between schools indicating very little school based CES-D clustering.
16. **Does the extent of clustering for school based CES-D school increase during the first three years of high school consistent with a school effect on student depressive symptomatology?** At intake to high school the CES-D intraclass correlation coefficient is less than 1 per cent and although this increases to Year 9 and Year 10 consistent with a school effect, the size of this effect is very small.

Limitations

> It is very easy to accept the hypothesis that there is no DIF. To accomplish this one merely has to run poor studies with smallish sample sizes and use weak statistical methods. Thus, to be credible, a claim to have found 'no DIF' must mean a careful study with as large a sample size as could be found that uses the most powerful statistical procedures available to analyze these data. (Holland & Wainer, 1993, p. 31)

A key finding from the present study is that although there is CED-D DIF across gender and year levels the impact of this DIF at the total scale level is rather minor. Cognisant of the concerns expressed by Holland and Wainer (shown above) it can be confidently stated that the present study uses data collected from a large sample of young adolescents and employs the most powerful statistical techniques that are currently available for the analysis of DIF. Nonetheless every study has limitations and the present one is no exception. There are two key areas of limitations to the gender and year level DIF analyses. The first concerns deficiencies in the EDED dataset itself and the second relates to the manner in which the dataset was used in this study.

The sample size of the EDED program is both a strength and weakness to the present study. On the positive side, large numbers of students took part in the program and participation rates were high, around 85 per cent. This quite high response rate was achieved through a passive consent process. Students (or their parents) who did not wish to take part in EDED were required actively to choose not to participate by

returning a form to the school. If this form was not returned it was assumed that consent had been given. Through this type of consent process the student participation rate in the EDED program was considerably higher than the rates typically achieved in other large scale mental health surveys of school students.

As in most other large scale mental health surveys of school students very little is known about the students who did not take part in the program. Anectodately, school staff reported that there were very few so-called 'conscientious objectors' to EDED and that most failures to participate were the result of school absences on the day of the survey or other school engagements such as music lessons at the time the survey. This evidence, combined with the overall high participation rate tends to suggest that a fairly representative sample of school students completed questionnaires. Unfortunately the high student participation rate of the EDED that was achieved through the passive consent process did not translate to an equally high student cooperation rate.

The EDED program was both a research and mental health screening intervention project and students were required to provide personal identifying (initials and dates of birth) on the survey forms. Students were aware that high scores on the questionnaire (which included the CES-D) would be used to identify them to school counsellors. These demand characteristics for faking 'bad' or for faking 'good' might have exerted quite powerful effects on the responses that students gave to the CES-D. To avoid the possibility of being identified a sizeable minority of students failed to record their correct initials and date of birth but otherwise seemed to provide valid responses to the scales comprising the questionnaire.

Students who failed to record accurate identifying information on their questionnaires frustrated attempts to match students longitudinally across the three waves of the program. For the purposes of the present study, however, this was preferable to an alternative strategy adopted by a very small number of students who provided nonsense responses to most of the questionnaire. The impression of EDED program staff was that rates of non-compliance with the identification process were higher for boys than they were for girls, and boys were more likely to provide nonsense responses to the questionnaire. Weaknesses of these sorts in the EDED dataset are inevitable in any large school based mental health survey but arguably are not fatal to the present study.

It is important to remember that overall mean levels of depressive symptomatology reported by the sample are in line with other studies, as is the gender effect size. In addition, the key CES-D analyses reported in the present study are performed at the item level and it is very unlikely that students, either boys or girls, responded differentially to individual CES-D items for reasons of avoiding identification. For example, it seems implausible to entertain that boys would be less likely than girls for the same level of depressive symptomatology to acknowledge the presence of Item 2 (*Appetite*) in the belief that this symptom (as opposed to any other) was more likely to trigger the identification criteria. In summary, while it is appropriate to acknowledge that the EDED dataset used in the present study is not perfect, the present author believes that these deficiencies do not materially affect the findings of the study.

The second main area of potential concern to the findings of the study relates to the manner in which the EDED dataset is used in the analyses. For many DIF analyses students responses are amalgamated across year levels. These analyses therefore proceed as if the sample comprised responses from 2306 Year 8 students, 2275 different Year 9 students and 2158 different Year 10 students. In reality the dataset was clustered at the student level with CES-D responses obtained longitudinally from

approximately the same 2300 students three years in a row. The reason for amalgamating across year levels was that the very large dataset this produced enabled the appropriate SEM estimation techniques (weighted least squares) to be applied.

A drawback to this approach is that the observations in the dataset are not truly independent and this clustering at the student level was ignored. Even though the CES-D is widely regarded as tapping depressive symptoms in adolescents that are quite volatile, transient and fluctuating it is possible that the wave to wave correlation of CES-D total scores (at the individual level) is in the order of 0.60 to 0.70 indicating a substantial level of dependence. The effect of this amalgamation on some of the analyses is likely to be quite mild but on others, particularly those involving variances, standard errors and model fit indices caution is clearly warranted. It is acknowledged that the amalgamation occurred as a last resort and that the findings require replication.

In a similar vein, a further potential criticism of the DIF analyses concerns the fact that the school based clustered nature of the EDED dataset is ignored. This might appear to be quite contradictory given that in the introduction to this thesis it was argued that previous analyses of school student CES-D data might be in error because researchers have ignored this clustering. This criticism is valid although this study has been able to show that the effect of the clustering on standard traditional statistical techniques is most likely very minor. For the advanced IRT and SEM analyses performed in the present study very little is known about the effect of clustering for these types of analyses. At this point of time, and one suspects for some time yet, it is not possible to perform a multilevel multiple group threshold analysis with any mainstream SEM software package.

Several limitations to the analysis of possible school effects on depressive symptomatology should also be acknowledged. The number of schools (26) is relatively small and it is known that estimates of the proportion of variance at the school level may be underestimated in small samples using multilevel techniques (Draper, 1995; Morris, 1995). On the other hand, both schools from the public and private sectors were included in the study and this would normally be expected to increase the estimate of between school variation.

The school sample comprised a non-random sample of schools from one State of Australia and, although there is no reason to doubt that the result would be generalisable to other similar samples of schools, caution is required. The magnitude of the Australian school effect for academic achievement is similar to other Western developed countries (Peaker, 1975; Rowe, Hill & Holmes-Smith, 1995) and by analogy, the results for mental health constructs such as depressive symptomatology may also be similar.

Classroom data were not collected and some recent research evidence suggests that for student achievement (particularly perhaps when measured by teacher ratings rather than standardised tests) the variation between classes may be very much larger than the variation between schools. Although the studies examining this issue have yielded contradictory findings (see discussion by Hill & Rowe, 1996) it is possible that in the present sample the classroom level may have been associated with a larger variation in student depressive symptomatology.

A strength of the present analysis of school effects is the longitudinal design using repeated measures of a cohort of students nested within schools. This design allows students and schools to be used as their own controls and avoided the need to rely on statistically controlling for student background factors (Rowe et al., 1995). It is important to remember however that students in the study were not randomly

allocated to the 26 schools - rather the design is best conceptualised as a quasi-experimental study where each school is seen as a treatment group (Bryk & Raudenbush, 1992).

Methodological conclusions

An important methodological contribution of the present study is that it provides a concrete example of both IRT and SEM techniques to the investigation of DIF in a widely used psychological scale. The SEM analyses performed with M*plus* are unique in the substantive literature and address the main weaknesses identified by influential methodologists in the application of traditional factor analysis to ordinal rating scales. Most importantly, the SEM analyses using M*plus* have allowed a fuller test of CES-D measurement invariance across gender and year levels than has previously been undertaken.

In performing the analyses, the value of using both IRT and SEM as complementary approaches to the investigation of DIF became evident. The IRT analyses with the TestGraf software were very quick and easy to perform and yielded a considerable amount of simple to interpret information about the CES-D at the item and scale level. Items with possible DIF could be identified visually and the reason for this DIF was able to be established at the response option level. The decision to employ a nonparametric IRT technique proved wise because as it turned out many items would not have been modelled efficiently using a parametric approach based around the logistic function. On this basis it can be recommended that non-parametric approaches should be given consideration in any future IRT analyses of the CES-D.

The SEM analyses with M*plus* were quite difficult to undertake with many practical and conceptual problems to be addressed. The M*plus* software is relatively new and although arguably best suited to the questions of this study, it has not yet benefited from many years of testing. The discovery of a minor bug in the program should not be taken to diminish the considerable statistical advances that have been incorporated in M*plus* nor detract from the very friendly user interface. SEM threshold models are also new and cookbook type examples of these analyses are not currently available. For most substantive researchers, informal channels of assistance such as e-mail discussion groups and personal contacts with methodologists specialising in SEM measurement invariance issues will be necessary to perform correctly the analyses.

Given the complexity of the SEM analyses some simple IRT analyses to fall back on proved indispensable. The types of IRT and SEM analyses performed in the present study are also complementary at a more fundamental level. The IRT approach adopted in the present study obtained item parameters from separate calibrations of each group which were then equated using the total score (or more strictly the ML estimate of total score) as a basis for linking the groups. Conversely, the SEM analyses calibrated the items for both groups simultaneously and then linked the two groups based on a marker or anchor item, which was assumed *a priori* to be DIF free.

This is the reason that the 'marker item selection problem' arose for the SEM analyses but not the IRT. Each approach has its advantages and disadvantages (see Orlando & Marshall, 2002) and arguably there is merit in using both in a study of DIF. One clear benefit in the present study was that the results from the IRT analyses were useful, along with theoretical considerations, in guiding the choice of the marker item for the SEM analyses. In addition, the consistency in the results across the analyses indicates that the findings are not dependent on any one approach or

technique. That is, the majority of items found to exhibit strong DIF were identified across both IRT and SEM analyses and this supports a view that the findings are robust.

The problem of analysing nonnormal data with standard factor analysis or SEM techniques is longstanding. This problem is particularly troublesome for depression researchers because when rating scales such as the CES-D, the BDI or the CDI are used in community populations many respondents will be asymptomatic for any symptom in the period under review (in the case of the CES-D in the last two weeks). This means that the data distributions obtained from these instruments will be very skewed. Solutions such as normalising data prior to analysis or using scaled chi-square statistics do not provide a complete solution to the problem.

Until recently the problem of factor analysis with nonnormal data has been viewed as intractable (see Long & Brekke, 1999) and very little research has addressed how serious this problem might be for substantive SEM analyses in real datasets. The present study compared Maximum Likelihood (ML) and Weighted Least Squares (WLS) estimation techniques using two mainstream SEM packages (LISREL and M*plus*). Preliminary normality tests indicate that as expected every CES-D item fails to exhibit univariate normality and therefore the even more strict assumption of multivariate normality is not met. Importantly, it appears that the assumption of bivariate normality required for a LISREL WLS analysis could be met, but this judgement rests on fairly generous criteria proposed by the developer of LISREL based on an unpublished simulation study.

The results from the comparison of the ML technique (which treats CES-D items as continuous variables) and the WLS technique (which treats the CES-D items as ordinal variables) indicate a clear superiority for the WLS approach. Using the four factor model as a basis for testing, WLS estimation produces a better fitting model with higher factor loadings and higher factor correlations compared with ML. Whether previous analyses that have used ML might have arrived at invalid conclusions depends on many considerations including the nature of the questions addressed. At the very least, the present study demonstrates that they have been less than optimal and therefore a sole reliance on the ML technique for future SEM analyses of the CES-D seems unwise.

The threshold model available in M*plus* with WLS estimation allows for a more comprehensive analysis of CES-D measurement invariance than has previously been possible. This is important because several items that demonstrate metric invariance (equal factor loadings) failed to show scalar invariance (equal thresholds) across gender or year level groups. For example, Item 12 (*Happy*) demonstrates gender metric invariance but fails a test of scalar invariance. As discussed earlier, the DIF (with expected item scores for boys about one third of a point higher than for girls at median CES-D total scores) for this item is also evident from the IRT results suggesting that the finding is robust.

It is important to emphasise that the DIF for Item 12 (*Happy*) would not have been identified from the SEM analyses if scalar invariance had not been tested. It is also true that the DIF for this item has not been found in earlier gender SEM analyses that have only tested for metric invariance. A plausible reason therefore why more items are detected in the present study as showing DIF compared with earlier studies is simply because a more comprehensive analysis of measurement invariance is undertaken. This implies that future SEM measurement invariance studies should also include tests of scalar invariance.

From a methodological perspective the analysis of possible school effects on depressive symptomatology contributes to an improved understanding of the extent of clustering in student mental health data and the reliability of a widely used instrument (the CES-D) when used as an aggregate measure of school level student depressive symptomatology. In the present sample of schools the intraclass correlation coefficients are very small indicating that traditional statistical procedures (which do not take into account clustering effects) could have been applied without serious error for the analysis of student (Level 1) background factors.

For the future, however, the ease of use and wide availability of multilevel statistical software places an onus on researchers to consider the improvements offered by this approach for the analysis of school clustered child and adolescent mental health data. The estimates of school depressive symptomatology are only modestly reliable - around 0.50 overall across the three year levels. This reliability is a function of both the psychometric proprieties of the CES-D and the number of students sampled from each school. Future research examining school effects on student depressive symptomatology (using the CES-D) therefore may need to sample a larger number of students per school than was the case in the present study.

Substantive conclusions

Using very powerful statistical techniques in a large community sample of young adolescents the present study finds gender and to a lesser extent age (between 13 to 15 years) DIF in many CES-D items. Most (but not all) items showing gender DIF serve to increase scores for girls and most (but not all) items showing age DIF serve to increase scores for older as opposed to younger adolescents. Despite this DIF total scores across gender and age are relatively unbiased. In the present sample girls show higher overall levels of depressive symptomatology than boys and this remains true even when gender DIF is taken into account. Similarly across age, latent mean analyses show that the impact of year level DIF is very minor and for all intents and purposes the CES-D seems to provide an unbiased measure of depressive symptomatology across early adolescence.

The present age measurement invariance results are consistent with the only other study to examine this issue in a sample of adolescents. This study (Roberts et al., 1990) used a sample of adolescents aged between 15 and 17 years and found that CES-D factor loadings were invariant across age groups. The present study provides an important replication of Roberts et al. (1990) because it has tested for measurement invariance in a sample of young adolescents (13 to 15 years) where measurement invariance problems might be more evident and in addition because more levels of measurement invariance were tested. Based on the present results it appears that the inconsistent findings in the CES-D literature about increases in depressive symptomatology during early adolescence do not stem from a lack of measurement invariance in the scale itself.

Previous research suggests that the largest adolescent gender differences occur in symptoms of affective distress (e.g. Compass, 1997; Craig & Van Natta, 1979; Newmann, 1984; Silverstein et al., 1995). This finding has been taken to indicate that higher female scores might simply reflect the presence of transient symptoms of depressed affect rather than to a depressive syndrome. In the present study the largest observed gender differences also occurred across items comprising the Depressed Affect factor (see Table 10 and Table 12). In itself this is not evidence of bias and as it turns out four out of the seven items showing DIF across both the IRT and SEM

analyses relate to somatic symptoms. Three of these items (Item1: *Bothered*, Item 2: *Appetite* and Item 11: *Sleep*) increase scores for girls with the fourth item (Item 7: *Effort*) increasing scores for boys.

Given that CES-D total scores very much reflect Depressed Affect scores (see the nested factor model) these results suggest that for approximately equivalent levels of depressed affect, girls report higher levels of somatic symptoms (with the notable exception of Item 7: *Effort*) than boys. This finding provides indirect support to the hypothesis that higher levels of female depressive symptomatology result because females exhibit a syndrome of 'anxious somatic depression' – depression accompanied by anxiety and somatic symptomatology (Silverstein et al., 1995). Arguably therefore while it is true that high female total scores are at least partly a function of excess scores with respect to depressed affect this should not be lightly dismissed as reflecting nothing more than the transient symptoms of depressed affect because from the present results the high scores also appear indicative of a depressive syndrome.

More generally the results from the present study support the current practice of the majority of researchers who simply sum all 20 CES-D items and than use the computed total score in further gender analyses. On the other hand caution is warranted where gender analyses are performed with a single CES-D item used as a proxy measure for a particular construct. For example, scores on Item 8 (*Hopeful*) can be used as a measure (albeit a rather crude one) of hopelessness. If the size of any gender effect found in these single item studies is large then given that for most items showing DIF the magnitude of this is rather small then the finding might still be robust. But if the size of the gender effect is small or marginal and it so happened that an item with relatively large gender DIF was employed then it is possible that the result may not be valid.

Earlier it was noted that one research team (Aseltine et al., 1998) has adopted the practice of leaving out Item 17 (*Cry*) in their CES-D analyses. In the present sample girl's raw scores on Item 17 (*Cry*) are on average approximately one third of a CES-D point higher than boys scores. On the basis of the IRT and SEM analyses this study estimates that the size of the gender bias (at the total score level) is equal to around one half of a CES-D point. This means that not including Item 17 (*Cry*) in the calculation of a total score effectively negates most of the gender bias. On the negative side however, this practice will slightly alter the balance between the symptom groups covered by the CES-D and in addition alter the psychometric properties of the scale.

It will be recalled that the CES-D comprises items addressing Depressed Affect (seven items), Positive Affect (four items), Somatic (seven items) and Interpersonal (two items). Consistent with descriptions provided in DSM-IV and ICD-10 that crying (or tearfulness) is indicative of depressed mood, previous CES-D factor analyses have found that this item loads to the Depressed Affect factor. In the present study, using the traditional four factor model, the factor loading of this item to the Depressed Affect factor is a very satisfactory 0.84. To omit Item 17 (*Cry*) item in the calculation of the total score therefore slightly diminishes the CES-Ds emphasis on depressed affect.

Omitting Item 17 (*Cry*) item in the calculation of the total score will also slightly alter the psychometric properties of the CES-D. The present IRT analyses reveal that this item is not very discriminating across low to moderate levels of depressive symptomatology but is effective (for both boys and girls) across high levels of depressive symptomatology. This means that the omission of Item 17 (*Cry*) reduces

the ability of the CES-D to discriminate between individuals experiencing high levels of depressive symptomatology. It can also be noted that other similar self-report depression scales include an item referring to crying. For example, Item 10 in the CDI asks the respondent to report whether they have felt like crying every day, many days, or once in a while during the past two weeks. Similarly, Item 10 in the BDI asks how often respondents have cried during the past week.

The magnitude of the gender DIF, which can be thought of as the amount of systematic error by gender in the CES-D, is minimal compared with the variation in scores attributable to random or chance errors. The lack of measurement precision in the CES-D is brought into stark relief by the results from the IRT analyses. For a student scoring 10 on the CES-D, their true score (using a 95% confidence interval) can only be estimated to be somewhere between 5 to 15. In a similar fashion the SEM analyses showed that less than one half of the variation in CES-D scores is accounted for by the construct of depressive symptomatology. On this argument the gender bias introduced by the majority of CES-D researchers who simply add all 20 items is relatively trivial compared with the lack of precision inherent in the scores themselves.

Although the gender bias in the CES-D might be able to be shown to be practically unimportant it can be noted that in the educational testing literature a very uncompromising attitude is taken towards items showing DIF. For example Berk (1982) argues that even if items with DIF cancel themselves out and only account for a minute proportion of total variance biases of any kind are socially and psychometrically undesirable and should be eliminated. Consistent with this approach large scale educational testing programs such as the Educational Testing Service in the United States implement extensive so-called ‘fairness reviews’ in which test items are examined for inappropriate stereotyping, sexist or racist language and for DIF. New test items for standardised college and graduate admission tests are continually being developed and it is a relatively simple and inexpensive matter for items to be dropped or added to tests.

For psychological scales and tests the question of what to do with items showing DIF is less clear cut. Psychological test revision (such as with the MMPI) is an onerous and costly business (see Butcher, 2000; Reise, Waller & Comrey, 2000; Silverstein & Nelson, 2000 for a discussion). The complexities raised with revising the CES-D are not quite of the same order as those associated with a widely used standardised test such as the MMPI but it is clear from the literature that recommendations made by researchers from single studies to drop or modify items of well established scales are most unlikely to be adopted. For this reason the present author resists the temptation to create a ‘gender balanced’ version of the CES-D destined for obscurity.

Understanding gender differences in adolescent depressive symptomatology is likely to be a continuing area of psychological research. The CES-D is widely used in this research and the present results suggest that by and large it is adequate for this purpose. Having said this it should be recognised that the CES-D was not designed specifically with gender comparisons in adolescent populations in mind. In addition, the CES-D was developed without the benefit of the sophisticated statistical methods that are available today for examining DIF in scales. Without disrespect to the developer of the CES-D, and acknowledging that the construction process at the time was first rate, the fairly clean bill of health given to the CES-D by the results from the present study may owe more to good fortune than to good design.

The present results suggest that the development of a new self-report depression scale specifically designed for adolescents and free from gender DIF might be of value.

The findings from the present study would be of use for this exercise but despite their statistical sophistication, the present results are at the level of descriptive differences. Prior to the development of any new scale future research is needed to improve an understanding of what might cause gender DIF in self-report depression scales. Randomised DIF studies are used by educational researchers to establish why some items show DIF and others do not. These types of experimental DIF studies may be of considerable value for better understanding of gender differences in the expression of depressive symptomatology and aid the development of a gender neutral self-report depression scale.

A further suggestion for future research concerns addressing the current lack of measurement precision in the CES-D. Previous attempts to improve the effectiveness of self-report depression scales have encountered serious limitations but one area that could be profitably explored is computerised adaptive testing (CAT). In CAT different sets of questions are administered to different individuals depending on each individuals level on the trait being measured (Weiss, 1985). For example students with total CES-D scores of 10 would be administered only items which provide maximum information at this trait level. CAT has been adopted in personality testing with some success (see Handel, Ben-Porath & Watt, 1999 for an example with the MMPI) and given the widespread availability of computers in schools, CAT for measuring adolescent depressive symptomatology is worthy of further study.

This is the first study to investigate possible school effects on student levels of depressive symptomatology using HLM techniques. These possible school effects are important for a number of reasons not the least of which is the pressing need for the development of effective prevention programs for childhood and adolescent depression (Kovacs, 1997). Schools are places where whole populations of young people can be accessed easily and present ideal opportunities for preventative mental health programs. For these reasons schools are assuming greater prominence for the delivery of child and adolescent mental health services. Given the critical importance of these efforts, schools need to be offered programs with a high likelihood of success.

The present findings from this study in a natural setting of CES-D school effects suggest that there may be severe limitations on the amount of improvement possible from so-called ‘whole of school’ interventions. This is because school level characteristics appear to have only very weak effects on student depressive symptomatology. Individual level psychological factors on the other hand were found to have by far the greatest influence on depressive symptomatology. On this basis, programs directed towards individual level factors (e.g. example those that teach cognitive and behavioural skills in the classroom) have much greater potential for alleviating depressive symptomatology. Future studies are required to confirm this unexpected finding and ideally they would include classroom data where clustering effects might be stronger.

12

References

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/ 4-18 and 1991 profile*. Burlington VT: University of Vermont, Department of Psychiatry.

Aiken, L. S., Stein, J. A., & Bentler, P. M. (1994). Structural equation analyses of clinical subpopulation differences and comparative treatment outcomes: Characterizing the daily lives of drug addicts. *Journal of Consulting and Clinical Psychology, 62*(3), 488-499.

Aitkin, M., & Longford, N. (1985). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society*, 54, 1-29.

Allgood-Merten, B., Lewinsohn, P. M., & Hops, H. (1990). Sex differences and adolescent depression. *Journal of Abnormal Psychology, 99*(1), 55-63.

Allison, S., Roeger, L., Martin, G., & Keeves, J. (2001). Gender differences in the relationship between depression and suicidal ideation in young adolescents. *Australian and New Zealand Journal of Psychiatry, 35*, 498-503.

Altman, N. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46*, 175-185.

Altmann, E. O., & Gotlib, I. H. (1988). The social behavior of depressed children: An observational study. *Journal of Abnormal Child Psychology, 16*(1), 29-44.

Amemiya, Y., & Anderson, T. W. (1990). Asymptotic chi-square tests for a large class of factor analysis models. *Annals of Statistics, 18*, 1453-1463.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed). Washington, DC: Author.

Andresen, E., Malmgren, J., Carter, W., & Patrick, D. (1994). Screening adults for depression in well older adults: Evaluation of a short form of the CES-D. *American Journal of Preventative Medicine, 10*(2), 77-84.

Andrews, G., Hall, W., Teesson, M., & Henderson, S. (1999). *The mental health of Australians*. Canberra: Department of Health and Aged Care.

Andrews, J. A., Lewinsohn, P. M., Hops, H., & Roberts, R. E. (1993). Psychometric properties of scales for the measurement of psychosocial variables associated with depression in adolescence. *Psychological Reports, 73*(3), 1019-1046.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Andrich, D., & Van Schoubroeck, L. (1989). The General Health Questionnaire: A psychometric analysis using latent trait theory. *Psychological Medicine, 19*(2), 469-485.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Angold, A., & Costello, E. J. (2001). The epidemiology of depression in children and adolescents. In I. M. E. Goodyer (Ed.), *The depressed child and adolescent* (2nd ed., pp. 143-178). New York, US: Cambridge University Press.

Angold, A., Erkanli, A., Silberg, J., Eaves, L., & Costello, E. J. (2002). Depression scale scores in 8-17-year-olds: Effects of age and gender. *Journal of Child Psychology & Psychiatry & Allied Disciplines, 43*(8), 1052-1063.

Angold, A., & Rutter, M. (1992). Effects of age and pubertal status on depression in a large clinical sample. *Development and Psychopathology, 4*(1), 5-28.

Arbuckle, J. L. (1997). *Amos users guide*. Chicago: SmallWaters.

Aseltine, R. H., Jr. (1996). Pathways linking parental divorce with adolescent depression. *Journal of Health and Social Behavior, 37*(2), 133-148.

Aseltine, R. H., & Gore, S. (1993). Mental health and social adaptation following the transition from high school. *Journal of Research on Adolescence, 3*(3), 247-270.

Aseltine, R. H., Gore, S., & Colten, M. E. (1994). Depression and the social developmental context of adolescence. *Journal of Personality and Social Psychology, 67*(2), 252-263.

Aseltine, R. H., Jr., Gore, S., & Colten, M. E. (1998). The co-occurrence of depression and substance abuse in late adolescence. *Development and Psychopathology, 10*(3), 549-570.

Avison, W. R., & McAlpine, D. D. (1992). Gender differences in symptoms of depression among adolescents. *Journal of Health and Social Behavior, 33*(2), 77-96.

Bagozzi, R. P. (1981). An examination of the validity of two models of attitude. *Multivariate Behavioral Research, 16*(3), 323-359.

Barnes, G. E., & Prosen, H. (1985). Parental death and depression. *Journal of Abnormal Psychology, 94*(1), 64-69.

Beals, J., Manson, S. M., Keane, E. M., & Dick, R. W. (1991). Factorial structure of the Center for Epidemiologic Studies--Depression Scale among American Indian college students. *Psychological Assessment, 3*(4), 623-627.

Bebbington, P. E. (1998). Sex and depression. *Psychological Medicine, 28*(1), 1-8.

Bebbington, P. E., Dunn, G., Jenkins, R., Lewis, G., Brugha, T., Farrell, M., & Meltzer, H. (1998). The influence of age and sex on the prevalence of depression conditions: Report from the National Survey of Psychiatric Morbidity. *Psychological Medicine, 28*(1), 9-19.

Beck, A. T. (1976). *Cognitive therapy and the emotional disorders*. New York: International Universities Press.

Beck, A. T., Steer, R. A., & Garbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review, 8*(1), 77-100.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561-571.

Beck, A. T., Weissman, A., Lester, D., & Trexler, L. (1974). The measurement of pessimism: The Hopelessness Scale. *Journal of Consulting and Clinical Psychology*, *42*(6), 861-865.

Bedi, R. P., Maraun, M. D., & Chrisjohn, R. D. (2001). A multisample item response theory analysis of the Beck Depression Inventory-1A. *Canadian Journal of Behavioural Science, 33*(3), 176-185.

Beeber, L. S., Shea, J., & McCorkle, R. (1998). The Center for Epidemiologic Studies Depression Scale as a measure of depressive symptoms in newly diagnosed patients. *Journal of Psychosocial Oncology, 16*(1), 1-20.

Bentler, P. M. (1990). Comparative fix indexes in structural models. *Psychological Bulletin, 107*(2), 238-246.

Bentler, P. M. (1995). *EQS structural equations programs manual*. Encino, CA: Multivariate Software, Inc.

Bentler, P. M., & Wu, E. J. C. (1995). *EQS for Windows: Users guide*. Encino, CA: Multivariate Software Inc.

Berganza, C. E., & Agular, G. (1992). Depression in Guatemalan adolescents. *Adolescence, 27*(108), 771-782.

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*(3), 467-477.

Bird, H. R. (1996). Epidemiology of childhood disorders in a cross-cultural context. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 37*(1), 35-49.

Birmaher, B., Ryan, N. D., Williamson, D. E., Brent, D. A., Kaufman, J., Dahl, R. E., Perel, J. M., & Nelson, B. (1996). Childhood and adolescent depression: A review of the past 10 years, Part I. *Journal of the American Academy of Child and Adolescent Psychiatry, 35*(11), 1427-1439.

Bland, R. C., Orn, H., & Newman, S. C. (1988). Lifetime prevalence of psychiatric disorders in Edmonton. *Acta Psychiatrica Scandinavica, 77*, 24-32.

Blatt, S. J., Hart, B., Quinlan, D. M., Leadbeater, B., & Auerbach, J. (1993). Interpersonal and self-critical dysphoria and behavioral problems in adolescents. *Journal of Youth and Adolescence, 22*(3), 253-269.

Blazer, D. G. (1982). *Depression in later life*. St. Louis, MO: C. V. Mosby.

Block, J. H., Block, J., & Gjerde, P. F. (1986). The personality of children prior to divorce: A prospective study. *Child Development, 57*(4), 827-840.

Block, R. D. (1972). Estimating item parameters and latent ability when responses are sorted in two or more nominal categories. *Psychometrika, 37*, 29-51.

Boey, K. W. (1999). Cross-validation of a short form of the CES-D in Chinese elderly. *International Journal of Geriatric Psychiatry, 14*(8), 608-617.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY, USA: John Wiley and Sons.

Bollen, K. A. & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods and Research*, *21*, 205-229.

Booth, M. L., & Samdal, O. (1997). Health-promoting schools in Australia: Models and measurement. *Australian and New Zealand Journal of Public Health, 21*, 365-370.

Bosker, R., & Witziers, B. (1995). *A meta analytical approach regarding school effectiveness: The true size of school effects and the effect size of educational leadership*. Netherlands: Reports - Research (143).

Breithaupt, K., & Zumbo, B. D. (2002). Sample invariance of the Structural Equation Model and the Item Response Model: A case study. *Structural Equation Modeling, 9*(3), 390-412.

Breslau, N. (1985). Depressive symptoms, major depression, and generalized anxiety: A comparison of self-reports on CES-D and results from diagnostic interviews. *Psychiatry Research, 15*(3), 219-229.

Brooks-Gunn, J., & Warren, M. P. (1989). Biological and social contributions to negative affect in young adolescent girls. *Child Development, 60*(1), 40-55.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In Bollen KA & L. JS (Eds.), *Testing structural equation models* (pp. 136-162). Thousand Oaks, CA: Sage.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park: Sage Publications.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *Hierarchical linear and nonlinear modeling with the HLM/2l and HLM/3l programs*. Chicago: Scientific Software International.

Bryk, A., & Thum, Y. (1989). The effects of high school organisation on dropping out: An exploratory investigation. *American Educational Research Journal, 26*(3), 353-383.

Butcher, J. N. (2000). Revising psychological tests: Lessons learned from the revision of the MMPI. *Psychological Assessment, 12*(3), 263-271.

Byles, J., Byrne, C., Boyle, M. H., & Offord, D. R. (1988). Ontario Child Health Study: Reliability and validity of the General Functioning subscale of the McMaster Family Assessment Device. *Family Process*, *27*(1), 97-104.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466.

Cairns, E., McWhirter, L., Duffy, U., & Barry, R. (1990). The stability of self-concept in late adolescence: Gender and situational effects. *Personality and Individual Differences, 11*(9), 937-944.

Callahan, C. M., & Wolinsky, F. D. (1994). The effect of gender and race on the measurement properties of the CES-D in older adults. *Medical Care, 32*(4), 341-356.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 397-413). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Chambers, W. J., Puig-Antich, J., Hirsch, M., Paez, P., Ambrosini, P. J., Tabrizi, M. A., & Davies, M. (1985). The assessment of affective disorders in children and adolescents by semistructured interview: Testretest reliability of the Schedule for Affective Disorders and Schizophrenia for School-Age Children, Present Episode Version. *Archives of General Psychiatry, 42*(7), 696-702.

Chan, A. C. M. (1996). Clinical validation of the Geriatric Depression Scale (GDS). *Journal of Aging and Health, 8*(2), 238-253.

Chan, D. (2000). Detection of differential item functioning on the Kirton Adaption-Innovation Inventory using multiple-group mean and covariance structure analyses. *Multivariate Behavioral Research, 35*(2), 169-199.

Chapleski, E., Lamphere, J., Kaczynski, R., Lichtenberg, P., & Dwyer, J. (1997). Structure of a depression measure among American Indian Elders: Confirmatory factor analysis of the CES-D scale. *Research on Aging, 19*(4), 462.

Chartier, G. M., & Lassen, M. K. (1994). Adolescent depression: Children's Depression Inventory norms, suicidal ideation, and (weak) gender effects. *Adolescence, 29*(116), 859-864.

Chen, H., Mechanic, D., & Hansell, S. (1998). A longitudinal study of self-awareness and depressed mood in adolescence. *Journal of Youth and Adolescence, 27*(6), 719-734.

Cheung, C. K., & Bagley, C. (1998). Validating an American scale in Hong Kong: The Center for Epidemiological Studies Depression Scale (CES-D). *Journal of Psychology, 132*(2), 169-186.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*(1), 1-27.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross Cultural Psychology, 31*(2), 187-212.

Christensen, H., Jorm, A., Mackinnon, A., Korten, A., Jacomb, P., Henderson, A., & Rogers, B. (1999). Age differences in depression and anxiety symptoms: A structural equation modelling analysis of data from a general population sample. *Psychological Medicine, 29*(2), 325-339.

Clark, D. C., Gibbons, R. D., Haviland, M. G., & Hendryx, M. S. (1993). Assessing the severity of depressive states in recently detoxified alcoholics. *Journal of Studies on Alcohol, 54*(1), 107-114.

Clark, V., Aneshensel, C. S., Frerichs, R., & Morgan, T. M. (1981). Analysis of effects of sex and age in response to items on the CES-D scale. *Psychiatric Research, 5*, 171-181.

Clarke, G., Hawkins, W., Murphy, M., & Sheeber, L. (1993). School-based primary prevention of depressive symptomatology in adolescents. Findings from two studies. *Adolescent Research, 8*(2), 183-204.

Clarke, G. N., Hawkins, W., Murphy, M., Sheeber, L. B., Lewinsohn, P. M., & Seeley, J. (1995). Targeted prevention of unipolar depressive disorder in an at-risk sample of high school adolescents: A randomized trial of group cognitive intervention. *Journal of the American Academy of Child and Adolescent Psychiatry, 34*(3), 312-321.

Clarke, G., Hops, H., Lewinsohn, P. M., Andrews, J., Seeley, J., & Williams, J. (1992). Cognitive-behavioral group treatment of adolescent depression: Prediction of outcome. *Behavior Therapy, 23*(3), 341-354.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cole, D. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology*, *55*(4), 584-594.

Cole, D. A. (1989). Psychopathology of adolescent suicide: Hopelessness, coping beliefs, and depression. *Journal of Abnormal Psychology, 98*(3), 248-255.

Cole, D. A., & Maxwell, S. E. (1985). Multitrait-multimethod comparisons across populations: A confirmatory factor analytic approach. *Multivariate Behavioral Research, 20*(4), 389-417.

Cole, D. A., Maxwell, S., Arvey, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modelling. *Psychological Bulletin, 114*(1), 174-184.

Cole, N. S. (1993). History and development of DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 25-29). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Colten, M. E., Gore, S., & Aseltine, R. H., Jr. (1991). The patterning of distress and disorder in a community sample of high school aged youth. In M. E. Colten & S. Gore (Eds.), *Adolescent stress: Causes and consequences. Social institutions and social change* (pp. 157-180). New York, NY, USA: Aldine De Gruyter.

Compas, B. E. (1987). Stress and life events during childhood and adolescence. *Clinical Psychology Review, 7*(3), 275-302.

Compas, B. E. (1997). Depression in children and adolescents. In E. J. Mash & L. G. Terdal (Eds.), *Assessment of childhood disorders (3rd ed.)* (pp. 197-229). New York, NY, USA: The Guildford Press.

Compas, B. E., Davis, G. E., & Forsythe, C. J. (1985). Characteristics of life events during adolescence. *American Journal of Community Psychology, 13*(6), 677-691.

Compas, B. E., Ey, S., & Grant, K. E. (1993). Taxonomy, assessment, and diagnosis of depression during adolescence. *Psychological Bulletin, 114*(2), 323-344.

Compas, B. E., Oppedisano, G., Connor, J. K., Gerhardt, C. A., Hinden, B. R., Achenbach, T. M., & Hammen, C. (1997). Gender differences in depressive symptoms in adolescence: Comparison of national samples of clinically referred and nonreferred youths. *Journal of Consulting and Clinical Psychology, 65*(4), 617-626.

Cooke, D. J., Michie, C., Hart, S. D., & Hare, R. D. (1999). Evaluating the Screening Version of the Hare Psychopathy Checklist--Revised (PCL:SV): An item response theory analysis. *Psychological Assessment, 11*(1), 3-13.

Costello, C. G. (1982). Social factors associated with depression: A retrospective community study. *Psychological Medicine, 12*(2), 329-339.

Coyne, J. C. (1994). Self-reported distress: Analog or ersatz depression? *Psychological Bulletin, 116*(1), 29-45.

Coyne, J. C., Downey, G., & Boergers, J. (1992). Depression in families: A systems perspective. In D. Cicchetti & S. L. Toth (Eds.), *Rochester symposium on developmental psychopathology. Vol. 4. Developmental perspectives on depression* (pp. 211-249). New York: University of Rochester Press.

Craig, T. J., & Van Natta, P. A. (1979). Influence of demographic characteristics on two measures of depressive symptoms. *Archives of General Psychiatry, 36*(2), 149-154.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

D'Agostino, R. B. (1986). Tests for the normal distribution. In R. B. D'Agostino & M. A. Stephens (Eds.), *Goodness-of-fit techniques.* (pp. 367-419). New York: Marcel Dekker.

Davidson, H., Feldman, P. H., & Crawford, S. (1994). Measuring depressive symptoms in the frail elderly. *Journals of Gerontology, 49*(4), 159-164.

Deardorff, W. W., & Funabiki, D. (1985). A diagnostic caution in screening for depressed college students. *Cognitive Therapy and Research, 9*(3), 277-284.

Devins, G., & Orme, C. (1985). Center for Epidemiologic Studies Depression Scale. In Keyser DJ & S. RC (Eds.), *Test critiques* (Vol. 2). Kansas, MO, USA: Test Corporation of America.

Devins, G. M., Orme, C. M., Costello, C. G., Binik, Y. M., Frizzell, B., Stam, H., & Pullin, W. (1988). Measuring depressive symptoms in illness populations: Psychometric properties of the Center for Epidemiologic Studies Depression (CES-D) scale. *Psychology and Health, 2*(2), 139-156.

Dick, R. W., Beals, J., Keane, E. M., & Manson, S. M. (1994). Factorial structure of the CES-D among American Indian adolescents. *Journal of Adolescence, 17*(1), 73-79.

Dick, R. W., Manson, S. M., & Beals, J. (1993). Alcohol use among male and female Native American adolescents: Patterns and correlates of student drinking in a boarding school. *Journal of Studies on Alcohol, 54*(2), 172-177.

Dierker, L. C., Albano, A. M., Clarke, G. N., Heimberg, R. G., Kendall, P. C., Merikangas, K. R., Lewinsohn, P. M., Offord, D. R., Kessler, R., & Kupfer, D. J. (2001). Screening for anxiety and depression in early adolescence. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*(8), 929-936.

Doerfler, L. A., Felner, R. D., Rowlison, R. T., Raley, P. A., & Evans, E. (1988). Depression in children and adolescents: A comparative analysis of the utility and construct validity of two assessment measures. *Journal of Consulting and Clinical Psychology, 56*(5), 769-772.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47*(2), 309-326.

Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research, 35*(1), 21-50.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. E. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioural Statistics, 20*(2), 115-147.

Dumenci, L., & Windle, M. (1996). A latent trait-state model of adolescent depression using the Center for Epidemiologic Studies-Depression Scale. *Multivariate Behavioral Research, 31*(3), 313-330.

Duncan-Jones, P., Grayson, D., & Moran, P. (1986). The utility of latent trait models in psychiatric epidemiology. *Psychological Medicine, 16*, 391-405.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341-349.

Fava, G. A. (1983). Assessing depressive symptoms across cultures: Italian validation of the CES-D Self-Rating Scale. *Journal of Clinical Psychology, 39*(2), 249-251.

Fechner-Bates, S., Coyne, J. C., & Schwenk, T. L. (1994). The relationship of self-reported distress to depressive disorders and other psychopathology. *Journal of Consulting and Clinical Psychology, 62*(3), 550-559.

Feiveson, A. (2001). *Variance decomposed in bernoulli model - Reply (Email)*. multilevel@mailbase.ac.uk [2001, 1 August].

Feldman, M., & Wilson, A. (1997). Adolescent suicidality in urban minorities and its relationship to conduct disorders, depression, and separation anxiety. *Journal of the American Academy of Child and Adolescent Psychiatry, 36*(1), 75-84.

Ferrando, P. J. (1996). Calibration of invariant item parameters in a continuous item response model using the extended Lisrel measurement submodel. *Multivariate Behavioral Research, 31*(4), 419-439.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286-299.

Funder, D. C. (1993). Judgments as data for personality and developmental psychology: Error versus accuracy. In D. C. Funder & R. D. Parke (Eds.), *Studying lives through time: Personality and development. APA science volumes* (pp. 121-146). Washington, DC, USA: American Psychological Association.

Gallo, J. J., Anthony, J. C., & Muthén, B. O. (1994). Age differences in the symptoms of depression: A latent trait analysis. *Journals of Gerontology, 49*(6), P251-P264.

Garland, A., & Zigler, E. (1993). Adolescent suicide prevention: Current research and social policy implications. *American Psychologist, 48*(2), 169-182.

Garrison, C. Z., Addy, C. L., Jackson, K. L., McKeown, R. E., & Waller, J. L. (1991a). A longitudinal study of suicidal ideation in young adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 30*(4), 597-603.

Garrison, C. Z., Addy, C. L., Jackson, K. L., McKeown, R. E., & Waller, J. L. (1991b). The CES-D as a screen for depression and other psychiatric disorders in adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 30*(4), 636-641.

Garrison, C. Z., Jackson, K. L., Addy, C. L., McKeown, R. E., & Waller, J. L. (1991c). Suicidal behaviours in young adolescents. *American Journal of Epidemiology, 133*, 1005-1014.

Garrison, C. Z., Jackson, K. L., Marsteller, F., McKeown, R., & Addy, C. L. (1990). A longitudinal study of depressive symptomatology in young adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 29*(4), 581-585.

Garrison, C. Z., Schluchter, M. D., Schoenbach, V. J., & Kaplan, B. K. (1989). Epidemiology of depressive symptoms in young adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 28*(3), 343-351.

Garrison, C. Z., Waller, J. L., Cuffe, S. P., McKeown, R. E., Addy, C. L., & Jackson, K. L. (1997). Incidence of major depressive disorder and dysthymia in young adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 36*(4), 458-465.

Ge, X., Lorenz, F. O., Conger, R. D., Elder, G. H., & Simons, R. L. (1994). Trajectories of stressful life events and depressive symptoms during adolescence. *Developmental Psychology, 30*(4), 467-483.

Gelin, M. N., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the Center for Epidemiologic Studies Depression scale. *Educational & Psychological Measurement, 63*(1), 65-74.

Gibbons, R. D., Clark, D. C., & Kupfer, D. J. (1993). Exactly what does the Hamilton Depression Rating Scale measure? *Journal of Psychiatric Research, 27*(3), 259-273.

Gibbons, R. D., Clark, D. C., VonAmmon Cavanaugh, S., & Davis, J. M. (1985). Application of modern psychometric theory in psychiatric research. *Journal of Psychiatric Research, 19*(1), 43-55.

Gjerde, P. F. (1995). Alternative pathways to chronic depressive symptoms in young adults: Gender differences in developmental trajectories. *Child Development, 66*(5), 1277-1300.

Gjerde, P. F., & Block, J. (1991). Preadolescent antecedents of depressive symptomatology at age 18: A prospective study. *Journal of Youth and Adolescence, 20*(2), 217-232.

Gjerde, P. F., & Block, J. (1996). A developmental perspective on depressive symptoms in adolescence: Gender differences in autocentric-allocentric modes of impulse regulation. In D. Cicchetti & S. L. Toth (Eds.), *Adolescence: Opportunities and challenges. Rochester symposium on developmental psychopathology, Vol. 7* (pp. 167-196). Rochester, NY, USA: University of Rochester Press.

Gjerde, P. F., Block, J., & Block, J. H. (1988). Depressive symptoms and personality during late adolescence: Gender differences in the externalization-internalization of symptom expression. *Journal of Abnormal Psychology, 97*(4), 475-486.

Gjerde, P. F., & Westenberg, P. M. (1998). Dysphoric adolescents as young adults: A prospective study of the psychological sequelae of depressed mood in adolescence. *Journal of Research on Adolescence, 8*(3), 377-402.

Gladstone, T. R. G., Kaslow, N. J., Seeley, J. R., & Lewinsohn, P. M. (1997). Sex differences, attributional style, and depressive symptoms among adolescents. *Journal of Abnormal Child Psychology, 25*(4), 297-305.

Gladstone, T. R. G., & Koenig, L. J. (1994). Sex differences in depression across the high school to college transition. *Journal of Youth and Adolescence, 23*(6), 643-669.

Golding, J. M., & Aneshensel, C. S. (1989). Factor structure of the Center for Epidemiologic Studies Depression Scale among Mexican Americans and non-Hispanic Whites. *Psychological Assessment, 1*(3), 163-168.

Goldstein, H. (1987). *Multilevel Models in Social and Educational Research*. London: Arnold.

Goldstein, H., & Thomas, S. (1996). Using examination results as indicators of school and college Performance. *Journal of the Royal Statistical Society, 159*, 149-163.

Gore, S., & Aseltine, R. H. (1995). Protective processes in adolescence: Matching stressors with social resources. *American Journal of Community Psychology, 23*(3), 301-327.

Gore, S., Aseltine, R. H., & Colten, M. E. (1992). Social structure, life stress and depressive symptoms in a high school-aged population. *Journal of Health and Social Behavior, 33*(2), 97-113.

Gore, S., Aseltine, R. H., & Colten, M. E. (1993). Gender, social-relational involvement, and depression. *Journal of Research on Adolescence, 3*(2), 101-125.

Gotlib, I. H. (1984). Depression and general psychopathology in university students. *Journal of Abnormal Psychology, 93*, 19-30.

Gotlib, I. H., & Cane, D. B. (1989). Self-report assessment of depression and anxiety. In P. C. Kendall & D. Watson (Eds.), *Anxiety and depression: Distinctive and overlapping features.* (pp. 131-169). San Diego, CA, USA: Academic Press.

Gotlib, I. H., Lewinsohn, P. M., & Seeley, J. R. (1995). Symptoms versus a diagnosis of depression: Differences in psychosocial functioning. *Journal of Consulting and Clinical Psychology, 63*(1), 90-100.

Gotlib, I. H., Lewinsohn, P. M., Seeley, J. R., Rohde, P., & Redner, J. E. (1993). Negative cognitions and attributional style in depressed adolescents: An examination of stability and specificity. *Journal of Abnormal Psychology, 102*(4), 607-615.

Gottfredson, L. S. (1994). The science and politics of race-norming. *American Psychologist, 49*(11), 955-963.

Greenberger, E., Chen, C., Tally, S. R., & Dong, Q. (2000). Family, peer, and individual correlates of depressive symptomatology among U.S. and Chinese adolescents. *Journal of Consulting and Clinical Psychology, 68*(2), 209-219.

Guarnaccia, P. J., Angel, R., & Worobey, J. L. (1989). The factor structure of the CES-D in the Hispanic health and nutrition examination survey: The influences of ethnicity, gender and language. *Social Science and Medicine, 29*(1), 85-94.

Gustafsson, J. E. (1992). The relevance of factor analysis for the study of group differences. *Multivariate Behavioral Research, 27*(2), 239-247.

Gustafsson, J. E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28*(4), 407-434.

Haaga, D. A., & Solomon, A. (1993). Impact of Kendall, Hollon, Beck, Hammen, and Ingram (1987) on treatment of the continuity issue in "depression" research. *Cognitive Therapy and Research, 17*(4), 313-324.

Hammen, C., & Rudolph, K. D. (1996). Childhood depression. In E. J. Mash & R. A. Barkley (Eds.), *Child psychopathology* (pp. 153-195). New York, NY, USA: The Guildford Press.

Handel, R. W., Ben-Porath, Y. S., & Watt, M. (1999). Computerized adaptive assessment with the MMPI-2 in a clinical setting. *Psychological Assessment, 11*(3), 369-380.

Hankin, B. L., Abramson, L. Y., Moffitt, T. E., Silva, P. A., McGee, R., & Angell, K. E. (1998). Development of depression from preadolescence to young adulthood: Emerging gender differences in a 10-year longitudinal study. *Journal of Abnormal Psychology, 107*(1), 128-140.

Hankin, B. L., Abramson, L. Y., & Siler, M. (2001). A prospective test of the hopelessness theory of depression in adolescence. *Cognitive Therapy and Research, 25*(5), 607-632.

Hann, D., Winter, K., & Jacobsen, P. (1999). Measurement of depressive symptoms in cancer patients: Evaluation of the Center for Epidemiological Studies Depression Scale (CES-D). *Journal of Psychosomatic Research, 46*(5), 437-443.

Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): 1. Construction of the schedule. *Journal of Psychology, 10*, 249-254.

Hays, R. D. (1998). Item response theory models. In M. J. Staquet & R. D. Hays & P. M. Fayers (Eds.), *Quality of life assessment in clinical trials: methods and practice*. Oxford; New York: Oxford University Press.

Helmes, E., & Nielson, W. R. (1998). An examination of the internal structure of the Center for Epidemiological Studies-Depression Scale in two medical samples. *Personality and Individual Differences, 25*(4), 735-743.

Hendryx, M. S., Haviland, M. G., Gibbons, R. D., & Clark, D. C. (1992). An application of item response theory to alexithymia assessment among abstinent alcoholics. *Journal of Personality Assessment, 58*(3), 506-515.

Herjanic, B., & Reich, W. (1997). Development of a structured psychiatric interview for children: Agreement between child and parent on individual symptoms. *Journal of Abnormal Child Psychology, 25*(1), 21-31.

Hertzog, C., & Nesselroade, J. R. (1987). Beyond autoregressive models: Some implications of the trait-state distinction for the structural modeling of developmental change. *Child Development, 58*(1), 93-109.

Hertzog, C., Van Alstine, J., Usala, P. D., Hultsch, D. F., & Dixon, R. (1990). Measurement properties of the Center for Epidemiological Studies Depression Scale (CES-D) in older populations. *Psychological Assessment, 2*(1), 64-72.

Hill, J. P., & Lynch, M. E. (1983). The intensification of gender-related role expectations during early adolescence. In J. Brooks-Gunn & A. C. Petersen (Eds.), *Girls at puberty: Biological and psychosocial perspectives*. New York: Plenum.

Hill, P. W., & Rowe, K. J. (1996). Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement, 7*(1), 1-34.

Hoijtink, H., Rooks, G., & Wilmink, F. W. (1999). Confirmatory factor analysis of items with a dichotomous response format using the multidimensional Rasch model. *Psychological Methods, 4*(3), 300-314.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc.

Holsen, I., Kraft, P., & Vitterso, J. (2000). Stability in depressed mood in adolescence: Results from a 6-year longitudinal panel study. *Journal of Youth and Adolescence, 29*(1), 61-78.

Hops, H., Lewinsohn, P. M., Andrews, J. A., & Roberts, R. E. (1990). Psychosocial correlates of depressive symptomatology among high school students. *Journal of Clinical Child Psychology, 19*(3), 211-220.

Horn, J., & McArdle, J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3), 117-144.

Horwath, E., Johnson, J., Klerman, G. L., & Weissman, M. M. (1992). Depressive symptoms as relative and attributable risk factors for first-onset major depression. *Archives of General Psychiatry, 49*(10), 817-823.

Horwath, E., Johnson, J., Klerman, G. L., & Weissman, M. M. (1994). What are the public health implications of subclinical depressive symptoms? *Psychiatric Quarterly, 65*(4), 323-337.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 158-176). Thousand Oaks, CA, USA: Sage Publications, Inc.

Hoyle, R. H., & Smith, G. T. (1994). Formulating clinical research hypotheses as structural equation models: A conceptual overview. *Journal of Consulting and Clinical Psychology, 62*(3), 429-440.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Hu, L. T., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure-analysis be trusted. *Psychological Bulletin, 112*, 351-362.

Iwata, N., Saito, K., & Roberts, R. E. (1994). Responses to a self-administered depression scale among younger adolescents in Japan. *Psychiatry Research, 53*(3), 275-287.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 32*, 443-482.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426.

Jöreskog, K. G. 2001, October 18(a). Analysis of ordinal variables 1: Preliminary analyses. Available: http://www.ssicentral.com/lisrel/column7.htm [2002, July 17].

Jöreskog, K. G. 2001, October 18(b). Analysis of ordinal variables 2: Cross-sectional data. Available: http://www.ssicentral.com/lisrel/column8.htm [2002, July 17].

Jöreskog, K. G. 2001, December 12(c). Analysis of ordinal variables 3: Longitudinal data. Available: http://www.ssicentral.com/lisrel/column9.htm [2002, July 17].

Jöreskog, K. G. 2002, February 5. Analysis of ordinal variables 4: Multiple groups. Available: http://www.ssicentral.com/lisrel/column10.htm [2002, July 17].

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36*(3), 347-387.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: Users reference guide*. Chicago: Scientific Software International.

Jorm, A. F. (1987). Sex and age differences in depression: A quantitative synthesis of published research. *Australian and New Zealand Journal of Psychiatry, 21*(1), 46-53

Kandel, D. B., & Davies, M. (1986). Adult sequelae of adolescent depressive symptoms. *Archives of General Psychiatry, 43*(3), 255-262.

Kandel, D. B., Raveis, V. H., & Davies, M. (1991). Suicidal ideation in adolescence: Depression, substance use, and other risk factors. *Journal of Youth and Adolescence, 20*(2), 289-309.

Kaplan, D. (1991). The behaviour of three weighted least squares estimators for structured means analysis with non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 44*(2), 333-346.

Keeves, J. P. (1997). Measures of variation. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: an international handbook* (2nd ed., pp. 580-592). New York, N.Y.: Pergamon.

Kendall, P. C., Cantwell, D. P., & Kazdin, A. E. (1989). Depression in children and adolescents: Assessment issues and recommendations. *Cognitive Therapy and Research, 13*(2), 109-146.

Kendall, P. C., & Flannery-Schroeder, E. C. (1995). Rigor, but not rigor mortis, in depression research. *Journal of Personality and Social Psychology, 68*(5), 892-894.

Kendall, P. C., Hollon, S. D., Beck, A. T., Hammen, C. L., & Ingram, R. E. (1987). Issues and recommendations regarding use of the Beck Depression Inventory. *Cognitive Therapy and Research, 11*(3), 289-299.

Kessler, R. C., McGonagle, K. A., Swartz, M., Blazer, D. G., & Nelson, C. B. (1993). Sex and depression in the National Comorbidity Survey: I. Lifetime prevalence, chronicity and recurrence. *Journal of Affective Disorders, 29*(2-3), 85-96.

Kessler, R. C., & McLeod, J. D. (1984). Sex differences in vulnerability to undesirable life events. *American Sociological Review, 49*(5), 620-631.

Kessler, R. C., Price, R. H., & Wortman, C. B. (1985). Social factors in psychopathology: Stress, social support, and coping processes. *Annual Review of Psychology, 36*, 531-572.

Killen, J. D., Hayward, C., Wilson, D. M., Taylor, C. B., Hammer, L. D., Litt, I., Simmonds, B., & Haydel, F. (1994). Factors associated with eating disorder symptoms in a community sample of 6th and 7th grade girls. *International Journal of Eating Disorders, 15*(4), 357-367.

King, D. A., & Buchwald, A. M. (1982). Sex differences in subclinical depression: Administration of the Beck Depression Inventory in public and private disclosure situations. *Journal of Personality and Social Psychology, 42*(5), 963-969.

Kleinman, A., & Good, B. (1985). *Culture and depression*. Berkeley, California: University of California Press.

Knight, R. G., Williams, S., McGee, R., & Olaman, S. (1997). Psychometric properties of the Centre for Epidemiologic Studies Depression Scale (CES-D) in a sample of women in middle life. *Behaviour Research and Therapy, 35*(4), 373-380.

Kovacs, M. (1989). Affective disorders in children and adolescents. *American Psychologist, 44*(2), 209-215.

Kovacs, M. (1992). *Children's Depression Inventory Manual*. New York: Multi-Health Systems.

Kovacs, M. (1997). Depressive disorders in childhood: An impressionistic landscape. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 38*(3), 287-298.

Kovacs, M., Feinberg, T., Crouse-Novak, M., Paulauskas, S., & Finkelstein, R. (1984). Depressive disorders in childhood. I. A longitudinal prospective study of characteristics and recovery. *Archives of General Psychiatry., 41*(3), 229-237.

Kovacs, M., Feinberg, T., Crouse-Novak, M., Paulauskas, S., Pollack, M., & Finkelstein, R. (1984). Depressive disorders in childhood. II. A longitudinal study of the risk for a subsequent major depression. *Archives of General Psychiatry., 41*(3), 635-644.

Kovacs, M., Goldston, D., & Gatsonis, C. (1993). Suicidal behaviors and childhood-onset depressive disorders: A longitudinal investigation. *Journal of the American Academy of Child and Adolescent Psychiatry, 32*(1), 8-20.

Kuo, W. H. (1984). Prevalence of depression among Asian-Americans. *Journal of Nervous and Mental Disease, 172*(8), 449-457.

Krueger, R. F., & Finger, M. S. (2001). Using item response theory to understand comorbidity among anxiety and unipolar mood disorders. *Psychological Assessment, 13*(1), 140-151.

Kutcher, S. P., & Marton, P. (1989). Parameters of adolescent depression: A review. *Psychiatric Clinics of North America, 12*(4), 895-918.

Labouvie, E., & Ruetsch, C. (1995). Wholes or parts? *Multivariate Behavioral Research, 30*(1), 121-123.

Langhinrichsen-Rohling, J., Lewinsohn, P., Rohde, P., Seeley, J., Monson, C. M., Meyer, K. A., & Langford, R. (1998). Gender differences in the suicide-related behaviors of adolescents and young adults. *Sex Roles, 39*(11-12), 839-854.

Larson, R. W., Raffaelli, M., Richards, M. H., Ham, M., & Jewell, L. (1990). Ecology of depression in late childhood and early adolescence: A profile of daily states and activities. *Journal of Abnormal Psychology, 99*(1), 92-102.

Lasko, D. S., Field, T. M., Gonzalez, K. P., Harding, J., Yando, R., & Bendell, D. (1996). Adolescent depressed mood and parental unhappiness. *Adolescence, 31*(121), 49-57.

Leadbeater, B. J., Blatt, S. J., & Quinlan, D. M. (1995). Gender-linked vulnerabilities to depressive symptoms, stress, and problem behaviors in adolescents. *Journal of Research on Adolescence, 5*(1), 1-29.

Lewinsohn, P. M., Clarke, G. N., Hops, H., & Andrews, J. A. (1990). Cognitive-behavioral treatment for depressed adolescents. *Behavior Therapy, 21*(4), 385-401.

Lewinsohn, P. M., Hops, H., Roberts, R. E., Seeley, J. R., & Andrews, J. A. (1993). Adolescent psychopathology: I. Prevalence and incidence of depression and other DSM-III--R disorders in high school students. *Journal of Abnormal Psychology, 102*(1), 133-144.

Lewinsohn, P. M., Joiner, T. E., Jr., & Rohde, P. (2001). Evaluation of cognitive diathesis-stress models in predicting major depressive disorder in adolescents. *Journal of Abnormal Psychology, 110*(2), 203-215.

Lewinsohn, P. M., Rohde, P., & Farrington, D. P. (2000). The OADP-CDS: A brief screener for adolescent conduct disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 39*(7), 888-895.

Lewinsohn, P. M., Rohde, P., & Seeley, J. R. (1994). Psychosocial risk factors for future adolescent suicide attempts. *Journal of Consulting and Clinical Psychology, 62*(2), 297-305.

Lewinsohn, P. M., Rohde, P., Seeley, J. R., & Fischer, S. A. (1993). Age-cohort changes in the lifetime occurrence of depression and other mental disorders. *Journal of Abnormal Psychology, 102*(1), 110-120.

Lewinsohn, P. M., Seeley, J. R., Roberts, R. E., & Allen, N. B. (1997). Center for Epidemiologic Studies Depression Scale (CES-D) as a screening instrument for depression among community-residing older adults. *Psychology and Aging, 12*(2), 277-287.

Li, F., Harmer, P., & Acock, A. (1996). The Task and Ego Orientation in Sport Questionnaire: Construct equivalence and mean differences across gender. *Research Quarterly for Exercise and Sport, 68*(2), 228-238.

Liang, J., Tran, T. V., Krause, N., & Markides, K. S. (1989). Generational differences in the structure of the CES-D scale in Mexican Americans. *Journals of Gerontology, 44*(3), S110-S120.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*(1), 53-76.

Lloyd, C. (1980). Life events and depressive disorder reviewed: II. Events as precipitating factors. *Archives of General Psychiatry, 37*(5), 541-548.

Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis (3rd Ed.)*. Mahwah, NJ, US: Lawrence Erlbaum Associates, Inc., Publishers.

Long, J. D., & Brekke, J. S. (1999). Longitudinal factor structure of the Brief Psychiatric Rating Scale in schizophrenia. *Psychological Assessment, 11*(4), 498-506.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lubin, B., & Himelstein, P. (1976). Reliability of the Depression Adjective Check Lists. *Perceptual and Motor Skills, 43*(3), 1037-1038.

Lubke, G. H., Dolan, C. V., & Kelderman, H. (2001). Investigating group differences on cognitive tests using Spearman's hypothesis: An evaluation of Jensen's method. *Multivariate Behavioral Research, 36*(3), 299-324.

Mackinnon, A., Jorm, A. F., Christensen, H., Scott, L. R., Henderson, A. S., & Korten, A. E. (1995). A latent trait analysis of the Eysenck Personality Questionnaire in an elderly community sample. *Personality and Individual Differences, 18*(6), 739-747.

Manson, S. M., Ackerson, L. M., Dick, R. W., Baron, A. E., & Fleming, C. M. (1990). Depressive symptoms among American Indian adolescents: Psychometric characteristics of the Center for Epidemiologic Studies Depression Scale (CES-D). *Psychological Assessment, 2*(3), 231-237.

Marciano, P. L., & Kazdin, A. E. (1994). Self-esteem, depression, hopelessness, and suicidal intent among psychiatrically disturbed inpatient children. *Journal of Clinical Child Psychology, 23*(2), 151-160.

Marcotte, D. (1996). Irrational beliefs and depression in adolescence. *Adolescence, 31*(124), 935-954.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*, 519-530.

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

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*(3), 313-349.

Marshall, S. C., Mungas, D., Weldon, M., Reed, B., & Haan, M. (1997). Differential item functioning in the Mini-Mental State Examination in English and Spanish speaking older adults. *Psychology and Aging, 12*(4), 718-725.

Martin, G., Roeger, L., Dadds, V., & Allison, S. (1997). *Early detection of emotional disorders in South Australia: The first two years.* Adelaide, South Australia: Southern Child and Adolescent Mental Health Service.

Masten, W. G., Caldwell-Colbert, A. T., Alcala, S. J., & Mijares, B. E. (1986). Reliability and validity of the Center for Epidemiological Studies Depression Scale. *Hispanic Journal of Behavioral Sciences, 8*(1), 77-84.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Mattison, R. E., Handford, H. A., Kales, H. C., Goodman, A. L., & McLaughlin, L. (1990). Four-year predictive value of the Children's Depression Inventory. *Psychological Assessment, 2*(2), 169-174.

Maydeu-Olivares, A. (2000). Review of M*PLUS*. *Multivariate Behavioral Research, 35*(4), 501-505.

McCallum, J., Mackinnon, A., Simons, L., & Simons, J. (1995). Measurement properties of the Center for Epidemiological Studies Depression Scale: An Australian community study of aged persons. *Journals of Gerontology: Series B: Psychological Sciences and Social Sciences, 50b*(3), S182-S189.

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64-82.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107*(2), 247-255.

McGaw, B., & Glass, G. V. (1980). Choice of the metric for effect size in meta-analysis. *American Educational Research Journal, 17*(3), 325-337.

McKeown, R. E., Garrison, C. Z., Cuffe, S. P., Waller, J. L., Jackson, K. L., & Addy, C. L. (1998). Incidence and predictors of suicidal behaviors in a longitudinal sample of young adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 37*(6), 612-619 l.

McKeown, R. E., Garrison, C. Z., Jackson, K. L., Cuffe, S. P., Addy, C. L., & Waller, J. L. (1997). Family structure and cohesion, and depressive symptoms in adolescents. *Journal of Research on Adolescence, 7*(3), 267-281.

Mechanic, D., & Hansell, S. (1987). Adolescent competence, psychological well-being, and self-assessed physical health. *Journal of Health and Social Behavior, 28*(4), 364-374.

Mecklin, C., and Mundfrom, DJ. (2002). An appraisal and bibliography of tests of multivariate normality. *International Statistical Review (forthcoming)*.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543.

Meredith, W. (1995). Two wrongs may not make a right. *Multivariate Behavioral Research, 30*(1), 89-94.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57*(2), 289-311.

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research, 30*(4), 577-605.

Millsap, R. E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research, 33*(3), 403-424.

Millsap, R. E. 2001, November 27. Ordinal variables and multigroup modeling. [Newsgroup]. Structural equation modeling discussion group. Available: http://bama.ua.edu/archives/semnet.html [2002, August 23].

Ministerial Council on Education, Employment, Training and Youth Affairs. (1998). *National Report on Schooling in Australia 1996*. Melbourne: Curriculum Corporation.

Moran, P. W., & Lambert, M. J. (1983). A review of current assessment tools for monitoring changes in depression. In M. J. Lambert & E. R. Christensen & S. S. DeJulio (Eds.), *The assessment of psychotherapy outcome* (pp. 304-355). New York: Wiley.

Morris, C. (1995). Hierarchical models for educational data: An overview. *Journal of Educational and Behavioural Statistics, 20*(2), 190-200.

Murray, C. J. L., & Lopez, A. D. (1996). *The global burden of disease: A comprehensive assessment of mortality and disability, injuries, and risk factors in 1990 and projected to 2020*. Cambridge: MA.

Muthén, B. O. (1978). Contributions to factor analysis of dichotomized variables. *Psychometrika, 43*, 551-560.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), 115-132.

Muthén, B. O. (1989a). Dichotomous factor analysis of symptom data. *Sociological Methods and Research*, *1*(18), 19-65.

Muthén, B. O. (1989b). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557-585.

Muthén, B. O. 2000, October 11. Measurement invariance. [Mplus Discussion Group]. Available: http://www.statmodel.com/index2.html [2001, January 7].

Muthén, B. O. 2001a, February 26. Multilevel data/Complex sample: Multiplegroup model. [Mplus Discussion Group]. Available: http://www.statmodel.com/index2.html [2002, April 15].

Muthén, B. O. 2001b, February 13. Mplus: best strategy with non-normal data. [Newsgroup]. Structural equation modeling discussion group. Available: http://bama.ua.edu/archives/semnet.html [2002, August 23].

Muthén, B. O. 2001c, December 4. Amos or Mplus. [Newsgroup]. Structural equation modeling discussion group. Available: http://bama.ua.edu/archives/semnet.html [2002, August 23].

Muthén, B. O., & Asparouhov, T. 2002, November 8. Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. Available: http://www.statmodel.com/mplus/examples/webnote.html [2000, November 19].

Muthén, B. O., & Kaplan, D. (1985). A comparison for some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*, 171-189.

Muthén, B. O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT based detection technique to mathematics achievement test items. *Journal of Education Measurement, 28*(1), 1-22.

Muthén, L. K., & Muthén, B. O. (1998). M*plus: The comprehensive modeling program for applied researchers*. Los Angeles, CA: Muthén & Muthén.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30*(4), 293-311.

National Health and Medical Research Council. (1996). *Effective school health promotion: Towards health promoting schools*. Canberra: NHMRC.

Nesselroade, J. R. (1995). " . . .and expectation fainted, longing for what it had not." Comments on Labouvie and Ruetsch's "Testing for equivalence . . ." *Multivariate Behavioral Research, 30*(1), 95-99.

Newmann, J. (1989). Aging and depression. *Psychology and Aging, 4*(2), 150-165.

Newmann, J. P. (1984). Sex differences in symptoms of depression: Clinical disorder or normal distress? *Journal of Health and Social Behavior, 25*(2), 136-159.

Nishide, T., & Natsuno, Y. (1997). The effects of family system functioning on the pupils' depressive mood. *Japanese Journal of Educational Psychology, 45*(4), 90-97.

Noh, S., Avison, W. R., & Kaspar, V. (1992). Depressive symptoms among Korean immigrants: Assessment of a translation of the Center for Epidemiologic Studies-Depression Scale. *Psychological Assessment, 4*(1), 84-91.

Nolen-Hoeksema, S. (1990). *Sex differences in depression*. Stanford, CA: Stanford University Press.

Nolen-Hoeksema, S. (1991). Responses to depression and their effects on the duration of depressive episodes. *Journal of Abnormal Psychology, 100*(4), 569-582.

Nolen-Hoeksema, S., Girgus, J. S., & Seligman, M. E. (1992). Predictors and consequences of childhood depressive symptoms: A 5-year longitudinal study. *Journal of Abnormal Psychology, 101*(3), 405-422.

Norton, E. C., Bieler, G. S., Ennett, S. T., & Zarkin, G. A. (1996). Analysis of prevention program effectiveness with clustered data using generalized estimating equations. *Journal of Consulting and Clinical Psychology, 64*(5), 919-926.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill Publishing.

Olsson, U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research, 14*(4), 485-500.

Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD Checklist: Detection and evaluation of impact. *Psychological Assessment, 14*(1), 50-59.

Orme, J. G., Reis, J., & Herz, E. J. (1986). Factorial and discriminant validity of the Center for Epidemiological Studies Depression (CES-D) scale. *Journal of Clinical Psychology, 42*(1), 28-33.

Osterlind, S. J. (1983). *Test item bias.* Beverly Hills: Sage Publications.

Paikoff, R. L., Brooks Gunn, J., & Warren, M. P. (1991). Effects of girls' hormonal status on depressive and aggressive symptoms over the course of one year. *Journal of Youth and Adolescence, 20*(2), 191-215.

Papassotiropoulos, A., Heun, R., & Maier, W. (1999). The impact of dementia on the detection of depression in elderly subjects from the general population. *Psychological Medicine, 29*(1), 113-120.

Parker, G., Tupling, H., & Brown, L. B. (1979). A parental bonding instrument. *British Journal of Medical Psychology, 52*(1), 1-10.

Paterniti, S., Verdier-Taillefer, M. H., Geneste, C., Bisserbe, J. C., & Alpeovitch, A. (2000). Low blood pressure and risk of depression in the elderly: A prospective community-based study. *British Journal of Psychiatry, 176*, 464-467.

Peaker, G. F. (1975). *An empirical study of education in twenty-one countries: A technical report*. Stockholm: International Association for the Evaluation of Educational Achievement (IEA).

Pearce, C. M., & Martin, G. (1994). Predicting suicide attempts among adolescents. *Acta Psychiatrica Scandinavica*, *90*(5), 324-328.

Peden, A. R., Hall, L. A., Rayens, M. K., Beebe, L. L., Hall, L. A., Rayens, M. K., & Beebe, L. L. (2000). Reducing negative thinking and depressive symptoms in college women. *Journal of Nursing Scholarship, 32*(2), 145-151.

Pentz, M. A., & Chou, C. P. (1994). Measurement invariance in longitudinal clinical research assuming change from development and intervention. *Journal of Consulting and Clinical Psychology, 62*(3), 450-462.

Petersen, A. C., Compas, B. E., & Brooks-Gunn, J. (1992). *Depression in adolescence: Current knowledge, research directions, and implications for programs and policy.* Washington DC: Carnegie Council on Adolescent Development.

Petersen, A. C., Compas, B. E., Brooks Gunn, J., Stemmler, M., Ey, S., & Grant, K. E. (1993). Depression in adolescence. *American Psychologist, 48*(2), 155-168.

Petersen, A. C., Kennedy, R. E., & Sarigiani, P. A. (1991). Coping with adolescence. In M. E. Colton & S. Gore (Eds.), *Adolescent stress: Causes and consequences* (pp. 93-110). New York: Aldine de Gruyter.

Petersen, A. C., Sarigiani, P. A., & Kennedy, R. E. (1991). Adolescent depression: Why more girls? *Journal of Youth and Adolescence, 20*(2), 247-271.

Post, R. M. (1992). Transduction of psychosocial stress into the neurobiology of recurrent affective disorder. *American Journal of Psychiatry, 149*(8), 999-1010.

Prescott, C. A., McArdle, J. J., Hishinuma, E. S., Johnson, R. C., Miyamoto, R. H., Andrade, N. N., Edman, J. L., Makini, G. K., Jr., Nahulu, L. B., Yuen, N. Y. C., & Carlton, B. S. (1998). Prediction of major depression and dysthymia from CES-D scores among ethnic minority adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 37*(5), 495-503.

Pumariega, A. J., Johnson, N. P., Sheridan, D., & Cuffe, S. P. (1996). The influence of race and gender on depressive and substance abuse symptoms in high-risk adolescents. *Cultural Diversity and Mental Health, 2*(2), 115-123.

Raczek, A. E., Ware, J. E., Bjorner, J. B., Gandek, B., Haley, S. M., Aaronson, N. K., Apolone, G., Bech, P., Brazier, J. E., Bullinger, M., & Sullivan, M. (1998). Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: Results from the IQOLA project. *Journal of Clinical Epidemiology, 51*(11), 1203-1214.

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385-401.

Radloff, L. S. (1991). The use of the Center for Epidemiologic Studies Depression Scale in adolescents and young adults. *Journal of Youth and Adolescence, 20*(2), 149-166.

Radloff, L. S., & Teri, L. (1986). Use of the Center for Epidemiological Studies-Depression Scale with older adults. *Clinical Gerontologist, 5*(1-2), 119-136.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517-529.

Ramsay, J. O. (1991). Kernel-smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611-630.

Ramsay, J. O. (2000). TESTGRAF: A program for the graphical analysis of multiple choice test and questionnaire data. Montreal, Quebec, Canada: McGill University.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA Press.

Rasmussen, J. L., & Dunlap, W. P. (1991). Dealing with nonnormal data: Parametric analysis of transformed data vs nonparametric analysis. *Educational and Psychological Measurement, 51*(4), 809-820.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25-36.

Reifman, A., & Windle, M. (1995). Adolescent suicidal behaviors as a function of depression, hopelessness, alcohol use, and social support: A longitudinal investigation. *American Journal of Community Psychology, 23*(3), 329-354.

Reinherz, H. Z., Giaconia, R. M., Pakiz, B., Silverman, A. M., Frost, A. K., & Lefkowitz, E. S. (1993). Psychosocial risks for major depression in late adolescence: A longitudinal community study. *Journal of the American Academy of Child and Adolescent Psychiatry, 32*(6), 1155-1164.

Reise, S. P., Smith, L., & Furr, R. M. (2001). Invariance on the NEO PI-R neuroticism scale. *Multivariate Behavioral Research, 36*(1), 83-110.

Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*(3), 287-297.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552-566.

Reynolds, C. R., & Harding, R. E. (1983). Outcome in two large sample studies of factorial similarity under six methods of comparison. *Educational and Psychological Measurement, 43*(3), 723-728.

Reynolds, W. M. (1992). Depression in children and adolescents. In W. M. Reynolds (Ed.), *Internalizing disorders in children and adolescents* (pp. 149-253). New York: Wiley-Interscience.

Rigdon, E. 1996, October 28. Normality and transforms. [Newsgroup]. Available: http://bama.ua.edu/archives/semnet.html [2002, August 23].

Rigdon, E. 2000, February 16. Multigroup with polychorics. [Newsgroup]. Available: http://bama.ua.edu/archives/semnet.html [2002, August 23].

Rigdon, E. 2001, June 1. Polychoric correlations. [Newsgroup]. Available: http://bama.ua.edu/archives/semnet.html [2002, August 23].

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23*(1), 51-67.

Roberts, R. E., Andrews, J. A., Lewinsohn, P. M., & Hops, H. (1990a). Assessment of depression in adolescents using the Center for Epidemiologic Studies Depression Scale. *Psychological Assessment, 2*(2), 122-128.

Roberts, R. E., & Chen, Y. W. (1995). Depressive symptoms and suicidal ideation among Mexican-origin and Anglo adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 34*(1), 81-90.

Roberts, R. E., Lewinsohn, P. M., & Seeley, J. R. (1991). Screening for adolescent depression: A comparison of depression scales. *Journal of the American Academy of Child and Adolescent Psychiatry, 30*(1), 58-66.

Roberts, R. E., Rhoades, H. M., & Vernon, S. W. (1990b). Using the CES-D scale to screen for depression and anxiety: Effects of language and ethnic status. *Psychiatry Research, 31*(1), 69-83.

Roberts, R. E., & Sobhan, M. (1992). Symptoms of depression in adolescence: A comparison of Anglo, African, and Hispanic Americans. *Journal of Youth and Adolescence, 21*(6), 639-651.

Roberts, R. E., & Vernon, S. W. (1983). The Center for Epidemiological Studies Depression Scale: Its use in a community sample. *American Journal of Psychiatry, 140*(1), 41-46.

Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance, 14*(2), 187-207.

Robins, L. N., & Regier, D. A. (1991). *Psychiatric disorders in America: The Epidemiological Catchment Area study*. New York: Free Press.

Roeger, L., Allison, S., Martin, G., Dadds, V., & Keeves, J. (2001). Adolescent depressive symptomatology: Improve schools or help students. *Australian Journal of Psychology, 53*(3), 134-139.

Rohde, P., Lewinsohn, P. M., & Seeley, J. R. (1994). Response of depressed adolescents to cognitive-behavioral treatment: Do differences in initial severity clarify the comparison of treatments? *Journal of Consulting and Clinical Psychology, 62*(4), 851-854.

Ross, K. N. (1988). Sampling. In J. P. Keeves (Ed*.), Educational research, methodology, and measurement: An international handbook* (pp. 527-537). Oxford, England UK: Pergamon Press.

Rothman, S. (2001). School absence and student background factors: A multilevel analysis. *International Education Journal*, *2*(1), (WWW) http://ed.sturt.flinders.edu.au/iej/welcome.html (2002, February 2).

Rowe, K., Hill, P., & Holmes-Smith, P. (1995). Methodological issues in educational performance and school effectiveness research: A discussion with worked examples. *Australian Journal of Education, 39*(3), 217-248.

Rumberger, R. W. (1995). Dropping out of middle school: A multilevel analysis of students and schools. *American Educational Research Journal, 32*(3), 583-625.

Rushton, J. P., & Chrisjohn, R. D. (1981). Extraversion, neuroticism, psychoticism and self-reported delinquency: Evidence from eight separate samples. *Personality and Individual Differences*, *2*(1), 11-20.

Rutter, M. (1983). School effects on pupil progress: Research findings and policy implications. *Child Development, 54*(1), 1-29.

Rutter, M., Giller, H., & Hagell, A. (1998). *Antisocial behavior by young people*. New York, NY, US: Cambridge University Press.

Ryan, N. D., Puig Antich, J., Ambrosini, P., Rabinovich, H., Robinson, D., Nelson, B., Iyengar, S., & Twomey, J. (1987). The clinical picture of major depression in children and adolescents. *Archives of General Psychiatry, 44*(10), 854-861.

Ryan, N. D., Williamson, D. E., Iyengar, S., Orvaschel, H., Reich, T., Dahl, R. E., & Puig-Atich, J. A. (1992). A secular increase in child and adolescent onset affective disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 31*(4), 600-605.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2), 100.

Santor, D. A., & Coyne, J. C. (1997). Shortening the CES-D to improve its ability to detect cases of depression. *Psychological Assessment, 9*(3), 233-243.

Santor, D. A., & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment, 10*(4), 345-359.

Santor, D. A., Ramsay, J. O., & Zuroff, D. C. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. *Psychological Assessment, 6*(3), 255-270.

Santor, D. A., Zuroff, D. C., Ramsay, J. O., Cervantes, P., & Palacios, J. (1995). Examining scale discriminability in the BDI and CES-D as a function of depressive severity. *Psychological Assessment, 7*(2), 131-139.

Satorra, A., & Bentler, P. M. (1994). Corrections to standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications to developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507-514.

Sawyer, M. G., Arney, F., M., Baghurst, P. A., Clark, J. J., Graetz, B. W., Kosky, R. J., Nurcombe, B., Patton, G. C., Prior, M. R., Raphael, B., Rey, J., Whaites, L. C., & Zubrick, S. R. (2000). *The mental health of young people in Australia: The child and adolescent component of the national survey of mental health and well being.* Canberra: AusInfo.

Sawyer, M. G., Arney, F., M., Baghurst, P. A., Clark, J. J., Graetz, B. W., Kosky, R. J., Nurcombe, B., Patton, G. C., Prior, M. R., Raphael, B., Rey, J., Whaites, L. C., & Zubrick, S. R. (2001). The mental health of young people in Australia: key findings from the child and adolescent component of the national survey of mental health and well being. *Australian and New Zealand Journal of Psychiatry, 35*, 806-814.

Sawyer, M., Clark, J., & Baghurst, P. (1993). Childhood emotional and behavioural problems: A comparison of children's reports with reports from parents and teachers. *Journal of Paediatric Child Health, 29*, 119-125.

Sawyer, M. G., Sarris, A., Baghurst, P. A., Cornish, C. A., & Kalucy, R. S. (1990). The prevalence of emotional and behaviour disorders and patterns of service utilisation in children and adolescents. *Australian and New Zealand Journal of Psychiatry, 24*(3), 323-330.

Schaie, K. W., & Hertzog, C. (1985). Measurement in the psychology of adulthood and aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (2nd ed., pp. 61-92). New York: Van Nostrand Reinhold.

Schaie, K., Maitland, S., Willis, S., & Intrieri, R. (1998). Longitudinal invariance of adult psychometric ability factor structures across 7 years. *Psychology and Aging, 13*(1), 8-20.

Scheuneman, J. D., & Bleistein, C. A. (1997). Item bias. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: an international handbook* (2nd ed., pp. 3043-3051). New York, N.Y.: Pergamon.

Scheuneman, J. D., & Bleistein, C. A. (1999). Item bias. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 220-234). Oxford: Pergamon.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53-61.

Schoenbach, V. J. (1982). Use of a symptom scale to study the prevalence of a depressive syndrome in young adolescents. *American Journal of Epidemiology, 115*, 791-800.

Segal, Z. V., Williams, J. M., Teasdale, J. D., & Gemar, M. (1996). A cognitive science perspective on kindling and episode sensitization in recurrent affective disorder. *Psychological Medicine, 26*(2), 371-380.

Seiffge-Krenke, I., & Stemmler, M. (2002). Factors contributing to gender differences in depressive symptoms: A test of three developmental models. *Journal of Youth and Adolescence, 31*(6), 405-417.

Shaffer, D., Fisher, P., Lucas, C., Dulcan, M., & Schwab-Stone, M. (2000). NIMH Diagnostic Interview Schedule for Children Version IV (NIMH DISC-IV): Description, differences from previous versions, and reliability of some common diagnoses. *Journal of the American Academy of Child and Adolescent Psychiatry, 39*(1), 28-38.

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential test functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-239). Hillsdale, NJ, USA: Lawrence Erlbaum Associates.

Sheeber, L., Hops, H., Alpert, A., Davis, B., & Andrews, J. (1997). Family support and conflict: Prospective relations to adolescent depression. *Journal of Abnormal Child Psychology, 25*(4), 333-344.

Sheehan, T. J., Fifield, J., Reisine, S., & Tennen, H. (1995). The measurement structure of the Center for Epidemiologic Studies Depression Scale. *Journal of Personality Assessment, 64*(3), 507-521.

Sheehan, M., Marshall, B., Cahill, H., Rowling, L., & Holdsworth, R. (1999). *SchoolMatters: Managing and mapping mental health in schools*. Canberra: Commonwealth Department of Health and Family Services.

Silburn, S. R., Zubrick, S. R., Garton, A., Gurrin, L., Burton, P., Dalby, R., Calton, J., Shepherd, C., & Lawrence, D. (1996). *Western Australian Child Health Survey: Family and Community Health*. Perth, Western Australia: Australian Bureau of Statistics and the TVW Television Institute for Child Health Research.

Silverstein, B., Caceres, J., Perdue, L., & Cimarolli, V. (1995). Gender differences in depressive symptomatology: The role played by "anxious somatic depression" associated with gender-related achievement concerns. *Sex Roles, 33*(9-10), 621-636.

Silverstein, M. L., & Nelson, L. D. (2000). Clinical and research implications of revising psychological tests. *Psychological Assessment, 12*(3), 298-303.

Simmons, R. G., Burgeson, R., Carlton-Ford, S., & Blyth, D. A. (1987). The impact of cumulative change in early adolescence. *Child Development, 58*(5), 1220-1234.

Simpson, E. H. (1951). Interpretation of interaction contingency tables. *Journal of the Royal Statistical Society, (Series B), 13*, 238-241.

Snedecor, G. W., & Cochran, W. G. (1967). *Statistical methods* (6th ed.). Ames, Iowa: Iowa State University Press.

Snijders, T. A. B. 1999, August 16. ICC in HGLM. [Newsgroup]. Available: http://www.jiscmail.ac.uk/lists/multilevel.html [2001, July 1].

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modelling.* London: Sage.

Soler, J., Perez Sola, V., Puigdemont, D., Perez Blanco, J., Figueres, M., & Alvarez, E. (1997). Validation study of the Center for Epidemiologic Studies-Depression (CES-D) in affective disorders in Spanish population. *Actas Luso Espanolas de Neurologia y Psiquiatria y Ciencias Afines, 25*(4), 243-249.

Steenkamp, J., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78-113.

Stommel, M., Given, B. A., Given, C. W., Kalaian, H. A., Schulz, R., & McCorkle, R. (1993). Gender bias in the measurement properties of the Center for Epidemiologic Studies Depression Scale (CES-D). *Psychiatry Research, 49*(3), 239-250.

Strauss, C. C., Smith, K., Frame, C., & Forehand, R. (1985). Personal and interpersonal characteristics associated with childhood obesity. *Journal of Paediatric Psychology, 10*(3), 337-343.

Suh, T., & Gallo, J. (1997). Symptom profiles of depression among general medical service users compared with specialty mental health service users. *Psychological Medicine, 27*(5), 1051-1063.

Susman, E. J., Inoff-Germain, G., Nottelmann, E. D., Loriaux, D. L., Cutler, G. B., & Chrousos, G. P. (1987). Hormones, emotional dispositions, and aggressive attributes in young adolescents. *Child Development*, *58*(4), 1114-1134.

Swanson, J. W., Linskey, A. O., Quintero-Salinas, R., Pumariega, A. J., & Holzer, C. E. (1992). A binational school survey of depressive symptoms, drug use, and suicidal ideation. *Journal of the American Academy of Child and Adolescent Psychiatry, 31*(4), 669-678.

Taylor, S. E., Repetti, R. L., & Seeman, T. (1997). Health psychology: What is an unhealthy environment and how does it get under the skin? *Annual Review of Psychology, 48*, 411-447.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 397-413). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Thorson, J. A., & Powell, F. C. (1993). The CES-D: Four or five factors? *Bulletin of the Psychonomic Society, 31*(6), 577-578.

Thrupp, M. (2001). Recent school effectiveness counter-critiques: problems and possibilities. *British Educational Research Journal, 27*(4), 443-457.

Tolor, A., & Murphy, V. M. (1985). Stress and depression in high school students. *Psychological Reports, 57*(2), 535-541.

Turner, R. J., & Marino, F. (1994). Social support and social structure: A descriptive epidemiology. *Journal of Health and Social Behavior, 35*(3), 193-212.

Twenge, J. M., & Nolen-Hoeksema, S. (2002). Age, gender, race, socioeconomic status, and birth cohort difference on the children's depression inventory: A meta-analysis. *Journal of Abnormal Psychology, 111*(4), 578-588.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-69.

van der Linden, W. J., & Hambleton, R. K. (1996). *Handbook of modern item response theory*. New York: Springer-Verlag.

Vernon, S. W., & Roberts, R. E. (1981). Measuring nonspecific psychological distress and other dimensions of psychopathology: Further observations on the problem. *Archives of General Psychiatry, 38*(11), 1239-1247.

Wade, T. J., Cairney, J., & Pevalin, D. (2002). Emergence of gender differences in depression during adolescence: National panel results from three countries. *Journal of the American Academy of Child and Adolescent Psychiatry, 41*(2), 190-198.

Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: An illustration with the MMPI. *Psychological Methods, 5*(1), 125-146.

Weich, S., Sloggett, A., & Lewis, G. (2001). Social roles and the gender difference in rates of the common mental disorders in Britain: A 7-year, population-based cohort study. *Psychological Medicine, 31*(6), 1055-1064.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*(6), 774-789.

Weissman, M. M., & Klerman, G. L. (1985). Gender and depression. *Trends in Neurosciences, 8*(9), 416-420.

Weissman, M. M., Orvaschel, H., & Padian, N. (1980). Children's symptoms and social functioning: Self-report scales. *Journal of Nervous and Mental Disease, 168*, 736-740.

Weissman, M. M., Sholomskas, D., Pottenger, M., Prusoff, B. A., Locke, B. Z. (1977). Assessing depressive symptoms in five psychiatric populations: A validation study. *American Journal of Epidemiology, 106*, 203-214.

Wells, J. E., Bushnell, J. A., Hornblow, A. R., Joyce, P. R., & Oakley-Browne, M. A. (1989). Christchurch Psychiatric Epidemiology Study: I. Methodology and lifetime prevalence for specific psychiatric disorders. *Australian and New Zealand Journal of Psychiatry, 23*(3), 315-326.

Wells, K. B., Burnam, M. A., Rogers, W., Hays, R., & Camp, P. (1992). The course of depression in adult outpatients: Results from the Medical Outcomes Study. *Archives of General Psychiatry, 49*(10), 788-794.

Wells, V. E., Klerman, G. L., & Deykin, E. Y. (1987). The prevalence of depressive symptoms in college students. *Social Psychiatry, 22*, 20-28.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56-75). Thousand Oaks, CA: Sage.

WichstrØm, L. (1999). The emergence of gender difference in depressed mood during adolescence: The role of intensified gender socialization. *Developmental Psychology, 35*(1), 232-245.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant & M. Windle (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC, US: American Psychological Association.

Wilhelm, K., & Parker, G. (1990). Reliability of the Parental Bonding Instrument and Intimate Bond Measure Scales. *Australian and New Zealand Journal of Psychiatry, 24*(1990), 199-202.

Windle, R. C., & Windle, M. (1997). An investigation of adolescents' substance use behaviors, depressed affect, and suicidal behaviors. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 38*(8), 921-929.

Wolff, S. (1993). *The school's potential for promoting mental health*. Geneva, Switzerland: Division of Mental Health, World Health Organization.

Wolk, S. I., & Weissman, M. M. (1995). Women and depression: An update. *American Psychiatric Press Review of Psychiatry, 14*, 227-259.

Wong, Y. L. I. (2000). Measurement properties of the Center for Epidemiologic Studies-Depression Scale in a homeless population. *Psychological Assessment, 12*(1), 69-76.

World Health Organization. (1992). *International classification of diseases and health-related problems (10th revision)*. Geneva: Author.

Ying, Y. W. (1988). Depressive symptomatology among Chinese-Americans as measured by the CES-D. *Journal of Clinical Psychology, 44*(5), 739-746.

Young, M. A., Fogg, L. F., Scheftner, W., Fawcett, J., Akiskal, H., & Maser, J. (1996). Stable trait components of hopelessness: Baseline and sensitivity to depression. *Journal of Abnormal Psychology, 105*(2), 155-165.

Zich, J. M., Attkisson, C. C., & Greenfield, T. K. (1990). Screening for depression in primary care clinics: The CES-D and the BDI. *International Journal of Psychiatry in Medicine, 20*(3), 259-277.

Zickar, M. J. (1998). Modeling item-level data with item response theory. *Current Directions in Psychological Science, 7*(4), 104-109.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20*(1), 71-87.

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology, 84*(4), 551-563.

Zimmerman, M., & Coryell, W. (1987). The Inventory to Diagnose Depression (IDD): A self-report scale to diagnose major depressive disorder. *Journal of Consulting and Clinical Psychology, 55*(1), 55-59.

Zimmerman, M., & Coryell, W. (1994). Screening for major depressive disorder in the community: A comparison of measures. *Psychological Assessment, 6*(1), 71-74.

Zonderman, A. B. (1995). Symptoms of depression are not a risk for cancer morbidity and mortality. In M. Stein & A. Baum (Eds.), *Chronic diseases. Perspectives in behavioral medicine* (pp. 169-179). Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc.

Zung, W. W. (1965). A self-rating depression scale. *Archives of General Psychiatry, 12*(1), 63-70.

A

Review of CES-D studies in adolescent samples

Author	Year	Sample	Comments
Schoenbach	1982	U.S. (North Carolina) junior high school students. N = 384, Age: 12-15.	Gender means not shown. Used a syndrome orientated approach to case finding. Found higher prevalence of black male cases.
Tolor & Murphy	1985	U.S. (Connecticut) high school students. N = 285, Age: 13-17.	Boys = 14.66, SD = 9.11; Girls = 18.48, SD = 10.81. Six month test-retest: Boys r = 0.50; Girls = 0.62.
Mechanic & Hansell	1987	U.S. (New Jersey) high school students. N = 1057, Age: 13 at Wave 1 of study.	Gender means not shown but females reported to be higher than males. Data from first two waves of longitudinal study presented.
Wells, Klerman & Deykin	1987	U.S. (Boston) college students. N = 424, Age: 16-19.	Gender means not shown. No difference between proportion of boys and girls with scores > 16. Item correlations to total score range 0.33 to 0.69. Rank order of items similar to adult sample and similar between low and high scorers.
Doerfler, Felner, Rowlison, Raley & Evans	1988	U.S. (rural Southern) high school students. N = 1207, Grade: 4-12.	At Grade 8: Boys = 15.43, SD = 9.55; Girls = 18.41, SD = 11.94. Correlation with CDI r = 0.58 and with anxiety measure 0.52. Correlation with parent and teacher ratings poor (0.10 – 0.30).

Author	Year	Sample	Comments
Gjerde, Block & Block	1988	U.S. (California) adolescents. N = 106, Age: 18.	Modified CES-D - questions intermixed in long survey. Boys = 19.77, SD = 10.75; Girls = 22.50, SD = 11.10. Dysthymic young males more negatively evaluated than dysthymic young females.
Garrison, Schluchter, Schoenbach & Kaplan	1989	U.S. (North Carolina) junior high school students. N = 677, Age: 12-15.	White Boys = 15.21, SD = 9.24; White Girls = 16.54, SD = 9.40. Alpha = 0.86. Three month test-retest r = 0.61. Minority race, lower SES and family constellation associated with higher CES-D scores.
Allgood-Merten, Lewinsohn & Hops	1990	U.S. (Oregon) high school students. N = 802, Age: 13-18.	Gender means not shown but girls reported to have higher means than boys. Overall mean = 19.12. Alpha = 0.91. One month test-retest r = 0.59
Garrison, Jackson, Marsteller, McKeown & Addy	1990	U.S. (Southeast) high school students. N = 550, Age: 12-13.	White boys = 13.98, SD = 8.52; White girls = 15.80, SD = 9.58. One and two year test-retest r = 0.53 and 0.35. Black students had higher scores than White students. Best predictor of subsequent CES-D score was previous year's score.
Hops, Lewinsohn, Andrews & Roberts	1990	U.S. (Oregon) high students. N = 2160, Age: 16.	Boys range: 13.4 to 19.3; Girls range: 18.0 to 21.0. CES-D scores correlated with a wide range of difficulties: anxiety, suicidal ideation etc.
Manson, Ackerson, Dick, Baron & Fleming	1990	U.S. (Southeastern) Indian boarding school adolescents. N = 188, Age: 15-17.	Boys = 16.7, SD = 8.0; Girls = 21.7, SD = 10.0. Alpha = 0.82. Three factor model preferred to four factor model. No gender difference observed in regard to positive affect items.
Roberts, Andrews, Lewinsohn & Hops	1990a	U.S. (Oregon) high school students. N = 2160, Age: 16.	Gender means not shown. Alpha > 0.87. One month test re-test correlation r > 0.50. Four factor CES-D tested in CFA and provided good fit to data. Factor loadings invariant across gender except for items *Cry* and *Appetite*. Mean rank order of items similar for boys and girls.
Colten, Gore & Aseltine	1991	U.S. (Boston) high school students. N = 1033, Age: 15-18.	Second wave results of longitudinal study. Boys = 17.9; Girls = 29.2. Girls with strong interpersonal caring orientation experienced higher CES-D scores.
Garrison, Addy, Jackson, McKeown & Waller	1991a	U.S. (South Carolina) high school students. N = 1073, Age: 12-14.	Gender means not shown but girls reported to have higher means than boys. CES-D score predicted suicidal behaviour.

Author	Year	Sample	Comments
Garrison, Addy, Jackson, McKeown & Waller	1991b	U.S. (South Carolina) high school students. N = 2465, Age: 12-14	Gender means not shown. Mean CES-D scores higher in group with clinical depression. Optimal cut-points for screening 12 for males and 22 for females.
Garrison, Jackson, Addy, McKeown & Waller	1991c	U.S. (South Carolina) high school students. N = 226, Age: 12-14.	Diagnostic sample of high scorers. White Boys = 18.90, SD = 11.20; White Girls = 26.84, SD = 13.24. Significant correlation between depression and suicidal ideation and attempts.
Gjerde & Block	1991	U.S. (California) adolescents. N = 106, Age: 16.	Modified CES-D - questions intermixed in long survey. Boys = 19.77, SD = 10.75; Girls = 22.50, SD = 11.10.
Paikoff, Brooks-Gunn & Warren	1991	U.S. (New York) female high school students. N = 72, Age: 14.	Girls = 32, SD = 9.13. CES-D scored incorrectly using (1,2,3,4) response format. Hormonal levels in girls associated with affective expression.
Radloff	1991	U.S. (Maryland) high school students. N = 637, Grade: 10-12.	Gender means not shown but overall means of 16.60, SD = 9.19 and 17.88, SD = 10.31 reported. Alpha: 0.85 – 0.86. Scores in junior high school sample said to be inflated by transient symptoms and an excess of interpersonal and affective symptoms.
Roberts, Lewinsohn & Seeley	1991	U.S. (Oregon) high school students. N = 1710, Age: 16.	Boys = 15.70; Girls = 18.12. Overall mean = 16.98, SD = 10.5. Alpha = 0.89. One month test-retest r = 0.48. Psychometric properties of the CES-D and its specificity to detect cases of clinical depression in adolescent samples similar to when used with adults.
Avison & McAlpine	1992	Canadian high school students. N = 306, Age: 17.	Boys = 15.45, SD = 9.82; Girls = 18.98, SD = 11.86. Alpha = 0.90. Elevated CES-D scores in girls not the result of stressful experiences.
Berganza & Aguilar	1992	Guatemalan high school students. N = 339, Age: 15.	Used modified CES-D termed the CES-DCM. Boys = 15.69, SD = 7.43; Girls = 20.78, SD = 8.91. Social class not related to CES-D scores.
Gore, Aseltine & Colten	1992	U.S. (Boston) high school students. N = 1208, Grade: 9-11.	First wave results of longitudinal study. Gender means not shown. Girls from families with low educational backgrounds and children from single-parent families had highest CES-D scores.

Author	Year	Sample	Comments
Roberts & Sobhan	1992	U.S. (National Survey) N = 2250, Age: 12-17	Used 12 item modified version of CES-D. Estimated means for 20 item CES-D: Anglo Boys = 10.16; Anglo Girls = 11.50. Mexican American adolescents reported higher levels of depressive symptomatology than the Anglo majority.
Swanson, Linskey, Quintero-Salinas, Pumariega & Holzer	1992	U.S. (Texas) and Mexican high school students. N = 4157, Age: 12-17.	Gender means not shown but greater proportion of girls than boys with scores >16. Alpha = 0.88 and 0.80.
Andrews, Lewinsohn, Hops & Roberts	1993	U.S. (Oregon) high school students. N = 2378, Age: 14-18.	Gender means not shown but females higher than males. Alpha = 0.75. One month test-retest r = 0.57.
Aseltine & Gore	1993	U.S. (Boston) high school students. N = 1576, Grade: 9-11.	Second and third wave results of longitudinal study. Gender means not shown. Lower CES-D scores and improved relations with parents following graduation from high school.
Blatt, Hart, Quinlan, Leadbeater & Auerbach	1993	U.S. (New York) high school students. N = 610, Grade: 9-12.	Used CES-DC. Boys = 16.74; Girls = 25.25. CES-DC scores correlated with measures of problem behaviours.
Clarke, Hawkins, Murphy, & Sheeber.	1993	U.S. (Oregon) high school students N = 513, Age: 15.	Calculated values: Boys = 14.74, SD = 10.6; Girls = 19.36, SD = 13.1. A three and five session educational intervention failed to demonstrate long term (12 week) reductions in levels of depressive symptomatology.
Dick, Manson & Beals	1993	U.S. (Southeastern) Indian boarding school adolescents. N = 188, Age: 15-17.	Gender means not shown but total mean = 19.18, SD = 9.52. CES-D scores showed modest correlation to alcohol use.
Gore, Aseltine & Colten	1993	U.S. (Boston) high school students. N = 1208, Grade: 9-11.	First wave results of longitudinal study. Boys = 11.2, SD = 7.5; Girls = 14.6, SD = 9.1. Alpha = 0.87. Girls with strong interpersonal caring orientation or involvement in family stresses experienced higher CES-D scores.
Gotlib, Lewinsohn, Seeley, Rohde & Redner	1993	U.S. (Oregon) high school students. N = 1710, Age: 14-18.	Boys = 15.71; Girls = 18.14. Alpha = 0.89. Measures of negative cognitions and attributional style related to diagnosis of depression.
Aseltine, Gore & Colten	1994	U.S. (Boston) high school students. N = 1576, Grade: 9-11.	Gender means not shown. Overall means across three year longitudinal study; Wave 1: 12.9, Wave 2: 11.6, Wave 3: 9.9. *Cry* item not included in scale.

Author	Year	Sample	Comments
Dick, Beals, Keane & Manson	1994	U.S. (Southeastern) Indian boarding high school students. N = 188, Age: 15.	Gender means not shown. Overall mean = 18.82, SD = 10.75. Three factor model better than four factor model. No gender differences with respect to factor loadings.
Iwata, Saito & Roberts	1994	Japanese (Gotemba) high school students. N = 1500, Age: 12-15.	Gender means not shown. Alpha = 0.81. Symptom presence on negatively worded items more common for females but this pattern not clear for predominance or persistence. Three times as many females responded to the *Cry* item than males.
Killen, Hayward, Wilson, Taylor, Hammer, Litt, Simmonds & Haydel	1994	U.S. (California) middle school students. N = 939, Age: 13.	Girls with eating disorder 23.8, Girls without eating disorder 17.9. Boys means not shown.
Lewinsohn, Rohde & Seeley	1994	U.S. (Oregon) high school students. N = 1508, Age: 15.	Gender means not shown. CES-D scores associated with future suicide attempts.
Rohde, Lewinsohn & Seeley	1994	U.S. (Oregon) clinically depressed adolescents. N = 115, Age: 16.	Gender means not shown. Alpha = 0.86. Pre-treatment CES-D scores not associated with response to treatment.
Clarke, Hawkins, Murphy, Sheeber, Lewinsohn, & Seeley	1995	U.S. (Oregon) high school students N = 150, Age: 15.	Gender means not shown. Group of high CES-D scoring adolescents participated in a coping with stress program. Survival analyses indicated a reduction in rates of clinical depression in the 12 months following the intervention.
Gore & Aseltine	1995	U.S. (Boston) high school students. N = 1208, Grade: 9-11.	Gender means not shown but females reported to be higher than males. Alpha = 0.89. Friendship related changes associated with change in depressed mood. Peer support for boys buffered impact of stress but for girls amplified emotional response to stress.
Gotlib, Lewinsohn & Seeley	1995	U.S. (Oregon) high school students. N = 1709, Age: 16.	Gender means not shown. Adolescents with elevated CES-D scores but do not meet criteria for clinical depression at similar risk for dysfunction as adolescents who do meet criteria.
Reifman & Windle	1995	U.S. (New York) high school students. N = 662, Age: 16.	Gender means not shown. Alpha = 0.90. CES-D scores predicted later suicidal behaviour.

Author	Year	Sample	Comments
Roberts & Chen	1995	New Mexico (Las Cruces) high school students. N = 2614, Age: 11-14.	Anglo Boys = 13.1; Anglo Girls = 15.8. Alpha = 0.92. Adolescents of Mexican origin higher means than their Anglo counterparts. CES-D scores correlated with suicidal ideation, loneliness and use of English.
Silverstein, Caceres, Perdue & Cimarolli	1995	U.S. (New York) senior high school students. N = 175, Age: 17-19.	Gender means not shown. Female adolescents more likely than males to report high CES-D scores accompanied by anxiety and somatic symptoms but not more likely to report high levels of depressive symptomatology unaccompanied by these symptoms.
Aseltine	1996	U.S. (Boston) high school students. N = 942, Grade: 9-11.	Gender means not shown. Parental divorce associated with increased levels of depressive symptomatology.
Dumenci & Windle	1996	U.S. (New York) high students. N = 805, Age: 16.	Gender means not shown but reported to be significantly different.
Lasko, Field, Gonzalez, Harding, Yando, & Bendell	1996	U.S. high school students. N = 455, Age: 15.	Boys = 17.9; Girls = 24.0. Alpha = 0.79. One month test re-test r = 0.79. Sample three quarters Hispanic or Black. Less intimacy with parents associated with higher CES-D scores.
Marcotte	1996	Canadian (Quebec) high school students. N = 349, Age: 11-18.	Boys = 13.15, SD = 8.69; Girls = 18.18, SD = 10.98. Showed sharp increase in average CES-D scores at age 15. Found relationship between cognitive distortions and CES-D scores.
Pumariega, Johnson, Sheridan & Cuffe	1996	U.S. (South Carolina) adolescents in residential group homes. N = 299, Age 12-17.	Gender means not shown but a higher proportion of females with scores above 16 compared with males. CES-D scores not found to be correlated with number of placements in out-of-home programs.
Gladstone, Kaslow, Seeley & Lewinsohn	1997	U.S. (Oregon) senior high school student. N = 1661, Age: 17.	Gender means not shown. Alpha = 0.89. CES-D scores correlated with attributional style.
McKeown, Garrison, Jackson, Cuffe, Addy & Waller	1997	U.S. (South Carolina) high school students. N = 3191, Age: 12-14.	First two years of longitudinal data. Gender means presented by race and family structure. Females in all groups higher means than males. Best predictor of CES-D score at Year 2 was Year 1 CES-D score. Levels of perceived family emotional bonding predicted levels of depressive symptomatology.

Author	Year	Sample	Comments
Sheeber, Hops, Alpert, Davis & Andrews	1997	U.S. (Oregon) high school students. N = 421, Age: 16.	Boys = 15.40, SD = 9.80; Girls = 17.92, SD = 11.17. Less supportive and conflictual family environments associated with higher CES-D scores both concurrently and prospectively over a one year period.
Windle & Windle	1997	U.S. (New York) high school students. N = 975, Age: 15.	Boys = 13.65, SD = 9.47; Girls = 16.11, SD = 10.55. Alpha > 0.90. Higher CES-D scores associated with suicidal ideation and suicide attempt.
Aseltine, Gore & Colten	1998	U.S. (Boston) high school students. N = 898, Grade: 9-11.	Results from longitudinal study (Waves 1 and 2) reported. Gender means reported in earlier paper. Risk for high CES-D scores associated with lack of support in family and peer supports.
Chen, Mechanic & Hansell	1998	U.S. (New Jersey) high school students. N = 479, Age: 13 at Wave 1 of study.	Gender means not shown but girls higher means than boys for first three years of study but not in the fourth year. No gender difference in the link between self-awareness and depressed mood.
Gjerde & Westenberg	1998	U.S. (California) adolescents. N = 106, Age: 18-23.	Modified CES-D - questions intermixed in long survey. At age 18 Boys = 19.77, SD = 10.75; Girls = 22.50, SD = 11.10. Elevated CES-D scores at age 18 predicted chronic depressive symptoms and suicidal ideation at age 23.
Langhinrichsen-Rohling, Lewinsohn, Rohde, Seeley, Monson, Meyer & Langford	1998	U.S. (California) high school students. N = 206, Grade: all high school.	Boys = 7.20; Girls = 8.89. Males showed more risk-taking, injury producing and negative health behaviours than females.
McKeown, Garrison, Cuffe, Waller, Jackson & Addy	1998	U.S. (South Carolina) high school students. N = 359, Age: 12-14.	Sample from Garrison et al. (1991) study. No gender means shown but overall mean = 14.75. Adolescents engaging in suicidal behaviours had higher CES-D one year earlier than those who did not engage in such behaviours.
Prescott, McArdle, Hishinuma, Johnson, Miyamoto, Andrade, Edman, Makini, Nahulu, Yuen & Carlton	1998	Hawaiian high school students. N = 2500, Age: 16.	Gender means not shown but girls reported to have higher means than boys. Overall mean = 14.1. No evidence that different screening cut-points are required for males and females: no gender difference in the strength of association between CES-D score and the probability of a diagnosis of depression.

Author	Year	Sample	Comments
Greenberger, Chen, Tally & Dong	2000	U.S. (Los Angles) and Chinese (Tianjin) high school students. N = 703, Age: 17	Modified version of CES-D used (five point scale). Females higher means than males in both samples. Method for calculating means not clear (range between 1.80 and 2.06) – these may represent item means rather than total scores.
Holsen, Kraft & Vitterso	2000	Norwegian high school students. N = 538, Age: 13-19.	Longitudinal study. Gender means not shown. CES-D strongly correlated (0.82) with a depressed mood scale. Girls had higher depressed mood scores than boys. Girls scores increased to peak in mid-adolescence but boys scores relatively stable.
Lewinsohn, Rohde & Farrington	2000	U.S. (Oregon) senior high school students N = 1709, Age: 14-18.	Gender means not shown. CES-D screening characteristics for Conduct Disorder poor.
Roeger, Allison, Martin, Dadds & Keeves	2001	Australia (SA) high school students N = 2489, Age: 13 at Wave 1 of study	Gender means not shown. Sample drawn from 26 schools. Size of school effect shown to be small indicating that differences between school environments do not exert large influence on student mental health. Used the EDED data set of the present study.
Allison, Roeger, Martin & Keeves	2001	Australia (SA) high school students N = 2489 Age: 13 at Wave 1 of study	Boys = 11.4, SD = 9.0; Girls = 13.9, SD = 11.6. Higher CES-D scores associated with increased risk of suicidal ideation. At moderate CES-D scores females at greater risk of ideation than males. Used the EDED data set of the present study.

B

Cumulative proportions

Table 65 Proportion of boys and girls endorsing response options from CES-D items (Year 8)

	CES-D Item	Response Option	% Boys	% Girls	Cumulative % Boys	Cumulative % Girls	Difference Total
1	Bothered	0	78.1	66.8			
		1	15.4	22.2	21.9	33.3	**11.4**
		2	3.5	8.0	6.5	11.1	4.6
		3	3.0	3.1	3.0	3.1	0.1
2	Appetite	0	80.5	63.2			
		1	13.2	21.3	19.5	36.9	**17.4**
		2	4.3	9.0	6.3	15.6	9.3
		3	2.0	6.6	2.0	6.6	4.6
3	Blues	0	78.6	68.7			
		1	13.4	15.6	21.4	31.4	**10.0**
		2	4.2	9.2	8.0	15.8	7.8
		3	3.8	6.6	3.8	6.6	2.8
4	Good	0	42.8	35.3			
		1	25.7	28.0	57.2	64.7	7.5
		2	14.6	19.1	31.5	36.7	5.2
		3	16.9	17.6	16.9	17.6	0.7

Table 66 Proportion of boys and girls endorsing response options from CES-D items (Year 8) (continued)

CES-D Item	Response Option	% Boys	% Girls	Cumulative % Boys	Cumulative % Girls	Difference Total
5 Mind	0	45.0	45.7			
	1	32.1	31.1	55.0	54.4	-0.6
	2	14.5	14.3	22.9	23.3	0.4
	3	8.4	9.0	8.4	9.0	0.6
6 Depress	0	65.9	56.8			
	1	22.0	22.5	34.1	43.2	9.1
	2	7.2	11.6	12.1	20.7	8.6
	3	4.9	9.1	4.9	9.1	4.2
7 Effort	0	36.4	45.3			
	1	23.7	25.5	63.5	54.7	-8.8
	2	23.2	18.4	39.8	29.2	**-10.6**
	3	16.6	10.8	16.6	10.8	-5.8
8 Hopeful	0	32.4	28.3			
	1	28.5	32.9	67.6	71.7	4.1
	2	20.3	22.6	39.1	38.8	-0.3
	3	18.8	16.2	18.8	16.2	-2.6
9 Failure	0	82.6	78.4			
	1	9.0	10.1	17.3	21.6	4.3
	2	4.8	5.7	8.3	11.5	3.2
	3	3.5	5.8	3.5	5.8	2.3
10 Fearful	0	82.6	75.9			
	1	11.1	15.1	17.4	24.1	6.7
	2	4.3	5.4	6.3	9.0	2.7
	3	2.0	3.6	2.0	3.6	1.6
11 Sleep	0	65.7	53.5			
	1	20.1	25.0	34.3	46.5	**12.2**
	2	8.6	11.7	14.2	21.5	7.3
	3	5.6	9.8	5.6	9.8	4.2

Table 67 Proportion of boys and girls endorsing response options from CES-D items (Year 8) (continued)

CES-D Item	Response Option	% Boys	% Girls	Cumulative % Boys	Cumulative % Girls	Difference Total
12 Happy	0	49.5	52.3			
	1	31.0	24.2	50.5	47.7	-2.8
	2	10.7	14.5	19.5	23.5	4.0
	3	8.8	9.0	8.8	9.0	0.2
13 Talk	0	63.5	60.3			
	1	22.6	22.6	36.6	39.7	3.1
	2	9.9	11.7	14.0	17.1	3.1
	3	4.1	5.4	4.1	5.4	1.3
14 Lonely	0	76.1	68.6			
	1	14.9	16.1	23.9	31.4	7.5
	2	5.1	8.1	9.0	15.3	6.3
	3	3.9	7.2	3.9	7.2	3.3
15 Unfriendly	0	63.6	67.4			
	1	23.3	21.0	36.4	32.6	-3.8
	2	8.6	7.9	13.1	11.6	-1.5
	3	4.5	3.7	4.5	3.7	-0.8
16 Enjoy	0	50.6	47.7			
	1	27.2	26.5	49.4	52.4	3.0
	2	12.2	14.1	22.2	25.9	3.7
	3	10.0	11.8	10.0	11.8	1.8
17 Cry	0	90.5	75.2			
	1	5.8	15.3	9.6	24.7	**15.1**
	2	2.1	5.4	3.8	9.4	5.6
	3	1.7	4.0	1.7	4.0	2.3
18 Sad	0	72.8	59.2			
	1	19.0	25.0	27.2	40.8	**13.6**
	2	5.2	8.4	8.2	15.8	7.6
	3	3.0	7.4	3.0	7.4	4.4

Table 68 Proportion of boys and girls endorsing response options from CES-D items (Year 8) (continued)

CES-D Item	Response Option	% Boys	% Girls	Cumulative % Boys	Cumulative % Girls	Difference Total
19 Dislike	0	67.7	57.8			
	1	20.5	25.6	32.4	42.2	**9.8**
	2	7.6	8.8	11.9	16.6	4.7
	3	4.3	7.8	4.3	7.8	3.5
20 Get-going	0	65.8	60.8			
	1	22.6	24.5	34.2	39.3	5.1
	2	7.4	9.1	11.6	14.8	3.2
	3	4.2	5.7	4.2	5.7	1.5

Cumulative difference totals ≥10 shown in bold

Table 69 Proportion of boys and girls endorsing response options from CES-D items (Year 9)

	CES-D Item	Response Option	% Boys	% Girls	Cumulative % Boys	Cumulative % Girls	Difference Total
1	Bothered	0	78.8	60.8			
		1	15.6	25.7	21.3	39.2	**17.9**
		2	3.4	9.1	5.7	13.5	7.8
		3	2.3	4.4	2.3	4.4	2.1
2	Appetite	0	82.3	64.0			
		1	11.0	20.6	17.6	36.0	**18.4**
		2	4.4	10.5	6.6	15.4	8.8
		3	2.2	4.9	2.2	4.9	2.7
3	Blues	0	80.8	64.9			
		1	11.3	19.9	19.2	35.1	**15.9**
		2	5.0	8.7	7.9	15.2	7.3
		3	2.9	6.5	2.9	6.5	3.6
4	Good	0	45.5	35.8			
		1	26.8	29.9	54.4	64.2	**9.8**
		2	13.7	20.2	27.6	34.3	6.7
		3	13.9	14.1	13.9	14.1	0.2
5	Mind	0	41.8	38.3			
		1	33.5	32.1	58.3	61.7	3.4
		2	16.9	18.7	24.8	29.6	4.8
		3	7.9	10.9	7.9	10.9	3.0
6	Depress	0	70.4	54.6			
		1	18.0	25.1	29.6	45.3	**15.7**
		2	7.2	13.1	11.6	20.2	8.6
		3	4.4	7.1	4.4	7.1	2.7
7	Effort	0	42.3	47.0			
		1	23.4	31.2	57.7	53.0	-4.7
		2	20.8	14.1	34.3	21.8	**-12.5**
		3	13.5	7.7	13.5	7.7	-5.8

Table 70 Proportion of boys and girls endorsing response options from CES-D items (Year 9) (continued)

CES-D Item	Response Option	% Boys	% Girls	Cumulative % Boys	Cumulative % Girls	Difference Total
8 Hopeful	0	32.3	28.2			
	1	34.3	35.4	67.7	71.9	4.2
	2	18.5	22.7	33.4	36.5	3.1
	3	14.9	13.8	14.9	13.8	-1.1
9 Failure	0	83.5	78.1			
	1	10.2	13.5	16.5	21.9	5.4
	2	3.7	4.2	6.3	8.4	2.1
	3	2.6	4.2	2.6	4.2	1.6
10 Fearful	0	82.3	77.2			
	1	12.3	15.8	17.6	22.7	5.1
	2	3.9	4.5	5.3	6.9	1.6
	3	1.4	2.4	1.4	2.4	1.0
11 Sleep	0	64.9	53.4			
	1	20.5	26.9	35.1	46.6	**11.5**
	2	8.9	11.4	14.6	19.7	5.1
	3	5.7	8.3	5.7	8.3	2.6
12 Happy	0	47.4	49.6			
	1	32.4	31.0	52.6	50.5	-2.1
	2	11.9	13.1	20.2	19.5	-0.7
	3	8.3	6.4	8.3	6.4	-1.9
13 Talk	0	64.9	58.4			
	1	23.3	27.5	35.2	41.7	6.5
	2	8.5	10.3	11.9	14.2	2.3
	3	3.4	3.9	3.4	3.9	0.5
14 Lonely	0	78.0	66.1			
	1	14.1	20.3	22.0	33.9	**11.9**
	2	4.7	8.5	7.9	13.6	5.7
	3	3.2	5.1	3.2	5.1	1.9

Table 71 Proportion of boys and girls endorsing response options from CES-D items (Year 9) (continued)

CES-D Item	Response Option	% Boys	% Girls	Cumulative % Boys	Cumulative % Girls	Difference Total
15 Unfriendly	0	67.6	70.6			
	1	23.6	20.5	32.4	29.4	-3.0
	2	6.2	6.1	8.8	8.9	0.1
	3	2.6	2.8	2.6	2.8	0.2
16 Enjoy	0	49.4	46.8			
	1	28.5	28.8	50.6	53.2	2.6
	2	12.0	14.4	22.1	24.4	2.3
	3	10.1	10.0	10.1	10.0	-0.1
17 Cry	0	92.7	72.7			
	1	4.6	17.5	7.4	27.2	**19.8**
	2	1.5	7.2	2.8	9.7	6.9
	3	1.3	2.5	1.3	2.5	1.2
18 Sad	0	77.7	58.7			
	1	15.3	27.0	22.3	41.3	**19.0**
	2	4.5	10.2	7.0	14.3	7.3
	3	2.5	4.1	2.5	4.1	1.6
19 Dislike	0	69.4	61.5			
	1	22.6	25.9	30.6	38.5	7.9
	2	5.2	7.2	8.0	12.6	4.6
	3	2.8	5.4	2.8	5.4	2.6
20 Get-going	0	63.9	58.6			
	1	24.4	28.3	36.1	41.4	5.3
	2	7.0	8.7	11.7	13.1	1.4
	3	4.7	4.4	4.7	4.4	-0.3

Cumulative difference totals ≥10 shown in bold

Table 72 Proportion of boys and girls endorsing response options from CES-D items (Year 10)

	CES-D Item	Response Option	% Boys	% Girls	Cumulative % Boys	Cumulative % Girls	Difference Total
1	Bothered	0	74.6	54.8			
		1	17.3	30.7	25.5	45.3	**19.8**
		2	5.1	10.4	8.2	14.6	6.4
		3	3.1	4.2	3.1	4.2	1.1
2	Appetite	0	81.6	56.7			
		1	12.0	26.1	18.4	43.4	**25.0**
		2	3.8	11.7	6.4	17.3	**10.9**
		3	2.6	5.6	2.6	5.6	3.0
3	Blues	0	82.1	62.5			
		1	10.8	20.1	17.9	37.5	**19.6**
		2	4.8	10.7	7.1	17.4	**10.3**
		3	2.3	6.7	2.3	6.7	4.4
4	Good	0	47.6	35.3			
		1	28.5	31.8	52.3	64.7	**12.4**
		2	12.8	19.9	23.8	32.9	9.1
		3	11.0	13.0	11.0	13.0	2.0
5	Mind	0	39.6	31.6			
		1	36.4	37.5	60.4	68.4	8.0
		2	16.4	19.6	24.0	30.9	6.9
		3	7.6	11.3	7.6	11.3	3.7
6	Depress	0	70.2	50.7			
		1	18.9	29.2	29.9	49.4	**19.5**
		2	7.4	13.2	11.0	20.2	9.2
		3	3.6	7.0	3.6	7.0	3.4
7	Effort	0	46.9	49.3			
		1	25.0	29.0	53.1	50.7	-2.4
		2	17.6	14.5	28.1	21.7	-6.4
		3	10.5	7.2	10.5	7.2	-3.3

Table 73 Proportion of boys and girls endorsing response options from CES-D items (Year 10) (continued)

CES-D Item	Response Option	% Boys	% Girls	Cumulative % Boys	Cumulative % Girls	Difference Total
8 Hopeful	0	31.2	28.0			
	1	36.3	34.5	68.8	71.9	3.1
	2	17.9	23.7	32.5	37.4	4.9
	3	14.6	13.7	14.6	13.7	-0.9
9 Failure	0	86.3	77.0			
	1	8.4	13.9	13.6	23.0	9.4
	2	3.2	4.8	5.2	9.1	3.9
	3	2.0	4.3	2.0	4.3	2.3
10 Fearful	0	84.6	76.7			
	1	10.3	15.6	15.4	23.3	7.9
	2	3.6	4.8	5.1	7.7	2.6
	3	1.5	2.9	1.5	2.9	1.4
11 Sleep	0	66.4	50.4			
	1	20.0	29.3	33.6	49.7	**16.1**
	2	8.7	11.6	13.6	20.4	6.8
	3	4.9	8.8	4.9	8.8	3.9
12 Happy	0	48.7	47.0			
	1	34.3	30.9	51.3	53.0	1.7
	2	11.2	16.1	17.0	22.1	5.1
	3	5.8	6.0	5.8	6.0	0.2
13 Talk	0	64.4	54.8			
	1	24.3	32.3	35.6	45.2	**9.6**
	2	9.0	9.5	11.3	12.9	1.6
	3	2.3	3.4	2.3	3.4	1.1
14 Lonely	0	77.2	64.2			
	1	14.8	22.6	22.8	35.9	**13.1**
	2	5.7	8.9	8.0	13.3	5.3
	3	2.3	4.4	2.3	4.4	2.1

Table 74 Proportion of boys and girls endorsing response options from CES-D items (Year 10) (continued)

CES-D Item	Response Option	% Boys	% Girls	Cumulative % Boys	Cumulative % Girls	Difference Total
15 Unfriendly	0	73.1	74.2			
	1	19.4	17.3	26.9	25.8	-1.1
	2	4.8	5.6	7.5	8.5	1.0
	3	2.7	2.9	2.7	2.9	0.2
16 Enjoy	0	48.8	44.6			
	1	31.6	31.7	51.2	55.4	4.2
	2	11.4	13.8	19.6	23.7	4.1
	3	8.2	9.9	8.2	9.9	1.7
17 Cry	0	94.1	71.8			
	1	3.7	19.1	5.9	28.2	**22.3**
	2	1.1	6.0	2.2	9.1	6.9
	3	1.1	3.1	1.1	3.1	2.0
18 Sad	0	77.5	53.2			
	1	16.1	31.0	22.4	46.8	**24.4**
	2	4.3	11.1	6.3	15.8	**9.5**
	3	2.0	4.7	2.0	4.7	2.7
19 Dislike	0	73.1	60.9			
	1	20.2	26.5	27.0	39.1	**12.1**
	2	4.3	7.8	6.8	12.6	5.8
	3	2.5	4.8	2.5	4.8	2.3
20 Get-going	0	63.0	54.7			
	1	25.3	28.9	37.0	45.4	8.4
	2	8.5	12.4	11.7	16.5	4.8
	3	3.2	4.1	3.2	4.1	0.9

Cumulative difference totals ≥10 shown in bold

Table 75 Proportion of boys and girls endorsing response options from CES-D items (all year levels)

	CES-D Item	Response Option	% Boys	% Girls	Cumulative % Boys	Cumulative % Girls	Difference Total
1	Bothered	0	77.2	61.0			
		1	16.1	26.0	22.9	39.0	**16.1**
		2	4.0	9.1	6.8	13.0	6.2
		3	2.8	3.9	2.8	3.9	1.1
2	Appetite	0	81.5	61.4			
		1	12.1	22.6	18.6	38.6	**20.0**
		2	4.2	10.3	6.5	16.0	**9.5**
		3	2.3	5.7	2.3	5.7	3.4
3	Blues	0	80.5	65.5			
		1	11.8	18.5	19.5	34.6	**15.1**
		2	4.7	9.5	7.7	16.1	8.4
		3	3.0	6.6	3.0	6.6	3.6
4	Good	0	45.3	35.4			
		1	27.0	29.8	54.7	64.5	**9.8**
		2	13.7	19.7	27.7	34.7	7.0
		3	14.0	15.0	14.0	15.0	1.0
5	Mind	0	42.2	38.8			
		1	34.0	33.4	57.9	61.2	3.3
		2	15.9	17.4	23.9	27.8	3.9
		3	8.0	10.4	8.0	10.4	2.4
6	Depress	0	68.8	54.2			
		1	19.6	25.5	31.1	45.8	**14.7**
		2	7.2	12.6	11.5	20.3	8.8
		3	4.3	7.7	4.3	7.7	3.4
7	Effort	0	41.8	47.1			
		1	24.0	28.5	58.1	52.8	-5.3
		2	20.5	15.7	34.1	24.3	**-9.8**
		3	13.6	8.6	13.6	8.6	-5.0

Table 76 Proportion of boys and girls endorsing response options from CES-D items (all year levels) (continued)

CES-D Item	Response Option	% Boys	% Girls	Cumulative % Boys	Cumulative % Girls	Difference Total
8 Hopeful	0	32.0	28.2			
	1	33.0	34.2	68.0	71.8	3.8
	2	18.9	23.0	35.0	37.6	2.6
	3	16.1	14.6	16.1	14.6	-1.5
9 Failure	0	84.2	77.9			
	1	9.2	12.4	15.8	22.1	6.3
	2	3.9	4.9	6.6	9.7	3.1
	3	2.7	4.8	2.7	4.8	2.1
10 Fearful	0	83.2	76.6			
	1	11.2	15.5	16.8	23.4	6.6
	2	4.0	4.9	5.6	7.9	2.3
	3	1.6	3.0	1.6	3.0	1.4
11 Sleep	0	65.6	52.5			
	1	20.2	27.0	34.3	47.5	**13.2**
	2	8.7	11.5	14.1	20.5	6.4
	3	5.4	9.0	5.4	9.0	3.6
12 Happy	0	48.5	49.7			
	1	32.6	28.6	51.5	50.3	-1.2
	2	11.3	14.6	18.9	21.7	2.8
	3	7.6	7.1	7.6	7.1	-0.5
13 Talk	0	64.3	57.9			
	1	23.4	27.3	35.8	42.1	6.3
	2	9.1	10.5	12.4	14.8	2.4
	3	3.3	4.3	3.3	4.3	1.0
14 Lonely	0	77.1	66.4			
	1	14.6	19.6	22.9	31.7	**10.8**
	2	5.2	8.5	8.3	14.1	5.8
	3	3.1	5.6	3.1	5.6	2.5

Table 77 Proportion of boys and girls endorsing response options from CES-D items (all year levels) (continued)

CES-D Item	Response Option	% Boys	% Girls	Cumulative % Boys	Cumulative % Girls	Difference Total
15 Unfriendly	0	68.0	70.6			
	1	22.1	19.7	32.0	29.3	-2.7
	2	6.6	6.5	9.9	9.6	-0.3
	3	3.3	3.1	3.3	3.1	-0.2
16 Enjoy	0	49.6	46.4			
	1	29.1	28.9	50.5	53.6	3.1
	2	11.9	14.1	21.4	24.7	3.3
	3	9.5	10.6	9.5	10.6	1.1
17 Cry	0	92.4	73.3			
	1	4.7	17.3	7.6	26.7	**19.1**
	2	1.6	6.2	2.9	9.4	6.5
	3	1.3	3.2	1.3	3.2	1.9
18 Sad	0	76.0	57.2			
	1	16.8	27.5	24.0	42.7	**18.7**
	2	4.7	9.8	7.2	15.2	8.0
	3	2.5	5.4	2.5	5.4	2.9
19 Dislike	0	70.0	60.0			
	1	21.1	26.0	30.0	39.9	**9.9**
	2	5.7	7.9	8.9	13.9	5.0
	3	3.2	6.0	3.2	6.0	2.8
20 Get-going	0	64.3	58.1			
	1	24.1	27.2	35.7	41.9	6.2
	2	7.6	10.0	11.6	14.7	3.1
	3	4.0	4.7	4.0	4.7	0.7

Cumulative difference totals ≥10 shown in bold

Table 78 Proportion of low and high scoring cases endorsing response options from CES-D items (Boys – All year levels)

	CES-D Item	Response Option	% Low	% High	Cumulative % Low	Cumulative % High	Difference Total
1	Bothered	0	82.4	34.9			
		1	14.7	27.2	17.6	65.2	47.6
		2	2.0	20.2	2.9	38.0	35.1
		3	0.9	17.8	0.9	17.8	16.9
2	Appetite	0	85.8	46.9			
		1	10.4	25.7	14.2	53.1	38.9
		2	2.7	16.1	3.8	27.4	23.6
		3	1.1	11.3	1.1	11.3	10.2
3	Blues	0	87.5	24.5			
		1	9.8	28.1	12.5	75.5	**63.0**
		2	1.9	26.7	2.7	47.4	44.7
		3	0.8	20.7	0.8	20.7	19.9
4	Good	0	49.3	8.4			
		1	27.9	24.5	50.7	87.1	36.4
		2	11.3	34.6	22.8	66.9	44.1
		3	11.5	32.5	11.5	33.7	22.2
5	Mind	0	46.4	8.4			
		1	35.1	24.5	53.6	91.6	38.0
		2	13.6	34.6	18.5	67.1	48.6
		3	4.9	32.5	4.9	32.5	27.6
6	Depress	0	76.4	7.9			
		1	19.0	24.5	23.6	92.1	**68.5**
		2	3.8	35.1	4.6	67.6	**63.0**
		3	0.8	32.5	0.8	32.5	31.7
7	Effort	0	44.7	18.8			
		1	23.1	31.7	55.3	81.3	26.0
		2	19.2	31.3	32.2	49.6	17.4
		3	13.0	18.3	13.0	18.3	5.3

Table 79 Proportion of low and high scoring cases endorsing response options from CES-D items (Boys – All year levels) (continued)

CES-D Item	Response Option	% Low	% High	Cumulative % Low	Cumulative % High	Difference Total
8 Hopeful	0	34.3	13.7			
	1	34.6	20.0	65.7	86.4	20.7
	2	17.2	32.7	31.1	66.4	35.3
	3	13.9	33.7	13.9	33.7	19.8
9 Failure	0	91.3	27.4			
	1	6.6	29.8	8.7	72.6	**63.9**
	2	1.4	23.8	2.1	42.8	40.7
	3	0.7	19.0	0.7	19.0	18.3
10 Fearful	0	88.8	37.7			
	1	9.0	29.6	11.2	62.3	**51.1**
	2	1.5	23.3	2.2	32.7	30.5
	3	0.7	9.4	0.7	9.4	8.7
11 Sleep	0	70.6	25.7			
	1	19.6	25.2	29.3	74.3	45.0
	2	6.9	23.1	9.7	49.1	39.4
	3	2.8	26.0	2.8	26.0	23.2
12 Happy	0	53.5	8.4			
	1	34.2	19.5	46.5	91.6	45.1
	2	7.5	41.6	12.3	72.1	**59.8**
	3	4.8	30.5	4.8	30.5	25.7
13 Talk	0	69.6	21.4			
	1	22.2	32.5	30.3	78.6	48.3
	2	6.5	29.8	8.1	46.1	38.0
	3	1.6	16.3	1.6	16.3	14.7
14 Lonely	0	84.7	15.9			
	1	12.7	29.8	15.3	84.1	**68.8**
	2	1.9	31.7	2.6	54.3	**51.7**
	3	0.7	22.6	0.7	22.6	21.9

Table 80 Proportion of low and high scoring cases endorsing response options from CES-D items (Boys – All year levels) (continued)

CES-D Item	Response Option	% Low	% High	Cumulative % Low	Cumulative % High	Difference Total
15 Unfriendly	0	73.5	24.3			
	1	21.1	30.5	26.5	75.7	49.2
	2	4.2	25.2	5.4	45.2	39.8
	3	1.2	20.0	1.2	20.0	18.8
16 Enjoy	0	54.7	8.4			
	1	30.6	17.1	45.2	91.6	46.4
	2	8.6	37.7	14.6	74.5	**59.9**
	3	6.0	36.8	6.0	36.8	30.8
17 Cry	0	96.3	61.3			
	1	3.0	18.0	3.7	38.7	35.0
	2	0.3	11.8	0.7	20.7	20.0
	3	0.4	8.9	0.4	8.9	8.5
18 Sad	0	83.7	14.4			
	1	14.5	35.8	16.3	85.6	**69.3**
	2	1.5	29.8	1.8	49.8	48.0
	3	0.3	20.0	0.3	20.0	19.7
19 Dislike	0	76.4	19.0			
	1	19.7	32.2	23.6	81.0	**57.4**
	2	3.1	26.4	3.9	48.8	44.9
	3	0.8	22.4	0.8	22.4	21.6
20 Get-going	0	70.4	15.1			
	1	22.7	34.9	29.6	84.9	**55.3**
	2	5.2	27.2	6.9	50.0	43.1
	3	1.7	22.8	1.7	22.8	21.1

Cumulative difference totals ≥50 shown in bold

Table 81 Proportion of low and high scoring cases endorsing response options from CES-D items (Girls – All year levels)

	CES-D Item	Response Option	% Low	% High	Cumulative % Low	Cumulative % High	Difference Total
1	Bothered	0	70.0	25.1			
		1	24.7	31.2	29.9	74.9	45.0
		2	4.6	26.9	5.2	43.7	38.5
		3	0.6	16.8	0.6	16.8	16.2
2	Appetite	0	69.4	29.7			
		1	21.5	26.7	30.6	70.2	39.6
		2	7.0	23.6	9.1	43.5	34.4
		3	2.1	19.9	2.1	19.9	17.8
3	Blues	0	78.5	13.5			
		1	16.4	26.9	21.5	86.5	**65.0**
		2	4.0	31.2	5.1	59.6	**54.5**
		3	1.1	28.4	1.1	28.4	27.3
4	Good	0	42.5	7.3			
		1	32.9	17.6	57.5	92.7	35.2
		2	16.7	31.7	24.6	75.1	**50.5**
		3	7.9	43.4	7.9	43.4	35.5
5	Mind	0	46.4	8.3			
		1	36.3	21.9	53.6	91.6	38.0
		2	13.0	35.0	17.3	69.7	**52.4**
		3	4.3	34.7	4.3	34.7	30.4
6	Depress	0	67.0	3.2			
		1	26.7	20.8	33.0	96.9	**63.9**
		2	5.8	39.4	6.3	76.1	**69.8**
		3	0.5	36.7	0.5	36.7	36.2
7	Effort	0	53.9	20.3			
		1	25.7	39.7	46.1	79.8	33.7
		2	13.5	24.8	20.4	40.1	19.7
		3	6.9	15.3	6.9	15.3	8.4

Table 82 Proportion of low and high scoring cases endorsing response options from CES-D items (Girls – All year levels) (continued)

CES-D Item	Response Option	% Low	% High	Cumulative % Low	Cumulative % High	Difference Total
8 Hopeful	0	32.4	11.5			
	1	37.6	21.1	67.7	88.5	20.8
	2	19.9	35.2	30.1	67.4	37.3
	3	10.2	32.2	10.2	32.2	22.0
9 Failure	0	90.2	28.9			
	1	7.9	30.4	9.7	71.1	**61.4**
	2	1.3	19.1	1.8	40.7	38.9
	3	0.5	21.6	0.5	21.6	21.1
10 Fearful	0	86.8	35.9			
	1	11.4	32.1	13.2	64.2	**51.0**
	2	1.2	19.6	1.8	32.1	30.3
	3	0.6	12.5	0.6	12.5	11.9
11 Sleep	0	60.2	21.9			
	1	27.6	24.8	39.9	78.2	38.3
	2	8.4	24.3	12.3	53.4	41.1
	3	3.9	29.1	3.9	29.1	25.2
12 Happy	0	60.9	5.0			
	1	30.2	22.4	39.1	94.9	**55.8**
	2	6.7	45.8	8.9	72.5	**63.6**
	3	2.2	26.7	2.2	26.7	24.5
13 Talk	0	66.8	22.4			
	1	25.1	36.2	33.2	77.6	44.4
	2	6.9	24.8	8.1	41.4	33.3
	3	1.2	16.6	1.2	16.6	15.4
14 Lonely	0	79.2	15.1			
	1	17.4	28.2	20.8	84.8	**64.0**
	2	2.6	31.7	3.4	56.6	**53.2**
	3	0.8	24.9	0.8	24.9	24.1

Table 83 Proportion of low and high scoring cases endorsing response options from CES-D items (Girls – All year levels) (continued)

CES-D Item	Response Option	% Low	% High	Cumulative % Low	Cumulative % High	Difference Total
15 Unfriendly	0	78.5	39.2			
	1	17.0	30.4	21.5	60.8	39.3
	2	3.2	19.9	4.5	30.4	25.9
	3	1.3	10.5	1.3	10.5	9.2
16 Enjoy	0	56.8	5.3			
	1	31.4	18.9	43.3	94.6	**51.3**
	2	8.1	38.0	11.9	75.7	**63.8**
	3	3.8	37.7	3.8	37.7	33.9
17 Cry	0	83.4	33.1			
	1	13.8	31.1	16.6	67.0	**50.4**
	2	2.5	20.8	2.8	35.9	33.1
	3	0.3	15.1	0.3	15.1	14.8
18 Sad	0	69.8	6.8			
	1	27.1	29.4	30.2	93.1	**62.9**
	2	2.6	38.5	3.1	63.7	**60.6**
	3	0.5	25.2	0.5	25.2	24.7
19 Dislike	0	70.5	18.3			
	1	25.0	30.1	29.5	81.8	**52.3**
	2	3.5	25.6	4.5	51.7	47.2
	3	1.0	26.1	1.0	26.1	25.1
20 Get-going	0	69.2	14.0			
	1	24.9	36.0	30.7	86.0	**55.3**
	2	5.0	29.7	5.8	50.0	44.2
	3	0.8	20.3	0.8	20.3	19.5

Cumulative difference totals ≥50 shown in bold

C

Mplus syntax examples

Program 3.1 CFA of the CES-D four factor model

```
DATA:
FILE IS bg123.DAT;
FORMAT IS 1F2.0 22F1.0;
VARIABLE: NAMES ARE sch gender wave
                dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
                dep10 dep11 dep12 dep13 dep14 dep15 dep16
                dep17 dep18 dep19 dep20;
CATEGORICAL ARE     dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
                dep10 dep11 dep12 dep13 dep14 dep15 dep16
                dep17 dep18 dep19 dep20;
USEVARIABLES ARE dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
                dep10 dep11 dep12 dep13 dep14 dep15 dep16
                dep17 dep18 dep19 dep20;

ANALYSIS:
   Type = General;
   Estimator = WLS;
   Model:      da by dep3* dep6@1 dep9 dep10 dep14 dep17 dep18;
                som by dep1* dep2 dep5 dep7 dep11 dep13 dep20@1;
                pos by dep4* dep8 dep12 dep16@1;
                inter by dep15* dep19@1;
   OUTPUT: stand;
```

Program 3.2 Second-order model of the CES-D

```
DATA:
FILE IS bg123.DAT;
FORMAT IS 1F2.0 22F1.0;
VARIABLE: NAMES ARE sch gender wave
                dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
                dep10 dep11 dep12 dep13 dep14 dep15 dep16
                dep17 dep18 dep19 dep20;
CATEGORICAL ARE     dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
                dep10 dep11 dep12 dep13 dep14 dep15 dep16
                dep17 dep18 dep19 dep20;
USEVARIABLES ARE dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
                dep10 dep11 dep12 dep13 dep14 dep15 dep16
                dep17 dep18 dep19 dep20;

ANALYSIS:
   Type = General;
   Estimator = WLS;
   Model:      da by dep3* dep6 dep9 dep10 dep14 dep17 dep18@1;
                som by dep1* dep2 dep5 dep7 dep11 dep13 dep20@1;
                pos by dep4* dep8 dep12 dep16@1;
                inter by dep15* dep19@1;
                sorder by da som pos inter;
                da@0.001;
!Note variance estimate of da constrained to prevent 'Heywood case'
   OUTPUT: stand;
```

Program 3.3 Nested CES-D model

```
DATA:
FILE IS bg123.DAT;
FORMAT IS 1F2.0 22F1.0;
VARIABLE: NAMES ARE sch gender wave
                dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
                dep10 dep11 dep12 dep13 dep14 dep15 dep16
                dep17 dep18 dep19 dep20;
CATEGORICAL ARE     dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
                dep10 dep11 dep12 dep13 dep14 dep15 dep16
                dep17 dep18 dep19 dep20;
USEVARIABLES ARE dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
                dep10 dep11 dep12 dep13 dep14 dep15 dep16
                dep17 dep18 dep19 dep20;

ANALYSIS:
   Type = General;
   Estimator = WLS;
   Model:      som by dep1* dep2 dep5 dep7 dep11 dep13 dep20 (1);
                pos by dep4* dep8 dep12 dep16 (2);
                inter by dep15* dep19 (3);
sorder by dep3* dep6 dep9 dep10 dep14 dep17 dep18 dep1 dep2 dep5 dep7 dep11
dep13 dep20 dep4 dep8 dep12 dep16 dep15 dep19;
       som with pos@0;
       som with inter@0;
       pos with inter@0;
       sorder with som@0;
       sorder with pos@0;
       sorder with inter@0;
       som@1;
       pos@1;
       inter@1;
      sorder@1;
  Output: stand ;
```

Program 3.4 Invariance covariance model (M0)

```
DATA:
 FILE IS bg123.DAT;
 FORMAT IS 1F2.0 22F1.0;
 VARIABLE: NAMES ARE sch gender wave
              dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
              dep10 dep11 dep12 dep13 dep14 dep15 dep16
              dep17 dep18 dep19 dep20;
        CATEGORICAL ARE dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
              dep10 dep11 dep12 dep13 dep14 dep15 dep16
              dep17 dep18 dep19 dep20;
        USEVARIABLES ARE gender dep1 dep2 dep3 dep4 dep5
              dep6 dep7 dep8 dep9
              dep10 dep11 dep12 dep13 dep14 dep15 dep16
              dep17 dep18 dep19 dep20;
         Grouping is gender (1=boys 2=girls);

ANALYSIS:  Type = mgroup;
              Estimator = WLS;
Model:
dep1 with dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9 dep10 dep11 dep12 dep13
          dep14 dep15 dep16  dep17 dep18 dep19 dep20;
dep2 with  dep3 dep4 dep5 dep6 dep7 dep8 dep9 dep10 dep11 dep12 dep13 dep14
           dep15 dep16 dep17 dep18 dep19 dep20;
dep3 with  dep4 dep5 dep6 dep7 dep8 dep9 dep10 dep11 dep12 dep13 dep14 dep15
           dep16 dep17 dep18 dep19 dep20;
dep4 with  dep5 dep6 dep7 dep8 dep9 dep10 dep11 dep12 dep13 dep14 dep15 dep16
           dep17 dep18 dep19 dep20;
dep5 with dep6 dep7 dep8 dep9 dep10 dep11 dep12 dep13 dep14 dep15 dep16
          dep17 dep18 dep19 dep20;
dep6 with  dep7 dep8 dep9 dep10 dep11 dep12 dep13 dep14 dep15 dep16
           dep17 dep18 dep19 dep20;
dep7 with   dep8 dep9 dep10 dep11 dep12 dep13 dep14 dep15 dep16
            dep17 dep18 dep19 dep20;
dep8 with   dep9 dep10 dep11 dep12 dep13 dep14 dep15 dep16 dep17
```

```
                dep18 dep19 dep20;
dep9 with  dep10 dep11 dep12 dep13 dep14 dep15 dep16 dep17 dep18 dep19 dep20;
dep10 with  dep11 dep12 dep13 dep14 dep15 dep16 dep17 dep18 dep19 dep20;
dep11 with  dep12 dep13 dep14 dep15 dep16 dep17 dep18 dep19 dep20;
dep12 with  dep13 dep14 dep15 dep16 dep17 dep18 dep19 dep20;
dep13 with  dep14 dep15 dep16 dep17 dep18 dep19 dep20;
dep14 with  dep15 dep16 dep17 dep18 dep19 dep20;
dep15 with  dep16 dep17 dep18 dep19 dep20;
dep16 with  dep17 dep18 dep19 dep20;
dep17 with  dep18 dep19 dep20;
dep18 with  dep19 dep20;
dep19 with  dep20;
OUTPUT: stand;
```

Program 3.5 Configural invariance model (M1)

```
DATA:
 FILE IS bg123.DAT;
 FORMAT IS 1F2.0 22F1.0;
 VARIABLE: NAMES ARE sch gender wave
            dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
            dep10 dep11 dep12 dep13 dep14 dep15 dep16
            dep17 dep18 dep19 dep20;
        CATEGORICAL ARE dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
            dep10 dep11 dep12 dep13 dep14 dep15 dep16
            dep17 dep18 dep19 dep20;
        USEVARIABLES ARE gender dep1 dep2 dep3 dep4 dep5
            dep6 dep7 dep8 dep9 dep10 dep11 dep12 dep13 dep14
            dep15 dep16 dep17 dep18 dep19 dep20;
         Grouping is gender (1=boys 2=girls);


ANALYSIS:  Type = mgroup;
            Estimator = WLS;
Model: dep by dep1* dep2 dep3 dep4 dep5;
      dep by dep6 (1);
      dep by dep7 dep8 dep9 dep10 dep11 dep12
          dep13 dep14 dep15 dep16
          dep17 dep18 dep19 dep20;

     [DEP1$3] (2);
     [DEP2$3] (3);
     [DEP3$3] (4);
     [DEP4$3] (5);
     [DEP5$3] (6);
     [DEP6$3] (7);
     [DEP7$2] (8);
     [DEP8$3] (9);
     [DEP9$3] (10);
     [DEP10$3] (11);
```

```
    [DEP11$3] (12);
    [DEP12$3] (13);
    [DEP13$3] (14);
    [DEP14$3] (15);
    [DEP15$3] (16);
    [DEP16$3] (17);
    [DEP17$3] (18);
    [DEP18$3] (19);
    [DEP19$3] (20);
    [DEP20$3] (21);
    [DEP6$2]  (22);
    DEP@1;

Model girls:
    dep*;
    dep by dep1 dep2 dep3 dep4 dep5;
    dep by dep6 (1);
    dep by dep7 dep8 dep9 dep10 dep11 dep12
          dep13 dep14 dep15 dep16
          dep17 dep18 dep19 dep20;

    [DEP1$3] (2);
    [DEP2$3] (3);
    [DEP3$3] (4);
    [DEP4$3] (5);
    [DEP5$3] (6);
    [DEP6$3] (7);
    [DEP7$2] (8);
    [DEP8$3] (9);
    [DEP9$3] (10);
    [DEP10$3] (11);
    [DEP11$3] (12);
    [DEP12$3] (13);
    [DEP13$3] (14);
    [DEP14$3] (15);
    [DEP15$3] (16);
```

```
    [DEP16$3] (17);
    [DEP17$3] (18);
    [DEP18$3] (19);
    [DEP19$3] (20);
    [DEP20$3] (21);
    [DEP6$2]  (22);

  [DEP1$1* DEP1$2*  DEP2$1* DEP2$2*  DEP3$1* DEP3$2*
  DEP4$1* DEP4$2*  DEP5$1* DEP5$2*   DEP6$1*
  DEP7$1* DEP7$3*  DEP8$1* DEP8$2*   DEP9$1* DEP9$2*
  DEP10$1* DEP10$2* DEP11$1* DEP11$2* DEP12$1* DEP12$2*
  DEP13$1* DEP13$2* DEP14$1* DEP14$2* DEP15$1* DEP15$2*
  DEP16$1* DEP16$2*  DEP17$1* DEP17$2* DEP18$1* DEP18$2*
  DEP19$1* DEP19$2* DEP20$1* DEP20$2*];

OUTPUT: stand;
```

Program 3.6 Partial metric invariance model (M3)

```
DATA:
 FILE IS bg123.DAT;
 FORMAT IS 1F2.0 22F1.0;
 VARIABLE: NAMES ARE sch gender wave
            dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
            dep10 dep11 dep12 dep13 dep14 dep15 dep16
            dep17 dep18 dep19 dep20;
       CATEGORICAL ARE dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
            dep10 dep11 dep12 dep13 dep14 dep15 dep16
            dep17 dep18 dep19 dep20;
       USEVARIABLES ARE gender dep1 dep2 dep3 dep4 dep5
            dep6 dep7 dep8 dep9
            dep10 dep11 dep12 dep13 dep14 dep15 dep16
            dep17 dep18 dep19 dep20;
        Grouping is gender (1=boys 2=girls);


ANALYSIS:  Type = mgroup;
            Estimator = WLS;
Model: dep by dep1* dep2 dep3 dep4 dep5;
      dep by dep6 (1);
      dep by dep7 dep8 dep9 dep10 dep11 dep12
          dep13 dep14 dep15 dep16
          dep17 dep18 dep19 dep20;

     [DEP1$3] (2);
     [DEP2$3] (3);
     [DEP3$3] (4);
     [DEP4$3] (5);
     [DEP5$3] (6);
     [DEP6$3] (7);
     [DEP7$2] (8);
     [DEP8$3] (9);
     [DEP9$3] (10);
```

```
    [DEP10$3] (11);
    [DEP11$3] (12);
    [DEP12$3] (13);
    [DEP13$3] (14);
    [DEP14$3] (15);
    [DEP15$3] (16);
    [DEP16$3] (17);
    [DEP17$3] (18);
    [DEP18$3] (19);
    [DEP19$3] (20);
    [DEP20$3] (21);
    [DEP6$2]  (22);
    DEP@1;

Model girls:
   dep*;
   dep by dep6 (1);
   dep by dep1 dep2 dep15 dep17;
    [DEP1$3] (2);
    [DEP2$3] (3);
    [DEP3$3] (4);
    [DEP4$3] (5);
    [DEP5$3] (6);
    [DEP6$3] (7);
    [DEP7$2] (8);
    [DEP8$3] (9);
    [DEP9$3] (10);
    [DEP10$3] (11);
    [DEP11$3] (12);
    [DEP12$3] (13);
    [DEP13$3] (14);
    [DEP14$3] (15);
    [DEP15$3] (16);
    [DEP16$3] (17);
    [DEP17$3] (18);
    [DEP18$3] (19);
```

```
      [DEP19$3] (20);
      [DEP20$3] (21);
      [DEP6$2]  (22);

  [DEP1$1* DEP1$2*  DEP2$1* DEP2$2*  DEP3$1* DEP3$2*
  DEP4$1* DEP4$2*  DEP5$1* DEP5$2*   DEP6$1*
  DEP7$1* DEP7$3*  DEP8$1* DEP8$2*   DEP9$1* DEP9$2*
  DEP10$1* DEP10$2* DEP11$1* DEP11$2* DEP12$1* DEP12$2*
  DEP13$1* DEP13$2* DEP14$1* DEP14$2* DEP15$1* DEP15$2*
  DEP16$1* DEP16$2*  DEP17$1* DEP17$2* DEP18$1* DEP18$2*
  DEP19$1* DEP19$2* DEP20$1* DEP20$2*];
OUTPUT: stand;
```

Program 3.7 Final scalar invariance model (M5)

```
DATA:
 FILE IS bg123.DAT;
 FORMAT IS 1F2.0 22F1.0;
 VARIABLE: NAMES ARE sch gender wave
                dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
                dep10 dep11 dep12 dep13 dep14 dep15 dep16
                dep17 dep18 dep19 dep20;
        CATEGORICAL ARE dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
                dep10 dep11 dep12 dep13 dep14 dep15 dep16
                dep17 dep18 dep19 dep20;
        USEVARIABLES ARE gender dep1 dep2 dep3 dep4 dep5
                dep6 dep7 dep8 dep9 dep10 dep11 dep12 dep13 dep14 dep15
                dep16 dep17 dep18 dep19 dep20;
         Grouping is gender (1=boys 2=girls);


ANALYSIS:  Type = mgroup;
                Estimator = WLS;
 Model: dep by dep1* dep2 dep3 dep4 dep5;
      dep by dep6 (1);
      dep by dep7 dep8 dep9 dep10 dep11 dep12
           dep13 dep14 dep15 dep16
           dep17 dep18 dep19 dep20;

     [DEP1$3] (2);
     [DEP2$3] (3);
     [DEP3$3] (4);
     [DEP4$3] (5);
     [DEP5$3] (6);
     [DEP6$3] (7);
     [DEP7$2] (8);
     [DEP8$3] (9);
     [DEP9$3] (10);
     [DEP10$3] (11);
```

```
    [DEP11$3] (12);
    [DEP12$3] (13);
    [DEP13$3] (14);
    [DEP14$3] (15);
    [DEP15$3] (16);
    [DEP16$3] (17);
    [DEP17$3] (18);
    [DEP18$3] (19);
    [DEP19$3] (20);
    [DEP20$3] (21);
    [DEP6$2]  (22);
    DEP@1;

Model girls:
    dep*;
    dep by dep6 (1);
    dep by dep1 dep2 dep15 dep17;
    [DEP1$3] (2);
    [DEP2$3] (3);
    [DEP3$3] (4);
    [DEP4$3] (5);
    [DEP5$3] (6);
    [DEP6$3] (7);
    [DEP7$2] (8);
    [DEP8$3] (9);
    [DEP9$3] (10);
    [DEP10$3] (11);
    [DEP11$3] (12);
    [DEP12$3] (13);
    [DEP13$3] (14);
    [DEP14$3] (15);
    [DEP15$3] (16);
    [DEP16$3] (17);
    [DEP17$3] (18);
    [DEP18$3] (19);
    [DEP19$3] (20);
```

```
    [DEP20$3] (21);
    [DEP6$2]  (22);

[DEP1$1* DEP1$2*   DEP2$1* DEP2$2*
 DEP15$1* DEP15$2* DEP17$1* DEP17$2*
 DEP3$1*  DEP4$1*   DEP7$1* DEP11$1*
 DEP12$1* DEP18$1* DEP7$3*];
OUTPUT: stand;
```

Program 3.8 Invariant uniquenesses (M6)

```
DATA:
 FILE IS bg123.DAT;
 FORMAT IS 1F2.0 22F1.0;
 VARIABLE: NAMES ARE sch gender wave
              dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
              dep10 dep11 dep12 dep13 dep14 dep15 dep16
              dep17 dep18 dep19 dep20;
        CATEGORICAL ARE dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
              dep10 dep11 dep12 dep13 dep14 dep15 dep16
              dep17 dep18 dep19 dep20;
        USEVARIABLES ARE gender dep1 dep2 dep3 dep4 dep5
              dep6 dep7 dep8 dep9
              dep10 dep11 dep12 dep13 dep14 dep15 dep16
              dep17 dep18 dep19 dep20;
         Grouping is gender (1=boys 2=girls);

ANALYSIS:  Type = mgroup;
              Estimator = WLS;
 Model: dep by dep1* dep2 dep3 dep4 dep5;
      dep by dep6 (1);
      dep by dep7 dep8 dep9 dep10 dep11 dep12
          dep13 dep14 dep15 dep16
          dep17 dep18 dep19 dep20;

     [DEP1$3] (2);
     [DEP2$3] (3);
     [DEP3$3] (4);
     [DEP4$3] (5);
     [DEP5$3] (6);
     [DEP6$3] (7);
     [DEP7$2] (8);
     [DEP8$3] (9);
     [DEP9$3] (10);
     [DEP10$3] (11);
```

```
    [DEP11$3] (12);
    [DEP12$3] (13);
    [DEP13$3] (14);
    [DEP14$3] (15);
    [DEP15$3] (16);
    [DEP16$3] (17);
    [DEP17$3] (18);
    [DEP18$3] (19);
    [DEP19$3] (20);
    [DEP20$3] (21);
    [DEP6$2]  (22);
    DEP@1;

Model girls:
   dep*;
   dep by dep6 (1);
   dep by dep1 dep2 dep15 dep17;
    [DEP1$3] (2);
    [DEP2$3] (3);
    [DEP3$3] (4);
    [DEP4$3] (5);
    [DEP5$3] (6);
    [DEP6$3] (7);
    [DEP7$2] (8);
    [DEP8$3] (9);
    [DEP9$3] (10);
    [DEP10$3] (11);
    [DEP11$3] (12);
    [DEP12$3] (13);
    [DEP13$3] (14);
    [DEP14$3] (15);
    [DEP15$3] (16);
    [DEP16$3] (17);
    [DEP17$3] (18);
    [DEP18$3] (19);
    [DEP19$3] (20);
```

```
    [DEP20$3] (21);
    [DEP6$2]  (22);

 [DEP1$1* DEP1$2*   DEP2$1* DEP2$2*
  DEP15$1* DEP15$2* DEP17$1* DEP17$2*
  DEP3$1*  DEP4$1*   DEP7$1* DEP11$1*
  DEP12$1* DEP18$1* DEP7$3*];

 dep@1;

{dep3@1 dep4@1 dep5@1 dep6@1
 dep7@1 dep8@1 dep9@1 dep10@1 dep11@1 dep12@1
 dep13@1 dep14@1 dep16@1 dep18@1
 dep19@1 dep20@1};

 OUTPUT: stand;
```

Program 3.9 Invariant factor variances (M7)

```
DATA:
 FILE IS bg123.DAT;
 FORMAT IS 1F2.0 22F1.0;
 VARIABLE: NAMES ARE sch gender wave
                dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
                dep10 dep11 dep12 dep13 dep14 dep15 dep16
                dep17 dep18 dep19 dep20;
        CATEGORICAL ARE dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
                dep10 dep11 dep12 dep13 dep14 dep15 dep16
                dep17 dep18 dep19 dep20;
        USEVARIABLES ARE gender dep1 dep2 dep3 dep4 dep5
                dep6 dep7 dep8 dep9 dep10 dep11 dep12 dep13 dep14 dep15
               dep16 dep17 dep18 dep19 dep20;
         Grouping is gender (1=boys 2=girls);


ANALYSIS:  Type = mgroup;
                Estimator = WLS;
Model: dep by dep1* dep2 dep3 dep4 dep5;
       dep by dep6 (1);
       dep by dep7 dep8 dep9 dep10 dep11 dep12
            dep13 dep14 dep15 dep16
            dep17 dep18 dep19 dep20;

      [DEP1$3] (2);
      [DEP2$3] (3);
      [DEP3$3] (4);
      [DEP4$3] (5);
      [DEP5$3] (6);
      [DEP6$3] (7);
      [DEP7$2] (8);
      [DEP8$3] (9);
      [DEP9$3] (10);
      [DEP10$3] (11);
```

```
    [DEP11$3] (12);
    [DEP12$3] (13);
    [DEP13$3] (14);
    [DEP14$3] (15);
    [DEP15$3] (16);
    [DEP16$3] (17);
    [DEP17$3] (18);
    [DEP18$3] (19);
    [DEP19$3] (20);
    [DEP20$3] (21);
    [DEP6$2]  (22);
    DEP@1;

Model girls:
    dep*;
    dep by dep6 (1);
    dep by dep1 dep2 dep15 dep17;
    [DEP1$3] (2);
    [DEP2$3] (3);
    [DEP3$3] (4);
    [DEP4$3] (5);
    [DEP5$3] (6);
    [DEP6$3] (7);
    [DEP7$2] (8);
    [DEP8$3] (9);
    [DEP9$3] (10);
    [DEP10$3] (11);
    [DEP11$3] (12);
    [DEP12$3] (13);
    [DEP13$3] (14);
    [DEP14$3] (15);
    [DEP15$3] (16);
    [DEP16$3] (17);
    [DEP17$3] (18);
    [DEP18$3] (19);
    [DEP19$3] (20);
```

```
      [DEP20$3] (21);
      [DEP6$2]  (22);
     DEP@1;
[DEP1$1* DEP1$2*   DEP2$1* DEP2$2*
 DEP15$1* DEP15$2* DEP17$1* DEP17$2*
 DEP3$1*  DEP4$1*   DEP7$1* DEP11$1*
 DEP12$1* DEP18$1* DEP7$3*];
OUTPUT: stand;
```

Program 3.10 Invariant factor means (M8)

```
DATA:
 FILE IS bg123.DAT;
 FORMAT IS 1F2.0 22F1.0;
 VARIABLE: NAMES ARE sch gender wave
            dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
            dep10 dep11 dep12 dep13 dep14 dep15 dep16
            dep17 dep18 dep19 dep20;
       CATEGORICAL ARE dep1 dep2 dep3 dep4 dep5 dep6 dep7 dep8 dep9
            dep10 dep11 dep12 dep13 dep14 dep15 dep16
            dep17 dep18 dep19 dep20;
       USEVARIABLES ARE gender dep1 dep2 dep3 dep4 dep5
            dep6 dep7 dep8 dep9 dep10 dep11 dep12 dep13 dep14 dep15
            dep16 dep17 dep18 dep19 dep20;
        Grouping is gender (1=boys 2=girls);


ANALYSIS:  Type = mgroup;
            Estimator = WLS;
 Model: dep by dep1* dep2 dep3 dep4 dep5;
      dep by dep6 (1);
      dep by dep7 dep8 dep9 dep10 dep11 dep12
          dep13 dep14 dep15 dep16
          dep17 dep18 dep19 dep20;

     [DEP1$3] (2);
     [DEP2$3] (3);
     [DEP3$3] (4);
     [DEP4$3] (5);
     [DEP5$3] (6);
     [DEP6$3] (7);
     [DEP7$2] (8);
     [DEP8$3] (9);
     [DEP9$3] (10);
     [DEP10$3] (11);
```

```
    [DEP11$3] (12);
    [DEP12$3] (13);
    [DEP13$3] (14);
    [DEP14$3] (15);
    [DEP15$3] (16);
    [DEP16$3] (17);
    [DEP17$3] (18);
    [DEP18$3] (19);
    [DEP19$3] (20);
    [DEP20$3] (21);
    [DEP6$2]  (22);
    DEP@1;

Model girls:
   dep*;
   dep by dep6 (1);
   dep by dep1 dep2 dep15 dep17;
    [DEP1$3] (2);
    [DEP2$3] (3);
    [DEP3$3] (4);
    [DEP4$3] (5);
    [DEP5$3] (6);
    [DEP6$3] (7);
    [DEP7$2] (8);
    [DEP8$3] (9);
    [DEP9$3] (10);
    [DEP10$3] (11);
    [DEP11$3] (12);
    [DEP12$3] (13);
    [DEP13$3] (14);
    [DEP14$3] (15);
    [DEP15$3] (16);
    [DEP16$3] (17);
    [DEP17$3] (18);
    [DEP18$3] (19);
    [DEP19$3] (20);
```

```
      [DEP20$3] (21);
      [DEP6$2]  (22);
  [DEP1$1* DEP1$2*   DEP2$1* DEP2$2*
   DEP15$1* DEP15$2* DEP17$1* DEP17$2*
   DEP3$1*  DEP4$1*   DEP7$1* DEP11$1*
   DEP12$1* DEP18$1* DEP7$3*];
[dep@0];
OUTPUT: stand;
```

Shannon Research Press
Adelaide, South Australia
ISBN: 1-920736-09-3