1351.0.55.011

Australian Bureau of Statistics

Research Paper

Testing the Reliability of a Measure of Aboriginal Children's Mental Health

An Analysis Based on the Western Australian Aboriginal Child Health Survey

New Issue

Research Paper

Testing the Reliability of a Measure of Aboriginal Children's Mental Health

An Analysis Based on the Western Australian Aboriginal Child Health Survey

Stephen R. Zubrick and David Lawrence

Telethon Institute for Child Health Research
Perth, Western Australia

John de Maio and Nicholas Biddle

Australian Bureau of Statistics
Canberra, ACT

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) THURS 2 MAR 2006

ABS Catalogue no. 1351.0.55.011
ISBN 0 642 48206 3

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr Nicholas Biddle, Analytical Services Branch on Canberra (02) 6252 5061 or email <nicholas.biddle@abs.gov.au>.

CONTENTS

TESTING THE RELIABILITY OF A MEASURE OF ABORIGINAL CHILDREN'S MENTAL HEALTH: AN ANALYSIS BASED ON THE WESTERN AUSTRALIAN ABORIGINAL CHILD HEALTH SURVEY 2000-01

Stephen R. Zubrick and David Lawrence

Telethon Institute for Child Health Research

and

John de Maio and Nicholas Biddle

Australian Bureau of Statistics

ABSTRACT

This report details an analysis of the Western Australian Aboriginal Child Health Survey as they pertain to the measurement of mental health in Aboriginal children under the age of 18 years. In the paper we do not focus on the mental health outcomes of Aboriginal children in Western Australia (those interested in such issues should read the survey publication Zubrick, et al. 2004). Rather, we focus on testing the validity of applying the Strengths and Difficulties Questionnaire (SDQ) to this population.

We begin with a consideration of contextual issues that govern the measurement of health and mental health in particular. We then describe the basic data collection methods and present descriptions of the mental health variables that comprise the measures. The principal findings of the paper follow including a set of analyses of the psychometric characteristics of the measures based on structural equation modelling and multi-level modelling of carer and community clustering. We finish by summarising the limitations of the findings and provide concluding comments.

1. INTRODUCTION

Health status is a difficult concept to reliably and validly measure using sample surveys of households. It is usually not feasible to have a person with medical training present during an interview and therefore the presence of specific conditions and overall health must be based on the respondent's recollection and self-report. This can lead to under- and over-reporting or misreporting of certain conditions because that individual: (a) has not been or can not recall being tested for that condition; (b) can not recall the exact diagnosis given by a medical practitioner; or (c) if they had the condition at a certain point in time they have not been tested to see whether it is still present.

Reliability and validity of reporting is further threatened when:

- The conditions in question fall under the heading of mental health;
- The individual in question is a child or for some other reason needs to have someone else answer the survey questions on their behalf; or
- The questions and method of gathering the data are not reliable and/or valid given the language and cultural circumstances of the respondent.

All of the above conditions are likely to impact on the validity of data gathered from Aboriginal Australians. The analyses presented here assess the reliability and internal consistency of the Strengths and Difficulties Questionnaire (SDQ) (Goodman, 2001) – a measure of mental health commonly used for assessing children. SDQ data were gathered on children and young people from their carers, from their classroom school teachers, and from young people themselves where they were aged 12–17 years.

The data describing the mental health collected from carers of children aged 4 through 17 years are the subject of this report. The use of carer reported SDQ ratings should be kept in mind when interpreting the analysis and results in the remainder of this paper.

This paper is broadly intended for scientists, practitioners and policy-makers in the fields of health, family community services and education. While many of the tables and figures in this presentation include statistical summaries, the intent of the paper is not to describe the overall mental health of Aboriginal children in Western Australia. Rather the aim is to assess how well, in terms of validity and scale reliability, the SDQ can be used as an estimate of the social and emotional well-being of Aboriginal Australian children and young people living in a diverse set of circumstances. Readers interested in an analysis and discussion of the health of Aboriginal children in Western Australia should see Zubrick, et al. (2004).

1.1 The problem

The Strengths and Difficulties Questionnaire (SDQ) is a one-page questionnaire for assessing the psychological adjustment of children and youth (see Appendix A). Goodman (2001) notes that identical or nearly identical versions can be completed by parents or teachers of 3–16 year old children; or by young people aged 11–16 years themselves.

The SDQ comprises twenty-five questions, some positively worded, others negatively worded. Respondents use a 3-point Likert scale to indicate how far each attribute applies to the target child. Goodman reports that the twenty five items represent five underlying subscales of five items each. These comprise the

- *Emotional symptoms* scale;
- *Conduct problems* scale;
- *Hyperactivity* scale;
- *Peer problems* scale; and
- *Prosocial skills* scale.

The SDQ questions have been well tested and are known to be valid for the general population (see, Goodman, 1997 and Goodman, et al., 1998). However, this is the first time the SDQ has been administered to an Australian Aboriginal population and the first large scale attempt to measure mental health of Aboriginal children and young people in a diverse range of circumstances and settings. We focus on two main themes: (1) the internal reliability and consistency of the SDQ scale and subscales and (2) multi-level factors that potentially effect estimates of reliability and validity owing to the nature of the sample design and collection methods.

The reliability and consistency of the SDQ scale

In addressing internal reliability and consistency we pose the following questions:

- How well do the items on the SDQ measure 'global' social and emotional well-being?
- How well do each of the five observed items measure the theoretical underlying ('unobserved') factors of *Emotional symptoms*, *Conduct problems*, *Hyperactivity*, *Peer problems* and *Prosocial skills*?
- Are the items comprising the five selected subscales of the SDQ measuring the same facets of mental health in the same way for both boys and girls, young and old children, and those living in remote and less remote settings?
- Are there differences in carer-rated SDQ scores by Birth/non-Birth mother respondents or by Aboriginal/non-Aboriginal carers?

Multi-level effects in SDQ reports: Modelling of mental health outcomes

The WAACHS has many features of modern complex survey designs with stratification, multiple stages of selection and unequal selection probabilities. The data are clustered within families, where data on all children eligible to participate may have been collected from the same carer. The sample was also selected initially on the basis of census collection districts. Thus, hierarchical clustering of responses within families and within census collection districts occurs. This gives rise to variations in the information on an individual child that are instead attributable to the family or Census Collection District. Multi-level models are frequently used in the human and biological sciences for the modelling of hierarchically clustered populations.

In a practical sense children living in the same family are likely to be more similar than children selected using simple random sampling. Similarly, children living in the same small area may tend to be more similar if local environment plays a role in determining their mental health status. Multi-level models allow the determination of the proportion of variation in child mental health that is attributable to family level and small area level effects. These models can be extended to explore variation in mental health explained by other characteristics such as age, sex, gender, remoteness or physical health problems.

1.2 Scope of the survey and terminology used

While such an analysis of mental health would be useful for other population subgroups (for example Torres Strait Islanders) or geographic areas (other states or territories), unfortunately such children are beyond the scope of the survey. This creates a problem for the terminology used in the paper, especially when attempting to provide some context for the research.

Most ABS research on Aboriginal Australians is at a national level and hence provides information on both Aboriginal and Torres Strait Islander Australians. When citing such work in this paper we therefore refer to outcomes for Indigenous Australians. However, as the survey analysed was based in Western Australia and we are unable to make definitive statements on Torres Strait Islanders, when reporting results from the survey we refer only to Aboriginal Australians.

2. CONTEXT

Quantifying the health outcomes of Indigenous Australians is an important though difficult exercise. Although policy makers and researchers within the Indigenous and wider community need to know what health problems affect Indigenous Australians, identification issues (i.e. in the sense of Indigenous status) and differing conceptualisations of health make the measurement of Indigenous health challenging.

Despite these difficulties, there is a wide range of quantitative and qualitative evidence that suggests that Aboriginal Australians and other Indigenous groups suffer disproportionately from a number of health conditions.

2.1 Physical health

Mortality and morbidity rates for Indigenous Australians are substantially higher than those for the non-Indigenous population. ABS and AIHW (2005, page 148) report that life expectancy at birth for Indigenous Australians was roughly 17 years less than for the non-Indigenous Australians. Prevalence rates of certain diseases are also higher. These include, but are not limited to diabetes, heart disease, kidney disease and ear and hearing problems (ABS and AIHW, 2005, page 96).

These relatively high rates of mortality and morbidity are also present for the young Indigenous population. For example, For the period 1999–2003, Indigenous infant deaths (under one year) represented 6.2% of total Indigenous male deaths and 6.5% of total Indigenous female deaths. This is compared with 0.9% and 0.8% of the total for the respective non-Indigenous populations (ABS and AIHW 2005 page 149). [1] Furthermore Indigenous children and youths suffer disproportionately from a number of conditions including dental decay, skin sores and middle ear infections (ABS and AIHW 2005).

There are many possible reasons as to why the health of Indigenous Australians might be worse than for the general population, the most likely being their poor overall socioeconomic status. For a discussion of other possible reasons, see Gray, Hunter and Taylor (2002).

2.2 Mental health

In addition to the problems of identification common for all analysis of Indigenous outcomes (ABS and AIHW, 2005), there are extra difficulties present when measuring and comparing mental health. Some of these are due to the difficulties inherent in measuring mental health *per se*, while others are specific to the Indigenous population. Despite these difficulties, there is strong evidence that mental health

1 These figures refer to the Queensland, South Australia, Western Australia and the Northern Territory only.

problems in the Indigenous population are at least comparable, if not substantially worse than non-Indigenous Australians.

For example, ABS and AIHW (2005, page 131) reported that “There were more hospitalisations of Indigenous Australians than other Australians for most types of mental and behavioural disorders”. Furthermore, although there are many aspects of mental health that vary in their effect, suicide is a severe ‘down-stream’ indicator of mental health problems. According to ABS and AIHW (2005, pages 159-160), suicide rates are generally higher for Indigenous Australians than non-Indigenous Australians, particularly amongst the young. For example the suicide rate for those aged 0-24 years was three times higher for males and five times higher for females. [2]

Mental health is, however, a more encompassing concept than the presence or absence of specific diseases. According to the World Health Organisation, “Mental Health is not simply the absence of mental disorder or illness, but also includes a positive state of mental well-being” (World Health Organisation, 2004). However, because of the difficulty in having a cross-culturally relevant question(s) that is suitable to both Aboriginal and non-Aboriginal Australians, complete mental health information has not been available in the National Health Surveys (Australian Bureau of Statistics, 2002).

Encouragingly, recent progress has been made in establishing a comprehensive measure of the mental health of adult Indigenous Australians. Following extensive consultation and testing with a range of stakeholders, measures of social and emotional wellbeing (including *Kessler-10* items assessing psychological distress) have been included in the 2004-05 National Aboriginal and Torres Strait Islander Health Survey. However, measures of the spectrum of mental health distress in Aboriginal children are still lacking.

2 These figures refer to the Queensland, South Australia, Western Australia and the Northern Territory only.

3. DATA

3.1 The Western Australian Aboriginal Child Health Survey

The Western Australian Aboriginal Child Health Survey, a large scale epidemiological survey of the health and well-being of 5,289 Western Australian Aboriginal and Torres Strait Islander children, was undertaken by the Telethon Institute for Child Health Research (ICHR) in 2000–2001. The Survey's primary objective was to identify the developmental and environmental factors that enable competency and resiliency in Aboriginal children and young people. With this in mind the survey was designed to build an epidemiological knowledge base from which preventive strategies can be developed to promote and maintain healthy development and the social, emotional, academic and vocational well-being of young people. This is the first undertaking to gather comprehensive health, psycho-social and educational information on a population-based random sample of Aboriginal and Torres Strait Islander children in their families and in their communities.

Western Australia comprises over one third of the continental landmass of Australia. Families with Aboriginal children live in an enormously diverse range of communities distributed across the state. Some of these communities are small and discrete and are located in remote and isolated areas and may have associated 'out stations'. Other communities may be within towns or on the outskirts or fringes of towns, while still others are part of rural centres or urban areas. Some of these communities, particularly those that are isolated from larger population centres, have predominantly Aboriginal residents. City areas on the other hand have Aboriginal populations scattered more widely across urban areas. The north-west and centre of the state includes large tracts of desert and some of the most remote and sparsely populated areas in the world. The more populated south-west of the state includes extensive agricultural and forested areas with numerous small population centres.

Over two-thirds of the State's total population and one-third of the Aboriginal and Torres Strait Islander population resides in the metropolitan area of Perth. At the time of the Survey, the preliminary resident population of Western Australia was approximately 1.9 million people of which some 66,000 were estimated to be of Aboriginal or Torres Strait Islander descent.

The main survey commenced in May 2000 and was completed in June 2002. Families were eligible to be in the survey if they reported that there were "Aboriginal or Torres Strait Islander children or teenagers living at this address who are aged between 0 and 18 years". Dwellings were selected for screening using an area-based clustered multi-stage sample design. Interviewers enumerated 166,290 dwellings in 761 Census Collection Districts and randomly approached about 139,000 of these to determine if residents were eligible to participate in the survey. Using this method a random sample of 2,386 families with 6,209 eligible children was identified throughout metropolitan, rural and remote regions of Western Australia. A total of 1,999 of these families (84%) with 5,513 eligible children consented to participate in the survey. Interviewers gathered useable data on 5,289 (96%) of these participating children.

These 5,289 Aboriginal children with useable data are split between 1,296 children aged 0–3 years ('little kids') and 3,993 children aged 4–18 years ('big kids'). As SDQ information is only collected for those aged 4–18 years, our analysis is based primarily on data contained in the 'big kids' form.

Consent was also obtained from carers allowing young people aged 12–17 years to complete a separate questionnaire (Youth Self Report). This resulted in 1073 (73%) young people participating as independent respondents, without carer input. As well as covering much of the same health and well-being issues, this questionnaire also addressed several issues specific to youth, including school, peer groups, sex and drugs, leisure activities, family functioning, racism and mental health.

3.2 The sample

Table 3.1 shows the sample sizes by age and gender. It shows that there are relatively more males in the youngest age groups, but fewer males than females among older youths.

3.1 Sample sizes by age and gender

	4–11 years	*12–17 years*	*Total*
Male	1,324	690	2,014
Female	1,270	709	1,979
Total	2,594	1,399	3,993

At a later stage in this paper, we analyse and compare different population sub-groups. These comparisons include the Aboriginal status of the carer, their biological relationship to the child (i.e. birth mother versus other carer), and level of relative isolation – a measure of remoteness to urban and other service centres. Table 3.2 below gives the proportion of respondents in the sample for these characteristics by age and sex.

3.2 Age and sex of survey children

	4–11 years	*12–17 years*
Male		
Carer is Birth mother	80.1%	71.8%
Carer identifies as Aboriginal	86.4%	87.7%
Lives ex-metro regional, remote or very remote	37.8%	38.1%
Female		
Carer is Birth mother	83.1%	70.6%
Carer identifies as Aboriginal	87.1%	88.2%
Lives ex-metro regional, remote or very remote	37.8%	37.4%

Table 3.2 shows that the majority of children in the sample live with their Birth mother, although this proportion decreases as the child becomes older. The majority of children in the sample have Aboriginal carers and live in major cities or inner regional areas.

3.3 The Strengths and Difficulties Questionnaire (SDQ)

The SDQ is principally designed to be self-enumerated via paper and pencil. At the outset this was acknowledged to be of little value where respondents would have varying levels of literacy and access to English. Of necessity, the SDQ as used in the Survey required face-to-face administration with responses recorded by the interviewer.

Permission was obtained from Professor Robert Goodman (Goodman, 2000, personal correspondence) to assess the SDQ for its appropriateness of use in Australian Aboriginal populations. The SDQ was subsequently used in the pilot phases of the Survey. Field reports and data quality indicated that respondents generally felt the questionnaire items to be meaningful and relevant, to cover an appropriate range of both 'good' and 'bad' behaviours, but that the response categories as designed for use in mainstream, predominantly Western cultures, were ambiguous. As a result of piloting, the original response categories of the SDQ were altered. Table 3.3 summarises these changes.

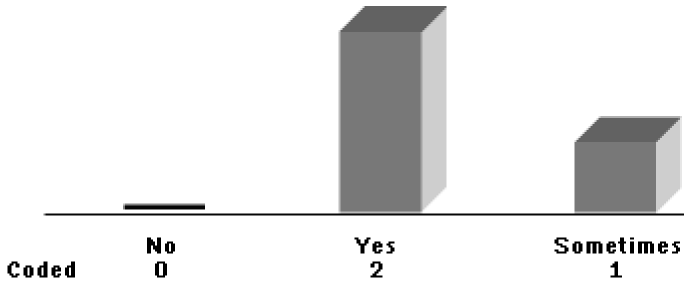
3.3 Changes to SDQ response categories for use in the WA Aboriginal Child Health Survey

Original response categories	*Response categories used in the WAACHS*	*Numerical coding*
Not true	No	0
Somewhat true	Sometimes	1
Certainly true	Yes	2

It is critical to note that the presentation order of the probe item and response categories conformed to the following procedure:

> **Instructions to the respondent**: "The next questions are about (child's name) behaviour and how (he/she) gets along with other people. Thinking about (child's name) behaviour over the past six months, that is, since (calendar event such as Christmas, a particular event in the community, etc.), (Item 1): has (he/she) been considerate of other people's feelings – No, Yes or Sometimes?"

3.4 Prompt card format for the Strengths and Difficulties Questionnaire



A single prompt card with visual prompts showing the relative 'size' (small, large and medium) of the response categories along with their labels (No, Yes and Sometimes) was provided during the administration of the 25 items (figure 3.4). This provided a relatively natural format of presenting the items and probing for response categories that corresponded with the way the pilot respondents reported thinking about their response. Pilot testing indicated that respondents felt that the answers to the probes were either "No" or "Yes" (these had the greatest salience in terms of judgement) and that "Sometimes" was another option. Notions such as "certainly" or "somewhat true" made little sense to respondents. In their views the answers were either "No" or "Yes" or "Sometimes".

Using this coding scheme the SDQ subscales and total scores were then calculated as per Goodman's directions (downloaded from www.sdq.info.com). Full details on how the SDQ is scored can be found in Appendix A.

The 25 items in the SDQ comprise five scales of five items each as shown in table 3.5. Each of the subscale scores can range from zero (no difficulties with any of the five items) to ten (maximum difficulties with all five items). As specified by Goodman, the total SDQ score is based upon the sum of the items on all scales except the *Prosocial skills* scale. Thus the total SDQ score, based on 20 items, can range from 0 to 40.

Univariate distributions for these variables are presented in Appendix B, along with the frequency distributions of the total SDQ score by age group and gender.

It should be noted that the items are scored on a coarse ordinal scale (0, 1 and 2) and that individual items are almost uniformly non-normal in their distributions. Similarly, the SDQ total score is manifestly positively skewed. These characteristics pose substantial challenges in selecting statistical methods for their analyses.

3.5 SDQ items and variable names used in later modelling*

SDQ Subscale	*Variable name*
Emotional symptoms scale	
Often complains of headaches, stomach aches or sickness	SOMATIC
Often seems worried	WORRIES
Often unhappy, sad or tearful	UNHAPPY
Nervous or clingy in new situations, easily lost confidence	CLINGY
Many fears, easily scared	AFRAID
Conduct problems scale	
Often has temper tantrums	TANTRUM
Usually done what adults told him/her to do	ROBEYS
Been in fights with other children or bullies them	FIGHTS
Often lies or cheats	LIES
Steals from home, school or elsewhere	STEALS
Hyperactivity scale	
Restless, overactive can not stay still for long	RESTLES
Constantly fidgeting or squirming	FIDGETY
Easily distracted, or poor concentration	DISTRAC
Able to stop and think things out before acting	RREFLECT
Good attention span and finished the things they start	RATTENDS
Peer problems scale	
Tends to play by themself	LONER
Has at least one good friend	RFRIEND
Generally liked by other children	RPOPULAR
Picked on or bullied by other children	BULLIED
Gets on better with adults than with other children	OLDBEST
Prosocial skills scale (not included in the Total Score)	
Considerate of other people's feelings	RCONSID
Readily shares with other children	RSHARES
Helpful if someone is hurt, upset or feeling ill	RCARING
Kind to younger children	RKIND
Often volunteers to help others	RHELPOUT

* Variables starting with 'R' have been reverse coded prior to data analyses. Thus higher scores signify behavioural or emotional difficulties.

4. INTERNAL RELIABILITY AND VALIDITY OF THE SDQ

One of the primary objectives of this paper is to assess the scale reliability (internal consistency) of the SDQ measurement model when it is applied in an Aboriginal context. That is to say, how well do each of the individual items measure the underlying latent variables (i.e. factors) that they are purported to measure and how well do the entire set of items measure, in a 'global' sense, mental health distress?

The principal statistical method used to address this question is confirmatory factor analysis (CFA). Initially, CFA is used to fit one-factor congeneric measurement models to the ordinal-scaled SDQ indicators. For each of the five SDQ measures, one-factor congeneric models were specified and fitted to the data to enable an assessment of how well each of the five observed indicators measure the unobservable latent variables (i.e. factors) underlying each subscale. We then assess reliability of the SDQ subscales based on various model goodness-of-fit statistics.

We start by building small congeneric models of each of the SDQ subscales and estimate these models as if the data were collected from a simple random sample. This ignores the clustering attributable to the carer in those situations where there was more than one child per carer. Subsequent models test for multi-level effects attributable to clustering – that is variance that is attributable to the fact that one carer may report on more than one child.

4.1 The single-level one factor congeneric model

The one-factor congeneric measurement model is described below (Joreskog and Sorbom, 1989, pp. 76–78).

$$X_i = \lambda_i \xi_1 + \delta_i \tag{1}$$

The five observed indicators contained in the *Emotional symptoms* subscale are used to illustrate the model. In this case:

X_i – observed variables (e.g. SOMATIC, WORRIES, UNHAPPY, CLINGY, AFRAID)

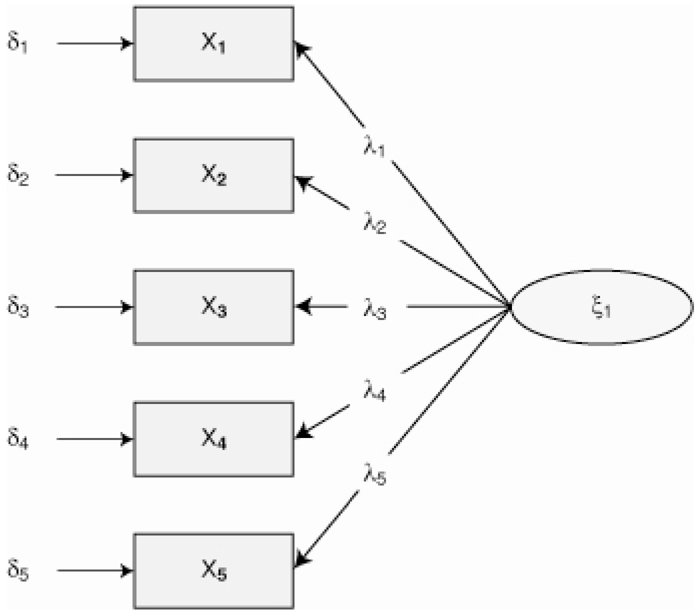
ξ_1 – unobserved latent variable (e.g. a factor called EMOTION)

δ_i – measurement errors in X_i

λ_i – regression coefficients in the relationships between each of the observed variables (X_i) and the unobserved ξ_1 (EMOTION).

The model described above can also be illustrated with a path diagram (see figure 4.1). The path diagram is a useful way to graphically display the pattern of relationships among sets of observable and unobservable variables (Dillon & Goldstein, 1984, page 433).

4.1 Graphical representation of the path diagram



4.2 Estimating structural equation models with ordinal data

Most researchers in applied statistics think in terms of modelling *individual* observations. In multiple regression analysis or ANOVA (Analysis of Variance), regression coefficients or the error variance estimates are derived from the minimisation of the sum of the squared differences between the predicted and observed dependent variable for each observation (Bollen, 1989).

In contrast to this approach, structural equation modelling emphasises *covariances* rather than cases. Rather than minimising functions of observed and predicted *individual* values, structural equation modelling minimises the difference between the sample covariances (i.e. the observed covariances) and the covariances predicted by the model. The observed covariances minus the predicted covariances form the residuals. Thus, researchers specify a model that they believe explain the observed data, and the data are fitted to this model and then statistically assessed for goodness-of-fit. There are several software programs that produce these analyses, one of which, Linear Structural Relations (LISREL) (Joreskog and Sorbom, 1996), is used here.

A critical feature of Structural Equation Modelling is that it most commonly assumes interval scale measures from which covariances and Pearson product moment correlations may be derived. In contrast the SDQ data are coarsely ordinal and markedly non-normal. The analysis of non-normally distributed and/or ordinal level data is much more problematic and the subject of considerable statistical debate.

Joreskog and Sorbom (1989, 1996) note that when some or all of the variables to be analysed are discrete or ordinal variables (as they are with the SDQ) then it is a misuse of LISREL methodology to: (1) assume these scores have interval scale properties, (2) compute a covariance matrix or a product-moment matrix for such scores, and (3) analyse such matrices using Maximum Likelihood methods (Joreskog and Sorbom, 1989, page 192). Under these circumstances, Joreskog and Sorbom propose using polychoric or polyserial correlations to replace covariances or Pearson correlations, and to assess the fit of models using such data via weighted least squares (WLS) with an appropriate weight matrix.

Hayduk (1987) is more cautious in his enthusiasm for such an approach, noting that the replacement of product moment correlations may be most prudent where the categorization process of the items has produced oppositely skewed categoric distributions in the items that serve as indicators of the underlying concepts. West, Finch and Curran (1995) in their review of structural equation modelling with non-normal variables note that factor loadings and factor correlations are subject to under-estimation particularly where there are few categories (two or three), the distributions are skewed (e.g. > 1.0) and there is differential skew across the items (West, Finch and Curran, 1995, page 64). In a re-assessment of the analysis of ordinal data, Hayduk (1996) concluded that while the analysis of ordered categorical data with maximum likelihood (ML) methods has returned results "better than anticipated" (page 213), coarsely ordered categories require use of procedures other than ML for estimation.

On balance the distributions of the SDQ data from the WAACHS show the items to be coarsely ordinal, that is there are only three possible response categories for each of the 25 items and that the distributions are markedly non-normal being skewed or U-shaped, and in some instances showing low (< 5%) response categories that effectively become zero in some sub-samples (see Appendix B). As a result we have adopted a cautious approach to estimating the internal reliability and validity of the SDQ. Our initial estimations use polychoric correlations with a weight matrix derived from the inverse of the asymptotic covariances as input to weighted least squares estimation (WLS).

4.3 Model results and interpretation

A single-level, one-factor congeneric measurement model is fitted to each of the SDQ subscales. The model results were obtained under a weighted least squares method of estimation based on polychloric correlation matrices. The path diagrams for each of the five models are reproduced in Appendix C, while the table below summarises the factor loadings for each SDQ subscale.

4.2 Factor loadings – one factor congeneric models

Emotion	*λ*	*Conduct*	*λ*	*Hyper*	*λ*	*Peer*	*λ*	*Prosocial*	*λ*
UNHAPPY	0.77	LIES	0.75	RESTLES	0.79	RFRIEND*	0.69	RCARING*	0.75
WORRIES	0.73	STEALS	0.73	FIDGETY	0.79	RPOPULAR*	0.56	RSHARES*	0.65
CLINGY	0.61	FIGHTS	0.62	DISTRAC	0.68	LONER	0.45	RKIND*	0.65
AFRAID	0.59	TANTRUM	0.56	RREFLECT*	0.57	BULLIED	0.38	RCONSID*	0.58
SOMATIC	0.51	ROBEYS*	0.51	RATTENDS*	0.56	OLDBEST	0.32	RHELPOUT*	0.55

* reverse coded

4.3.1 Regression model estimates

The estimated regression coefficients (λ_i's) give the magnitude of the expected change in the observed variable for a one-unit change in the unobservable variable. For example, if we were able to observe the latent variable EMOTION, a one unit change in this variable results in a 0.51 unit change in the observed variable SOMATIC (often complains of headaches, stomach aches). The other regression estimates on each of the observed variables can be interpreted in the same way. In the above path diagrams, the arrows do *not* represent direct causal influences in the usual sense. Rather in the sense, that *if the latent variables were observed* they would produce values of the observed indicators indicated by the regression estimates (Joreskog and Sorbom, 1989, page 77).

The t-statistics indicate that all the paths from the observed indicators to each respective latent variable are statistically significant.

The five models were also examined for theoretically inconsistent estimates. As Hair, et al. (1998) note, the three most common checks for offending estimates are:

- negative error variances,
- standardized coefficients exceeding or very close to 1.0, or
- very large standard errors.

No cases of any of these inconsistencies were found.

These results are generally satisfactory. As table 4.2 above shows, the majority of the estimated factor loadings are between 0.5 and 0.8, which we believe are acceptable loadings (in terms of the relationship between the observed indicators and the underlying unobservable constructs). Entries in each of the columns have been ordered by their strength of association with the underlying latent variable (i.e. factor). For example, problems with conduct are best measured by 'lying', 'stealing' and 'fighting', while 'tantrums' and '(dis)obeying' are less reliable measures of *Conduct problems*.

In contrast to most of the scales in table 4.2, the *Peer problems* scale is less well measured and shows considerable variability in the strength of association between the items and the underlying latent variable. Three of the indicators in the *Peer problems* subscale have factor loadings less than 0.5 ("tends to play by themself", "picked on or bullied", and "getting on better with adults than with other children"). Some of this undoubtedly reflects the wide variation in ages of the children (4–16 years) and the developmental appropriateness of the items for those ages. We more formally test for reliability of each of the five subscales by examining goodness-of-fit statistics in Section 4.4.

4.3 Factor analysis: Factor loading comparisons (a)

Emotion	*λ*	*Conduct*	*λ*	*Hyper*	*λ*	*Peer*	*λ*	*Prosocial*	*λ*
UNHAPPY	0.83	LIES	0.75	RESTLES	0.79	RFRIEND*	0.65	RCARING*	0.65
	0.61		0.63		0.65		(b)		0.61
	0.60		0.64		0.66		0.64		0.67
WORRIES	0.73	STEALS	0.72	FIDGETY	0.82	RPOPULAR*	0.64	RSHARES*	0.65
	0.60		0.66		0.63		(b)		0.59
	0.69		0.52		0.65		0.61		0.53
CLINGY	0.65	FIGHTS	0.70	DISTRAC	0.75	LONER	0.45	RKIND*	0.68
	0.49		0.65		0.62		0.58		0.59
	0.66		0.61		0.77		0.56		0.56
AFRAID	0.63	TANTRUM	0.65	RREFLECT*	0.65	BULLIED	0.72	RCONSID*	0.66
	0.40		0.36		0.51		(b)		0.49
	0.71		0.54		0.64		0.47		0.58
SOMATIC	0.51	ROBEYS*	0.67	RATTENDS*	0.62	OLDBEST	0.26	RHELPOUT*	0.51
	0.47		(b)		0.50		0.67		0.56
	0.47		0.43		0.72		0.56		0.68

(a) Upper figures are based on the WAACHS data and fitted in LISREL via weighted least squares estimation (N=3993, chi-square=2080.2, p<0.01, df=265, RMSEA=0.041, AGFI=0.98, RMR=0.12). Middle figures are based on WAACHS data fitted to a five-factor principal components analysis with varimax rotation. Lower figures are based upon data reported by Goodman (2001) on a sample of 10,434 3–16 year old British children fitted to a five-factor principal components analysis with varimax rotation.

(b) Variable did not load on this factor.

4.3.2 An international comparison

We also wanted to know the extent to which the WAACHS parent-reported data fit Goodman's reported factor structure (Goodman, 2001). Goodman reported the results of an unspecified factor analysis on 9,998 British children analysed using SPSS (see table 4.3). Assuming that Goodman conducted a principal components analysis, (while there is no mention of which analytical technique Goodman employs, he does note that he applied a varimax rotation), we undertook a similar analysis with the WAACHS data.

Principal Component Analysis of the British data

Goodman's unspecified factor analysis extracted six factors with eigenvalues in excess of 1.0 accounting for 45.9% of the common factor variance. Diagnostic screening indicated excellent factorability (KMO 0.87) and with final communalities ranging from 0.32 (TANTRUM) to 0.58 (FRIEND). Goodman also noted that factor analysis of his parent report data produced six factors (Goodman, 2001, page 1339) but that the last factor had an eigenvalue of 1.02.

Principal Component Analysis of the WAACHS

To test the WAACHS data against Goodman's reported factor analysis, the data were fitted to a five-factor solution (using PCA) and interpreted with a varimax solution. [As in Goodman's analysis, the WAACHS had a sixth factor with a low eigenvalue (1.06)]. To replicate Goodman's analysis, we excluded this last factor from further analysis.

The five-factor solution accounted for 41.6% of the common factor variance with communalities that ranged from 0.27 (SOMATIC) to 0.50 (UNHAPPY). A total of 21 of the 25 variables loaded on factors corresponding to those reported by Goodman (table 4.3). The variables that did not load on their purported underlying factors included ROBEYS, RFRIEND, RPOPULAR and BULLIED. It is notable that the predominant lack of fit in the factor analysis occurred with the *Peer problems* factor. Generally though, there is reasonable similarity between the two factor analytic solutions with a generally similar pattern of factor loadings suggesting at least four factors of good comparability.

Structural equation modelling of the WAACHS

Table 4.3 also provides a direct test of Goodman's model using more appropriate analytic techniques. We used weighted least squares with polychoric correlations and an asymptotic covariance weight matrix to test the fit of the WAACHS data to Goodman's five-factor model. The results indicated an acceptable fit (N=3993, chi-square=2080.2, p<0.01, df=265, RMSEA=0.041, AGFI=0.98, RMR=0.12) with generally satisfactory factor loadings. The average loading across the 25 items was 0.65 and loadings ranged from 0.26 (OLDBEST) to 0.83 (UNHAPPY).

In general, the factor structure of the SDQ, when used with Western Australian Australian Aboriginal children, shows good similarity to the model proposed by Goodman (1991). Some variability is seen in the underlying factors for the *Peer problems* and *Prosocial skills* factors, but in the main the WAACHS data conform surprisingly well to the overall model.

4.4 Model goodness-of-fit

Just how well are each of these five behavioural constructs measured using data collected from the carers of Aboriginal children? We assess how well each model fits the data by employing various goodness-of-fit measures. Joreskog and Sorbom (1989, page 43) outline four measures which can be used to judge model fit. These are:

- Chi-square (χ^2)
- Goodness-of-fit Index (GFI)
- Adjusted Goodness-of-fit Index (AGFI)
- Root mean square residual (RMR)

Brown and Cudeck (1993) also propose the Root Mean Square Error of Approximation (RMSEA) measure as another means of assessing model fit.

Full details of each test can be found in Appendix D. Here, we present a summary of some minimum guidelines reported in the literature for acceptable model fit.

4.4 Goodness-of-fit statistics – Summary of minimum guidelines

Test	*Guideline*	*Reference*
Chi-square	Insignificant χ^2	Joreskog (1989)
GFI	GFI > 0.95	Fullarton, et al. (2003)
AGFI	AGFI > 0.80	Hair, et al. (1998)
RMR	RMR < 0.05	Hair, et al. (1998)
RMSEA	RMSEA < 0.10	Brown & Cudeck (1993)

4.5 Assessing the five SDQ subscales

The first step in assessing the reliability of each SDQ subscale is to assess the fit of the one-factor congeneric models using the five goodness-of-fit statistics described in Appendix D.

All five models are judged to be satisfactory based on the GFI and AGFI measures. At least 98% of the variation in each of the five unobservable constructs are explained by their respective set of five indicators.

The *Emotional symptoms*, *Hyperactivity* and *Peer problems* models have an RMR value above the recommended value of 0.05 suggested by Hair, et al. (1998). We note that the *Hyperactivity* model has an RMSEA estimate of 0.108 which is just at the upper bound of acceptability (Joreskog, 2001).

4.5.1 Scale reliability of each SDQ subscale

Our next step is to estimate a summary measure of reliability for each set of items underlying the SDQ subscales. This is done to assess whether the five specified indicators adequately represent each SDQ subscale.

Following Raykov (www.ssicentral.com [3]), scale (or construct) reliability is calculated as:

$$\rho_y = \frac{(\sum_{i=1}^{k} b_i)^2}{(\sum_{i=1}^{k} b_i)^2 + \sum_{i=1}^{k} \theta_{ii}}$$

where:

b_i = the construct loadings (i.e. the lambdas from Section 4.1), and

θ_{ii} = the indicator measurement error (i.e. the theta deltas from Section 4.1).

This coefficient is defined as the ratio of true variance in the indicators to its observed variance. With higher values indicating more 'precise' or 'consistent' measurement in the model. Hair, et al. (1998) recommend a level of at least 0.70 when assessing scale reliability using this measure. Readers familiar with Chronbach's alpha (α) should not confuse the measure used here with that of Chronbach's. There are several reasons that make the use of α unsuitable as a measure of internal reliability and readers are referred to Raykov (http://www.ssicentral.com/lisrel/mainlis.htm) for a discussion of this.

The scale reliability for each of the five SDQ subscales is reported in table 4.5.

4.5 SDQ scale reliability

SDQ subscale	*Subscale reliability*
Hyperactivity	0.813
Emotional symptoms	0.780
Conduct problems	0.774
Prosocial skills	0.774
Peer problems	0.604

Again the results here are generally satisfactory. Internal reliabilities are all relatively robust for the *Hyperactivity*, *Emotional symptoms*, *Conduct problems* and *Prosocial skills* subscales. When assessed against the recommended value of 0.70 the *Peer problems* subscale is the only one not to exceed 0.70, indicating that it performs more poorly in terms of its scale reliability.

3 http://www.ssicentral.com/lisrel/mainlis.htm

4.5.2 SDQ total scale and subscale reliability with respect to LORI

We further assessed scale reliability by calculating scale reliabilities for each of the five SDQ subscales by the classification of the Level of Relative Isolation (LORI) – a measure of geographic remoteness from population service centres.[4] As LORI levels increase, the level of relative isolation or remoteness increases (figure 4.7). Readers will find full details of this measure in Zubrick, et al. (2004). Scale reliabilities for each level of LORI are provided in table 4.6.

4.6 Scale reliabilities by Level of Relative Isolation (LORI)

Level of Relative Isolation	*Number of children*	*I Emotional symptoms*	*II Conduct problems*	*III Hyper-activity*	*IV Peer problems*	*V Prosocial skills*	*Four-factor (I, II, III, IV) total scale reliability*
None	1,214	0.709	0.651	0.752	0.428	0.626	0.952
Low	1,266	0.644	0.661	0.734	0.352	0.674	0.945
Moderate	715	0.662	0.641	0.649	NA*	0.593	0.944
High	416	0.606	0.722	0.631	0.469	0.548	0.952
Extreme	382	0.585	0.482	0.607	NA*	0.506	0.942
Total	3,993	0.780	0.774	0.813	0.604	0.774	0.935

NA* – the models underlying these calculations did not converge

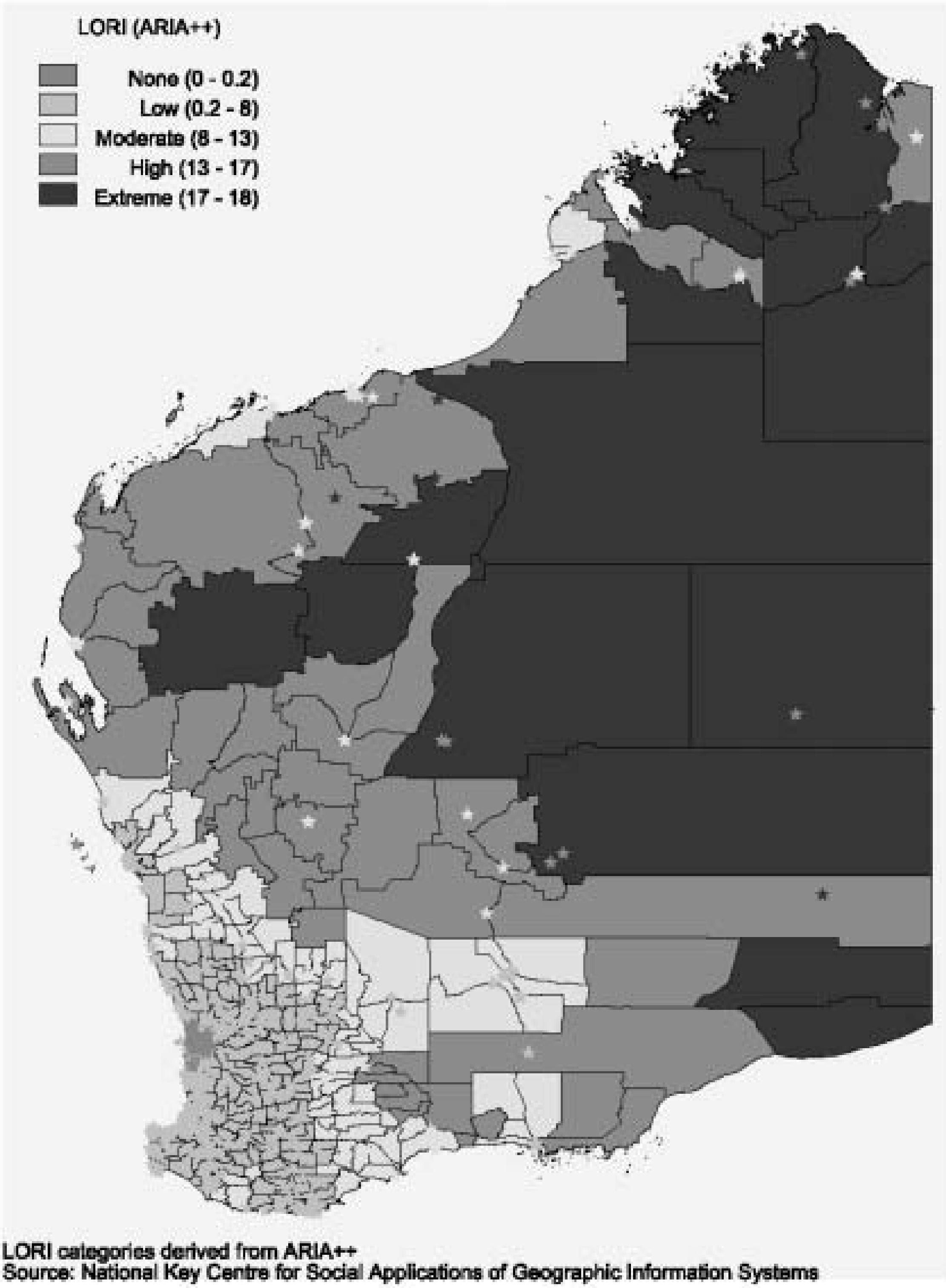
The overall individual scale reliabilities for the *Hyperactivity*, *Emotional symptoms*, *Conduct problems* and *Prosocial skills* subscales are relatively robust, ranging from 0.77 to 0.81. *Peer problems* has the lowest overall scale reliability when calculated across the entire sample reflecting underlying variability and non-convergence within some LORI levels (see below).

The total SDQ scale reliability is 0.935.

Within levels of relative isolation, data show that for each SDQ subscale, scale reliabilities decline as a child resides in a more remote locality. This possibly reflects differences in interview administration with a high proportion of respondents who spoke an Aboriginal language as a first language in areas of greater relative isolation, who required simultaneous translation during interview, and for whom some concepts were less salient to cultural and living circumstances. As with the overall scale reliabilities, the peer subscale performed poorly. Scale reliabilities could not be calculated for the peer subscale in the LORI categories of moderate and extreme as the models underlying these calculations did not converge.

4 Scale reliabilities by LORI status are calculated by running confirmatory factor models for each LORI category and SDQ subscale. For example, the scale reliability for the *Emotional symptoms* subscale for children living in LORI category 1 (0.709) is calculated using the factor loadings and measurement errors estimated under a one-factor congeneric model. This model is based on the polychloric correlation matrix generated from 1,214 children living in LORI category 1 and estimated via Weighted Least Squares.

4.7 WA Census Collection Districts – Level of Relative Isolation (LORI) categories



4.6 Summary: SDQ total scale and subscale analyses

The scale reliability of the five SDQ subscales have been assessed with reference to

- examination of their estimated factor loadings;
- various model fit statistics; and
- scale reliability as measured by Raykov.

Using 20 of the 25 items suggested by Goodman, the overall scale reliability of the SDQ across the sample and levels of relative isolation is on the order of 0.93. These total reliabilities based on 20 items are relatively stable at each level of relative isolation. However, at the subscale level there are noticeable variations in scale reliability. These variations are between each of the five underlying factors and between levels of relative isolation. Broadly speaking the *Emotional symptoms*, *Conduct problems* and *Hyperactivity* scales show relatively better scale reliability in the sense of magnitude, *Prosocial skills* somewhat less so, while the *Peer problems* subscale performs less well – particularly within levels of relative isolation.

On balance these findings suggest that the total SDQ score is likely to be an adequate measure of mental health distress and it is to this task we next turn.

5. SINGLE-LEVEL MULTIPLE FACTOR CONGENERIC MODELS

In the previous section we have examined each of the SDQ items and their relationship to the subscales that they are purported to measure. Results were generally satisfactory, particularly with respect to properties of the 20 item total scale. At the factor level, with the exception of *Peer problems* the other SDQ subscales appeared to be well measured. However, one of the purposes of using the SDQ is to derive a Total Score. To do this requires assessing how well the SDQ items fit this larger measurement model of mental health in Aboriginal children and young people.

The previous models described in Section 4 (one-factor congeneric models) generalises immediately to several sets of congeneric measures (see Joreskog and Sorbom, 1989). If the different latent variables $\xi_1, \xi_2, \ldots, \xi_n$ are all mutually uncorrelated, then each set of measures can be analysed separately as in the previous section. However in most cases, these latent variables correlate with each other and an overall analysis of the entire set of measures must be made.

We have no strong *a priori* hypothesis of the factor structure underlying the SDQ measurement model. For example, data are collected on five subscales but only four of these are used in the actual scoring of the SDQ. For this reason, in assessing Goodman's underlying model of strengths and difficulties on data collected from the carers of Aboriginal children we separately estimate three models:

- A five-factor congeneric model comprising five factors (*Emotional symptoms*, *Conduct problems*, *Hyperactivity*, *Peer problems* and *Prosocial skills* scales) and all 25 observed indicators.
- A four-factor congeneric model (*Emotional symptoms*, *Conduct problems*, *Hyperactivity* and *Peer problems*) with the 20 indicators used in recommended scoring the SDQ model.
- A preferred model with 16 indicators based on empirical results using the WAACHS data.

For each of these models, estimates are once again obtained under weighted least squares estimation, based on polychoric correlations (and asymptotic covariance matrices).

Some comment should be made on the reduced (16 item) model. Initial models were fitted via weighted least squares estimation on a 50% random sample from the WAACHS data using polychoric correlations and an appropriate weight matrix. After inspection of the standardised factor loadings (lambdas) those loadings that were above 0.59 were retained. Two exceptions were made to this. An additional item, SHARES, was retained on the *Prosocial skills* scale and BULLIED was retained over LONER on the *Peer problems* scale. In general the goal was to retain a set of items that strongly measured their underlying latent variables (in the sense of having lower proportions of error variance) with better properties for subsequent use in multi-level modelling.

5.1 Model goodness-of-fit

As with the earlier models, we assess how well these models fit the data with reference to the various diagnostic statistics described in Section 4.5.

5.1 Diagnostic statistics for the three multiple-factor congeneric models

Model	*GFI*	*AGFI*	*RMSEA*	*RMR*
Five-factor (25 items)	0.981	0.977	0.0414	0.118
Four-factor (20 items)	0.984	0.979	0.0437	0.104
Best fit (16 items)	0.989	0.983	0.0438	0.101

Overall, we conclude that each of the hypothesised models provides an adequate fit to the underlying data. The GFI, AGFI and RMSEA values all indicate that the models fit the data satisfactorily. Though the three models do have high RMR values that are above the recommended cut-off of 0.05.

5.2 Correlations among the latent variables

Extending the one factor congeneric model to analyse several sets of congeneric measures, allows estimates of the correlation between the unobservable latent variables. The correlation between each of the SDQ subscales for the three models are set out in the tables 5.2, 5.3 and 5.4 below. Each of the correlation estimates are statistically significant at conventional levels of significance.

5.2 Correlations among the five latent variables

	Emotional symptoms	*Conduct problems*	*Hyperactivity*	*Peer problems*	*Prosocial skills*
Emotional symptoms	1.000				
Conduct problems	0.679	1.000			
Hyperactivity	0.649	0.844	1.000		
Peer problems	0.771	0.766	0.622	1.000	
Prosocial skills	0.422	0.779	0.635	0.536	1.000

5.3 Correlations among the four latent variables

	Emotional symptoms	*Conduct problems*	*Hyperactivity*	*Peer problems*
Emotional symptoms	1.000			
Conduct problems	0.690	1.000		
Hyperactivity	0.646	0.837	1.000	
Peer problems	0.774	0.766	0.612	1.000

5.4 Correlations among the five latent variables – Best fit model

	Emotional symptoms	*Conduct problems*	*Hyperactivity*	*Peer problems*	*Prosocial skills*
Emotional symptoms	1.000				
Conduct problems	0.559	1.000			
Hyperactivity	0.664	0.657	1.000		
Peer problems	0.717	0.790	0.576	1.000	
Prosocial skills	0.369	0.659	0.478	0.532	1.000

We observe that the pattern of correlations between the latent variables is similar across the three models. The largest correlation between the unobservable constructs is between *Conduct problems* and *Hyperactivity* in both five- and four-factor models. The largest estimated correlation in the 'best fit' model is between the *Conduct problems* and *Peer problems* dimensions. These findings are in line with the expected correlation between these emotional and behavioural domains.

Path diagrams for each of the three models can be found in Appendix E.

5.3 Testing the best fit model across different populations

To this point, all the analyses presented above are based on a single sample. The focus now turns to models involving multiple samples. We do this to explore whether the best-fit measurement model (a model using 16 of the SDQ items) is equivalent (or invariant) across particular groups. In particular, we wish to know whether the items comprising the SDQ operate equivalently across different populations (for example, between boys and girls, or young and old children). The multi-sample analysis described below, by allowing us to explore whether the SDQ items are being interpreted in the same manner by the carers of children with different characteristics, is another step in assessing the reliability of the SDQ.

Our reasoning for using the best fit model over the four- and five-factor models to test across different populations is explained in Section 5 of the paper. We choose this model with less item error variance and better properties to estimate multi-sample analyses (i.e. models that stand a better chance of converging).

At the outset we should comment that we expect to see differences (i.e. non-equivalence) in carer perceptions of mental health and emotional problems with respect to boys and girls and across varying levels of relative isolation. These assessments are undertaken to establish these empirical properties. Thus assessment of the equivalence of the SDQ measurement model is undertaken across:

- boys and girls;
- young (under 11 years) and old (12–17 years);
- levels of relative isolation;
- Birth mother and non-Birth mother; and
- Aboriginal carer and non-Aboriginal carer.

In testing for the equivalence of the SDQ model across groups, we follow the approach of Byrne (1998) and test three hypotheses:

1. The number of underlying factors is equivalent across groups;
2. The factor loadings are equivalent; and
3. The factor covariances are equivalent.

However, instead of basing the analysis on covariance matrices (and maximum likelihood estimation) as Byrne (1998) does, we follow Joreskog's (2002) suggestion and compute mean vectors, covariance matrices and asymptotic covariance matrices for each of the population sub-groups we analyse. This approach allows the use of weighted least squares estimation to compare the population subgroups.

We illustrate the approach using gender as an example and then for brevity, we present summary results for the other groups.

5.4 Testing for the equivalence of a five-factor structure across gender

Our first step is to test for the equivalence of a five factor solution (with 16 indicators) in describing mental health across both boys and girls. This is done by combining the mean vectors, covariance matrices and asymptotic covariance matrices calculated separately for boys and girls into one LISREL input file. A two group, five factor baseline model (model 3 described in Section 5) is then estimated.

We first examine the estimated factor loadings across the two groups (in this case, boys and girls). The magnitude and pattern of factor loadings are similar across both groups. (The same comparison is made for the other population sub-groups – age, birth mother and Aboriginal carer. Once again we find the factor loadings are generally comparable across both groups. We do observe that the FRIEND indicator, has a stronger association with the latent variable *Peer problems* for the young, birth mother and non-Aboriginal carer groups.)

We then assess the validity of the five factor structure across both groups by examining the model's goodness-of-fit statistics. Based on

- $\chi^2(530)$ = 959.10
- RMSEA = 0.045
- GFI = 0.99

we conclude that based on the SDQ is adequately described by the hypothesised five-factor model fitted across boys and girls. As Byrne (1998) notes this in no way guarantees that the pattern and size of factor loadings is necessarily equivalent across both boys and girls. This hypothesis is tested in the next section.

5.5 Testing for the equivalence of factor loadings across gender

This tests the extent to which the strength of association between the individual items for both boys and girls is the same. To test this hypothesis we restate the model above to have equality constraints placed on all factor loadings across both groups. We test this hypothesis by comparing the difference in the chi-square measures between the two models with the change in the degrees of freedom associated with imposing the equality constraint in the second model.

Based on the chi-square measures we conclude that the factor loadings are not equal across gender ($\Delta\chi^2(27) = 580.22$). The 16 item, five-factor model is *not* measuring the same mental health aspects in exactly the same way for both boys and girls. In other words, SDQ items are being interpreted in different ways by the carers of Aboriginal children when it is applied to boys and girls. This is a finding common to many mental health instruments across the world and is in line with the common findings that carers of children place a different perceptual weight on the behaviours of boys and girls.

Given these results, we have undertaken some further analysis to determine which of the SDQ items are contributing to the inequality of factor loadings across boys and girls. Faced with testing all the possible combinations of the 16 indicators in the best fit model, the approach we take is to constrain the factor loadings in each SDQ subscale to be equal across population subgroups. For example, we first constrain WORRIES, SOMATIC and CLINGY (i.e. the *Emotional symptoms* items) to be equal across boys and girls and then compare the constrained model’s chi-square value to the baseline model. The other four subscales are compared in the same way. After testing these five constraints, we conclude that only the *Prosocial skills* subscale items are invariant across boys and girls.

5.6 Testing for the equivalence of factor covariances across gender

We test this hypothesis by imposing equality constraints on the covariances within the phi matrix (covariance matrix of unobservable variables). As Byrne (1998) notes, as each successive model is more restrictive than the former, this third hypothesis is tested by formulating a model in which both the factor loading matrix and factor covariances are constrained to be equal.

To test this hypothesis, we compare the fit of this third model with the first model (see Section 5.4). This gives a $\Delta\chi^2(37)$ value of 594.54. This is statistically significant, and therefore we conclude that the SDQ structure (factor loadings and covariances) are not the same for both boys and girls.

5.7 Testing for equivalence across other group characteristics

The same procedure is used to test for equivalence across

- Child's age (4–11 years vs. 12–17 years)
- Birth mother vs non-Birth mother.
- Aboriginal carer vs non-Aboriginal carer.

Table 5.5 summarises the results.

5.5 Summary of results for multi-sample analysis

Group	*Hypothesis*	χ^2	*df*	*p-value*	*Decision*
Boys and girls	1. Underlying factors equivalent	959.10	188		
	2. Factor loadings are equivalent	1,539.32	215	<0.00001	Reject
	3. Factor covariances are equivalent	1,553.64	225	<0.00001	Reject
Young and old	1. Underlying factors equivalent	965.81	188		
	2. Factor loadings are equivalent	4,181.05	215	<0.00001	Reject(a)
	3. Factor covariances are equivalent	4,209.50	225	<0.00001	Reject
Birth and non-Birthl mother	1. Underlying factors equivalent	1,112.69	188		
	2. Factor loadings are equivalent	3,987.59	215	<0.00001	Reject
	3. Factor covariances are equivalent	4,043.99	225	<0.00001	Reject
Aboriginal and non-Aboriginal carer*	1. Underlying factors equivalent	1,000.95	188		
	2. Factor loadings are equivalent	3,297.73	215	<0.00001	Reject(b)
	3. Factor covariances are equivalent	3,363.25	225	<0.00001	Reject

* Analysis based on a sample size of 3,964: 29 missing cases were excluded from the analysis.

(a),(b) – Models did not converge: chi-square estimates are preliminary.

Based on the probability values reported in the above table, we conclude that the SDQ is not being interpreted in the same manner by:

- The carers of boys (compared to the carers of girls);
- The carers of young children (compared to the carers of old children);
- Birth and non-Birth mothers; and
- Aboriginal and non-Aboriginal carers.

These procedures were applied to differences in relative isolation. The overall models did not converge. As a result, we further analysed levels of relative isolation by testing for equivalence between specific LORI categories. We started by testing for invariance between LORI level 1 (Perth metropolitan area) and LORI level 2 (rural town centres). Analysis supported the hypothesis of equal factor loadings (hypothesis 2) and covariances (hypothesis 3) at the $p < 0.05$ level in each case.

After accepting equivalence between level 1 and level 2, we then tested between LORI level 1 and LORI level 3 (moderate isolation). In this case, we concluded that both the pattern of factor loadings and covariances were not equivalent between LORI level 1 and LORI level 3. Based on these results we can say that the break in consistent interpretation of the SDQ occurs between those carers living in LORI level 1 (metropolitan and major rural town centres) and LORI level 3 (moderately isolated) localities.

5.8 Multi-level structural equation modelling

The previous analyses were based on single level models fitting structural equation models to the full sample of 3,993 Aboriginal children. This approach ignores the fact that there is clustering attributable to the carer in situations where there is more than one child per carer. We know that children living in the same family are likely to be more similar than children selected using simple random sampling.

As we noted in Section 4.2, the use of polychloric correlations with a weight matrix derived from the inverse of the asymptotic covariances as input to weighted least squares estimation is the most appropriate technique for analysis of ordinal data. Ideally, we would like to use this technique to fit multi-level structural equation models that takes into account the clustering attributable at both the child and family level. However, our attempts to fit multi-level structural equation models are constrained by limits to our computing capacity.

We have fit multi-level structural equation models using covariance matrices and maximum likelihood estimation (not reported). Not surprisingly, given the literature surrounding the analysis of ordinal data, the results from this method have been less than pleasing (in terms of low factor loadings).

6. A COMPOSITE MEASURE OF MENTAL HEALTH

In Sections 4 and 5 of this paper, the internal reliability of Goodman's SDQ model was tested using Confirmatory Factor Analysis. The results from this analysis are generally pleasing, suggesting that the observed indicators are capturing the unobservable dimensions of mental health they purport to measure. Having validated the SDQ model, our next step is to undertake model based analyses in an attempt to explore the factors that explain variation in Aboriginal children's mental health outcomes.

However, before this can be done, we need to reduce the 25 SDQ indicators into a composite measure that can be used to fit multi-level models. As Rowe (2003) notes "most theories and models in applied psychosocial research are formulated in terms of hypothetical constructs (or *latent* variables) that are not directly measurable or observable. As a means of data reduction it is common place to compute latent or composite variables such as *achievement, personality, performance standard* and so on, each measured on dichotomous or Likert-type ordinal scales".

We construct a composite measure of mental health by following the approach recommended by Rowe (2003). Our measure is based on factor score regression information from our preferred five-factor, sixteen indicator congeneric model of mental health.

6.1 Factor score regression coefficients

Indicator	*Factor score regression*	*Proportionately weighted factor score*
WORRIES	0.228	0.05379
UNHAPPY	0.496	0.11701
CLINGY	0.156	0.03680
FIGHTS	0.255	0.06016
LIES	0.259	0.06110
STEALS	0.277	0.06535
RESTLES	0.326	0.07691
FIDGETY	0.371	0.08752
DISTRAC	0.212	0.05001
RFRIEND*	0.205	0.04836
RPOPULAR*	0.198	0.04671
BULLIED	0.265	0.06252
RCONSID*	0.240	0.05662
RSHARES*	0.222	0.05237
RKIND*	0.241	0.05685
RCARING*	0.288	0.06794
TOTAL	**4.239**	**1.00000**

* Reverse coded

The factor score regression coefficients provide the relative amount each item is contributing to the overall estimation of the scale. The following factor scores are estimated (column 2 of table 6.1).

A proportionately weighted scale score for the composite measure of mental health that takes into account the individual and joint measurement errors of the 16 indicators can now be computed as a continuous variable by calculating a proportionately weighted factor score regression coefficient for each of the indicators. For example, a proportionally weighted factor score for the variable WORRIES is calculated by dividing its regression score coefficient (0.228) by the sum of the factor scores (4.239) which gives a proportionally weighted score of 0.0537 (Column 3 of the table above). Proportional weights for the other 15 indicators are calculated in the same way. The final proportionally weighted composite score is calculated by summing the product of the raw score for each indicator by its associated proportionally weighted factor score for each child's observation.

As Rowe (2003) notes there are at least two benefits to using this approach over a unit weighted additive index of the indicators (simply summing up the item responses). First, unit weighted addition of indicators (e.g. Goodman's scoring system) in forming scale scores ignores the possibility that some indicators typically contribute more to the measurement of the composite than others. And second, unit weight addition of the indicators may invalidate the composite score if one or more of the indicators 'measure' a construct other than the one under consideration.

6.1 Properties of the composite measure

The WAACHS dataset also contains a number of other indicators of mental health. This extra information is used to examine the properties of the composite measure. Polyserial correlations [5] between our continuous composite measure and the following ordinal indicators are shown in the table below.

5 Polyserial correlations are calculated as the data to be analysed is continuous versus ordinal. See Rowe (2003) for further details.

6.2 Polyserial correlations: Composite measure with other indicators of mental health

Mental health indicator	*Polyserial correlation estimate*
Self harm (a)	0.515
Talked about death or suicide (a)	0.308
Attempted suicide (a)	0.385
Do you think the child has emotional or behavioural difficulties?	0.544
Eating problems	0.361
Sleeping problems	0.463
Has nightmares	0.363
Bed wetting	0.213
Inappropriate sexual behaviour	0.379

(a) Based on 3,690 observations – very young children excluded.

The composite measure most highly correlates with emotional difficulties, self harm and sleeping problems. All correlations are statistically significant at the $p < 0.01$ level. Given these results, our findings from Section 5, and the advantages of this approach described by Rowe above, this composite measure is used in the subsequent modelling described in Section 7.

We have also been able to examine the correlation of the composite measure with the use of mental health services as consent from carers was sought to access hospital records. Data on the use of mental health services by children and carers was obtained by linking survey responses with administrative health records. Consent rates for record linkage were very high. Approximately 97 per cent of primary carers and 92 per cent of secondary carers gave consent for their records to be linked. The correlation between our composite measure and the use of mental health services (by children) was found to be 0.306 (significant at the 0.05 level).

6.2 An alternative measure of mental health

Notwithstanding the advantages associated with the composite measure described above, we also consider an alternative measure of mental health based on the total SDQ score calculated according to the scoring system proposed by Goodman (namely, the sum of the items on all scales except the *Prosocial skills* scale). Our rationale for doing this is that Goodman's measure has been well tested and known to be valid for the general population. Furthermore, when reviewing the international literature, the SDQ total score has been widely used in assessing mental health outcomes. Therefore to aid in cross-national comparisons of our data, we also fit multi-level models using this measure (these alternative model estimates are discussed in Section 7). Model results using both measures are compared and contrasted.

At the outset, we note that the composite measure and Goodman's measure are highly correlated (0.92, p-value < 0.0001). Polyserial correlations of the other mental health indicators with the composite measure and Goodman's measure are reported in table 6.3 below.

6.3 Polyserial correlation analysis – Composite and Goodman's measure

	Polyserial correlation estimate	
Mental health indicator	*Composite measure*	*Goodman's measure*
Self harm (a)	0.515	0.532
Talked about death or suicide (a)	0.308	0.306
Attempted suicide (a)	0.385	0.380
Do you think the child has emotional or behavioural difficulties?	0.544	0.553
Eating problems	0.361	0.373
Sleeping problems	0.463	0.488
Has nightmares	0.363	0.414
Bed wetting	0.213	0.227
Inappropriate sexual behaviour	0.379	0.402

(a) Based on 3,690 observations – very young children excluded.

A comparison of the two measures reveals that the correlation estimates with the other mental health indicators are quite similar. The composite measure has a slightly higher correlation with the 'talked about death' and 'attempted suicide' indicators, while Goodman's measure is more highly correlated with the other indicators.

7. MULTI-LEVEL MODELLING

In the preceding analysis, single-level models have been fitted to the SDQ items. However, we have good reason to believe that the data should be looked at using a multi-level structure accounting for its hierarchical nature. For example, children are nested within carers, who are nested within households, which, in turn, are nested within communities. Not fully taking into account the structure of the data can lead to unsubstantiated conclusions. As Rowe (2003) notes, "... the existence of such clustering poses special problems that lead to several long-standing and troublesome obstacles to statistical conclusion validity. Failure to account for the multi-level nature of data, invariably leads to an increased probability of committing Type 1 errors (falsely rejecting the null hypothesis) with important ramifications for the substantive interpretation of findings and their related policy implications". The advantage of multi-level models is that by incorporating the multi-level structure of the data into the model allows both within and between level (e.g. carer) variation to be analysed.

The nature of the survey data thus presented several challenges for statistically appropriate analysis. Unlike data collected from a simple random sample, the survey children are clustered within families and communities. The sample was selected in three stages: Census Collection Districts (CDs), families and children. CDs were selected with probabilities of inclusion in the survey proportional to the number of Aboriginal and Torres Strait Islander children living in the CD. Once families had been selected, each Aboriginal and Torres Strait Islander child under the age of 18 years was selected in the survey. As a result of this selection hierarchy, the data for individual children in the survey sample violate one of the basic assumptions of traditional regression modelling: that the observations are independent.

For many data items, children within the same family are more likely to have the same characteristics than children chosen from throughout the state using simple random sampling. Multi-level, or hierarchical, modelling can be used to account for the hierarchical structure of the survey data (Goldstein, 2003). However, the analysis is further complicated because unequal probabilities of selection have been used. CDs have been selected into the sample with probabilities proportional to the number of in-scope children. Survey weights have also been developed to adjust for different levels of non-response by age group and family size. While there are techniques to model data collected from surveys where unequal weights are used, and a range of software available that can fit multi-level models, addressing both issues at the same time is a relatively new statistical challenge.

Pfeffermann, et al. (1998) proposed a technique, called Probability Weighted Iterative Generalised Least Squares (PWIGLS) that can fit a multi-level model accounting for a complex survey design. The PWIGLS technique as described by Pfeffermann, et al. fits a two-level model to a normally distributed continuous variable. We have adapted this technique for the WAACHS where we wanted to model a three-level hierarchy: children within families within communities. As many of the survey variables are binary indicators we have also adapted the PWIGLS technique to fit logistic regression models. These new techniques have been implemented within SAS software. As far as we know, this is the first time such techniques have been used in a full-scale survey.

In this section the models have been fitted accounting for both the hierarchical structure of the data, and the survey design and survey weights. Multi-level models are an ideal analytic tool for use in the survey, as they enable children's health and well-being to be described in terms of not only child level factors, but family and community level factors as well. The use of survey weights allows us to generalise the results of the models to the entire population of Aboriginal children in Western Australia.

The benefit of multi-level models over single-level models is that they provide potentially important information about the *context* in which each individual is living. For example, in a traditional single-level explanatory model of individual health outcomes, it is impossible to determine whether the effect of low socioeconomic status (SES) for individuals living in Sydney is the same for low SES individuals in Brisbane. It is possible to include indicator variables for the different areas in which people live – however this is impractical for a large number of areas. Multi-level models allow us to model the effects at both levels simultaneously (individual and area in this example) and compare the variance explained by both individual and contextual covariates. (See Rowe, 2003 for a fuller discussion).

7.1 Method

In this section, we first fit a two-level model (individual children and families), we then fit a three-level model (individual children, families, and Census Collection Districts). We are interested in exploring what we can learn by extending the single-level models described previously to account for the multi-level data structure. Multi-level models are fitted to analyse how much variation in mental health can be attributed to differences between carers (and later between Census Collection Districts).

Before doing this, our first step is to test whether the hierarchical nature of the data is significant. This is done by determining the proportion of variance in child mental health that is due to the carer differences. This is done by fitting the simple variance components model described below. In our case, we have 3,993 children clustered within 1,704 families.

Using the subscript i to refer to children and the subscript j for the carer, this model can be written in two parts (see Rowe, 2003):

A within-families, among children part –

$$Y_{ij} = B_{0j}x_0 + e_{ij} \tag{7.1}$$

and a between-families part –

$$B_{0j} = B_0 + u_{0j} \tag{7.2}$$

By combining equations (7.1) and (7.2), a single equation version of the model can be written as:

$$Y_{ij} = B_0 + B_{0j}x_0 + (u_{0j} + e_{ij}) \tag{7.3}$$

where

Y_{ij} – is the normalised composite SDQ score for child *i* in family *j*'s care;

B_0 – is the 'average' SDQ score of children in the sample of families;

B_{0ij} – the amount that the intercept term estimated for each family varies around the grand mean (B_0);

u_{0j} – is a residual that varies randomly between families.

x_0 – is a column of 1s; and

e_{ij} – is a random variable that is assumed to have a mean of 0 and represents the sum of all other influences on the response variable Y_{ij}.

Each of these terms is described in more detail in the results section that follows.

In the multi-level models that follow, we do not use the original SDQ score based on 20 items with a Total Score ranging between 0 and 40 as the dependent variable (Y_{ij}). Rather, a composite measure for the total SDQ score is constructed. This composite measure is calculated based on factor score regression coefficients from our preferred five-factor, sixteen item model described in Section 5. The full details of how the composite measure is calculated and the benefits of this approach over a simple summing of the SDQ responses have been discussed in the previous section.

We also normalise the composite measure as this is a key assumption of the linear model. (Appendix F discusses how this is done). The normalised composite measure of mental health is then used to fit the multi-level model described above.

7.2 Results

7.2.1 A two-level variance components model (Model 1)

The variance components model described above is estimated to determine the proportion of variance in carer rated SDQ scores due to between carer differences in the following form:

$$Y_{ij} = B_{0j}x_0 + u_{0j} + e_{ij} \tag{7.4}$$

The estimated model parameters and their standard errors (in parentheses) are:

$$B_{0j} = 0.48062 \quad (0.00798)$$
$$\sigma^2_{u0} = 0.07924 \quad (0.00327)$$
$$\sigma^2_{e} = 0.04078 \quad (0.00185)$$

7.1 Family level variation in child mental health

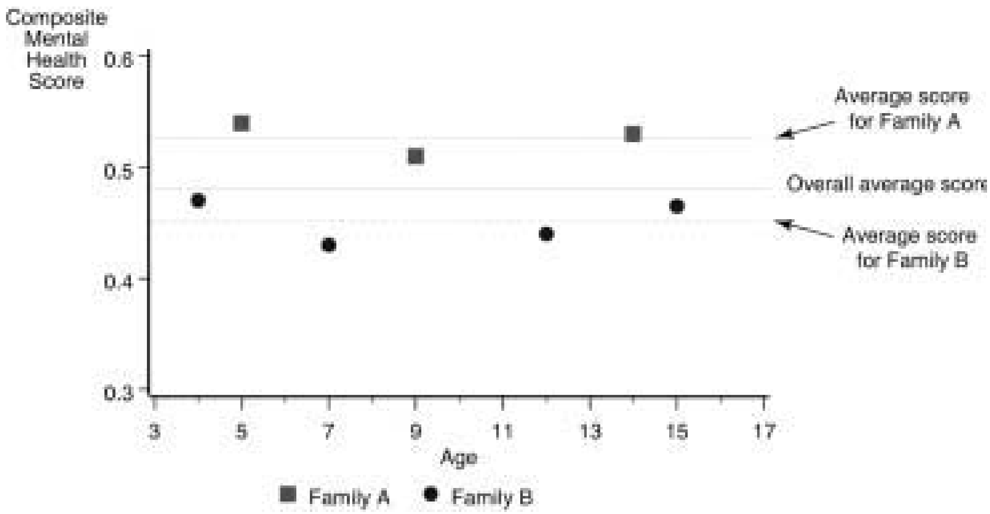


Figure 7.1 illustrates the general principal of the analysis with two hypothetical families. Family A has three children and Family B has four children. Each child's mental health score is plotted. As can be seen, there is a within family mean mental health score for each family. A total mean can also be constructed representing the overall average mental health score across all children and families. The figure also shows that children within families tend to be more similar (in the sense that their scores are less variable) than children between families.

The constant has a mean of 0.48, this can be interpreted as the overall mean 'mental health score' for all children. The 3,993 children are clustered within 1,704 families.

Using the family level (σ_{u0}^2) and child level (σ_e^2) variances estimated above, an intra-family correlation can be calculated as:

$$\rho = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_e^2} \qquad (7.5)$$

The model shows that the ratio of the parameter estimates to their standard errors for the family (σ_{u0}^2) and child level (σ_e^2) residual variances are both large and statistically significant, indicating stable variation at these levels. Using equation 5, of the total variance in the SDQ score, about two-thirds (66%) of it (0.07924/0.12002) is accounted for by family level effects and the other one-third is accounted for by child level effects.

This result implies that the mental health of children within families tends to be judged as more similar than the mental health of children in difference families. This is a reasonable result as children within families are more likely to be subject to similar behavioural influences.

7.2.2 A multi-level regression model (Model 2)

This estimate of family level clustering may be misleading if there are differences in mental health between younger and older children, or between boys and girls for example.

We explore whether this is the case in a multi-level modelling framework by extending the variance components model described above to control for other children and family characteristics. Specifically, we control for the following characteristics:

- Age (years) – X_{1ij}
- Male/female – X_{2ij}
- Levels of relative isolation (LORI) – X_{3ij}
- Birth/non-Birth mother – X_{4ij}
- Whether the child has a physical health problem – X_{5ij}

The model can be explicitly stated in the following form:

$$Y_{ij} = B_{0j}x_0 + B_{1j}X_{1ij} + B_{2j}X_{2ij} + B_{3j}X_{3ij} + B_{4j}X_{4ij} + B_{5j}X_{5ij} + u_{0j} + e_{ij} \quad (7.6)$$

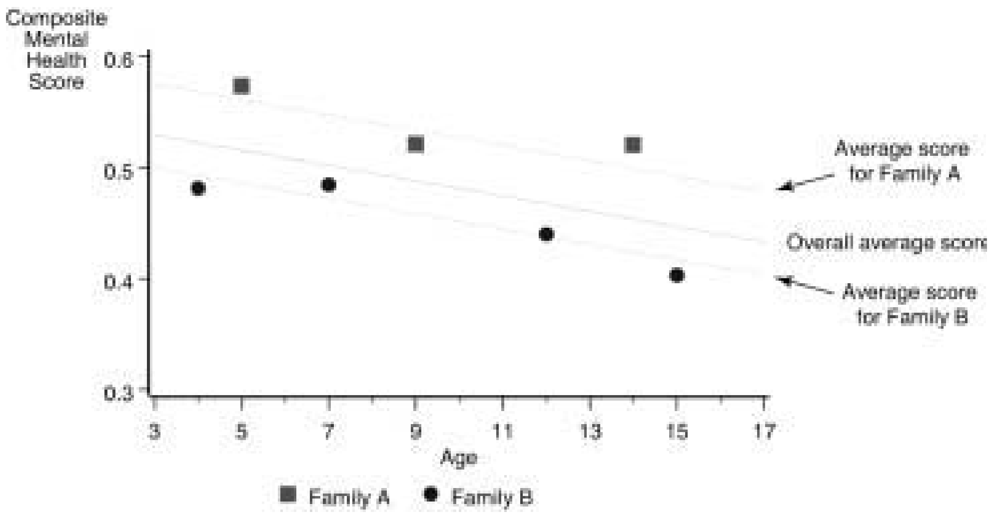
Model estimates are provided in table 7.2.

7.2 Two-level regression model

Explanatory variable	*Coefficient*	*Standard error*	*p-value*
Fixed part of the model –			
Constant	0.68582	0.01286	$p < 0.01$
Child level			
Age	–0.00766	0.00053	$p < 0.01$
Gender	–0.07528	0.00378	$p < 0.01$
Physical health problem	0.07766	0.00418	$p < 0.01$
Family level			
Birth mother	–0.01927	0.00615	$p < 0.01$
Remoteness	–0.01724	0.00298	$p < 0.01$
Random part of the model –			
Child level variance	0.05195	0.00046	$p < 0.01$
Family level variance	0.07068	0.00115	$p < 0.01$

An example of this model is shown graphically in figure 7.3. Once again we use a stylised two-family example to illustrate the key features of the model.

7.3 Two-level regression model



The total variation in mental health is the sum of these two components (0.1226). Of this about 58% (0.0707 / 0.1226) is due to differences between carers and the remainder (42%) is due to children.

Model 2 shows that we can adjust for the effects of age, level of relative isolation, gender, etc., but even in so doing, there are significant effects attributable to clustering at the individual and family level.

7.2.3 Three-level model with Census Collection District effects (Models 3 and 4)

To what extent do children living in the same Census Collection District have similar mental health?

We use Census Collection District (CD) level information to proxy for a child's neighbourhood[6]. As Snijders and Bosker (1999) note the three-level random intercept model is a straightforward extension of the two-level model. The dependent variable is now denoted by Y_{ijk}, referring to, child i, in carer j's care, in CD k. There are also three residuals as there is now variability at three levels.

In this case, the model is based on 3,993 children clustered with 1,704 families clustered within 530 Census Collection Districts (CDs). As before the response variable is the normalised composite measure of child mental health.

Table 7.4 contains the results from the three-level variance components model (Model 3) and the three-level model controlling for other child and family level covariates (Model 4).

7.4 Three-level regression model

	Model 3			*Model 4*		
Explanatory variable	*Coefficient*	*s.e.*	*p-value*	*Coefficient*	*s.e.*	*p-value*
Fixed part of the model –						
Constant	0.48110	0.01000	< 0.01	0.65320	0.03028	< 0.01
Child level						
Age				–0.00712	0.00145	< 0.01
Gender				–0.07592	0.00970	< 0.01
Physical health problem				0.07360	0.00982	< 0.01
Family level						
Birth mother				–0.01630	0.01545	0.2912
Census Collection district level						
Remoteness (a)						
Low				–0.01573	0.02286	0.4913
Moderate				–0.02539	0.02867	0.3758
High				–0.03778	0.03897	0.3323
Extreme				–0.05843	0.03301	0.0767
Random part of the model –						
Child level variance	0.04075	0.00196	< 0.01	0.03877	0.00184	< 0.01
Family level variance	0.05198	0.00315	< 0.01	0.04960	0.00290	< 0.01
Census Collection District level variance	0.03245	0.00323	< 0.01	0.03076	0.00301	< 0.01

(a) Reference category is LORI Level 1 (metro area)

6 Census Collection Districts (CDs) are an administrative unit and are not designed to explicitly capture a neighbourhood or community. In the absence of other data in the WAACHS, we use it as the best available measure of neighbourhoods.

The three-level model shows that the total variance in child mental health is 0.1251 (0.03245 + 0.05198 + 0.04075). About 26% of the variance in mental health is attributable to clustering at the CD level, 41% at the family level and 32% at the child level. This model can be extended to control for other child and family characteristics (Model 4). The same explanatory variables that were used in Model 2 are also used to fit this model. From Model 4, we can see that the total variance in the response variable is 0.1191 (the sum of the level one, two and three variances). Comparing these results to Model 3, we observe that variance in mental health at each level is very similar (about 26% at the CD level, 42% at the family level and 32% at the child level). The inclusion of additional explanatory variables makes little difference to the estimates of the amount of clustering that occurs at each of the three levels – that is, clustering is not due to age, level of relative isolation, gender, etc..

7.3 Sensitivity analysis – alternative response variable

The multi-level modelling throughout Section 7 has used a composite variable specifically constructed for this purpose (see Section 6). This composite measure, constructed with weights based on factor score regression coefficients, used sixteen of the original SDQ variables shown to provide a robust fit to an underlying five factor model of mental health.

To what extent do the results of the multi-level analysis change if the Total SDQ score as described by Goodman (i.e. using 20 of the items) is used as the response variable?

In tables 7.5 and 7.6, we report key multi-level model results for each of the two measures.

7.5 Sensitivity analysis – Two-level model (children and families)

Result	*Composite measure*	*Goodman's measure*
Child level variance	0.07068	18.1657
Family level variance	0.05195	30.5630
Intra family correlation	57.63 %	62.72 %
Fixed effects		
Age	negative	negative
Gender	negative	negative
Remoteness	negative	negative
Physical health	positive	positive
Birth mother	negative	negative

7.6 Sensitivity analysis – Three-level model (children, carers and region)

Result	*Composite measure*	*Goodman's measure*
Child level variance	0.03877	12.3223
Family level variance	0.04960	21.4226
Census Collection District level variance	0.03076	13.5546
Percentage of variance due to clustering at		
– child level	32.54 %	26.05 %
– family level	41.63 %	45.29 %
– Census Collection District level	25.82 %	28.65 %
Fixed effects		
Age	negative	negative
Gender	negative	negative
Remoteness		
– Low	negative (a)	negative (a)
– Moderate	negative (a)	negative
– High	negative (a)	negative (a)
– Extreme	negative (a)	negative
Physical health	positive	positive
Birth mother	negative (a)	negative (a)

(a) not statistically significant – 5% confidence level

When we examine tables 7.5 and 7.6, we can observe that we draw the same substantive conclusions regardless of which measure is used to model the mental health of Aboriginal children. Results from both measures suggest that most of the variability in children's mental health can be explained by differences between families.

8. LIMITATIONS

As noted in the introduction, the WAACHS data has many features of modern complex survey designs with stratification, multiple stages of selection and unequal selection probabilities. The analysis reported here has some important limitations.

Some models in this report failed to converge. Difficulties with these estimations occurred particularly where sub-samples were smaller (e.g. within Levels of Relative Isolation), or where the underlying construct demonstrated poor scale reliability (e.g. *Peer problems*). Under some circumstances this affected estimations of scale reliability for *Peer problems* within some Levels of Relative Isolation; it affected the multi-sample analysis of underlying factor, factor loading and covariance equivalency with respect to Levels of Relative Isolation.

Some of these problems were overcome by aggregating data to increase sub-sample sizes – for example, it was possible to assess underlying factor, factor loading and covariance equivalency with respect to Levels of Relative Isolation for metropolitan and rural centres vs more remote regions. It was also possible to assess scale reliabilities across the entire sample for all SDQ subscales and Levels of Relative Isolation for the Total Score.

None the less, problems with convergence undoubtedly reflect the underlying metric of the variables and as well the specific pattern of their association within sub-samples.

9. CONCLUSIONS

This paper has focussed on analysing the Strengths and Difficulties Questionnaire (Goodman, 2001) – a 25 question instrument that purports to capture five dimensions of mental health (*Emotional symptoms*, *Conduct problems*, *Hyperactivity*, *Peer problems* and *Prosocial skills*). It is the principal method used in the WAACHS to assess the mental health of Aboriginal children and young people aged 4–18 years

We have focused on two main research themes.

9.1 The internal validity and reliability of the SDQ scale.

The principal statistical methods used to assess the internal validity and reliability of the SDQ and its five subscales is Confirmatory Factor Analysis (CFA).

Initially, CFA is used to fit one-factor congeneric models to the ordinal scaled indicators. Five separate models are estimated – one for each subscale. The reliability of each subscale is assessed with reference to

- examination of the factor loadings,
- various diagnostic model fit statistics,
- scale reliability coefficient suggested by Raykov.

Assessed against these criteria, we find that the *Peer problems* subscale is the least reliable when applied in assessing mental health of Aboriginal children. The results for the other four subscales are generally pleasing. All four have a calculated scale reliability over 0.70. Furthermore, the WAACHS data show adequate congruence with data reported by Goodman (2001) on a representative sample of British children.

We have further assessed scale reliability by remoteness. For each of the five subscales, internal reliability declines as a child resides in a more remote locality. As with the overall scale reliabilities, the *Peer problems* subscale performs more poorly in terms of internal consistency when analysed by remoteness.

These single-level, single factor models are extended to allow the latent variables to correlate with each other (multiple factor congeneric models). Given that we have no strong a priori hypothesis of the factor structure underlying the SDQ measurement model (for example, data are collected on five subscales, but only four of these are used in the actual scoring of the SDQ) we separately estimate and test three models:

- A five-factor model comprising the five SDQ subscales and 25 observed indicators;

- A four-factor model (*Emotional symptoms*, *Conduct problems*, *Hyperactivity*, *Peer problems*) and 20 indicators based on the items used to calculate Goodman's SDQ total score.
- An optimal model of five factors and 16 observed indicators selected on the basis of the strength of their association with their respective underlying factor.

Based on various goodness-of-fit statistics, we conclude that each of the hypothesised models provides an adequate fit to the underlying data. We choose to use the 16-indicator model in hierarchical modelling. We felt this model retained a set of items that strongly measured their underlying latent variables (in the sense of having lower proportions of error variance) and exhibited better properties for multi-sample and multi-level modelling.

Further assessment of SDQ reliability was undertaken by running various multi-sample analyses. This permitted assessing model equivalency across particular groups (e.g. between boys and girls, young and old).

From this analysis, we find that carers' reports of their child's mental health and well-being varied with respect to:

- the child's age (aged 4–11 years vs 12–17 years),
- the level of relative isolation,
- the Birth and non-Birth status of the mother, and
- the Aboriginal and non-Aboriginal status of the carer.

9.2 Multi-level effects in SDQ reports: Modelling mental health outcomes

Several weighted multi-level models of mental health were also estimated. This allowed the estimation of variation in Aboriginal children's mental health due to differences between their carers or the community in which they reside. A composite measure of mental health based on our preferred five-factor, 16 indicator multiple congeneric model was constructed and used to model mental health within a multi-level framework. The three-level model shows that about 26% of the variance in mental health is attributable to clustering at the CD level, 41% at the family level and 32% at the child level. The inclusion of additional explanatory variables made no difference to the estimates of the amount of clustering that occurs at each of the three levels – that is, clustering is not due to age, level of relative isolation, gender, etc..

9.3 Using the SDQ to assess Aboriginal children's mental health

A range of statistical analyses have been undertaken to test the concurrent validity and scale reliability of the SDQ subscales and Total Score. Results from single-level congeneric models are generally satisfactory. Internal reliabilities for four of the five subscales (*Emotional symptoms*, *Conduct problems*, *Hyperactivity* and *Prosocial skills*) are all good (exceeding 0.70). Further analysis that allows the unobservable mental health dimensions to correlate with each other (i.e. multiple-factor congeneric models) indicate that the three hypothesised models provide a good fit to the underlying data. These results suggest that the observed indicators are capturing the unobservable dimensions of mental health they purport to measure. While there are undoubtedly steps that could be taken to improve the SDQ and its metric properties that would result in better scale reliability and efficiency, as used in the WAACHS, the SDQ Total Score provides a reasonable measure of mental health and well-being in Aboriginal Australian children and young people.

ACKNOWLEDGEMENTS

The authors would like to thank Francis Mitrou for assistance with data queries, data analysis and advice on models. Thanks are also due to Monika Ciurej for help with further data analysis. Responsibilities for mistakes or omissions remain with the authors.

REFERENCES

Australian Bureau of Statistics (2002) *National Health Survey: Aboriginal and Torres Strait Islander Results, Australia, 2001*, cat. no. 4715.0, ABS, Canberra.

Australian Bureau of Statistics (ABS) and Australian Institute of Health and Welfare (AIHW) (2005) *The Health and Welfare of Australia's Aboriginal and Torres Strait Islander Peoples*, ABS cat. no. 4704.0 and AIHW cat. no. IHW14, ABS, Canberra.

Bearden, W., Sharma, S. and Teel, J. (1982) 'Sample Size Effects on Chi-square and Other Statistics Used in Evaluating Causal Models', *Journal of Marketing Research*, vol. XIX (November 1982), pp. 425–30.

Bollen, K. (1989) *Structural Equations With Latent Variables*, John Wiley & Sons.

Browne, M.W. And Cudeck, R. (1993) 'Alternative Ways of Assessing Model Fit' in Bollen, K.A. and Long, J.S. (eds.) *Testing Structural Equation Models*, Sage.

Byrne, B.M. (1998) *Structural Equation Modelling with LISREL, PRELIS and SIMPLIS. Basic Concepts, Applications and Programming*, Lawrence Erlbaum Associates Publishing.

Dillon, W. and Goldstein, M. (1984) *Multivariate Analysis Methods and Applications*, John Wiley and Sons.

Fergusson, D., Hong, B., Horwood, J., Jensen, J. and Travers, P. (2003) *Living Standards of Older New Zealanders: A Technical Account*, http://www.msd.govt.nz/documents/publications/csre

Fullarton, S. (2002) *Student Engagement With School: Individual and School Level Influences*, Australian Council for Education Research.

Goldstein, H. (2003) *Multi-level Statistical Models*, Third Edition, London: Arnold.

Goodman, R. (2001) 'Psychometric Properties of the Strengths and Difficulties Questionnaire', *Journal of the American Academy of Child and Adolescent Psychiatry*, November 2001, pp. 1337–1345.

Goodman, R. (1997) 'The Strengths and Difficulties Questionnaire: A Research Note', *Journal of Child Psychology and Psychiatry*, vol. 38, pp. 581–586.

Goodman, R., Meltzer, H. and Bailey, V. (1998) 'The Strengths and Difficulties Questionnaire: A Pilot Study on the Validity of the Self-Report Version', *European Child and Adolscent Psychiatry*, vol. 7, pp. 125–30.

Gray, M.C., Hunter, B.H. and Taylor, J. (2002) *Health Expenditure, Income and Health Status Among Indigenous and Other Australians*, Research Monograph 21, Centre for Aboriginal Economic Policy Research, Canberra.

Hair, J.F., Anderson, R.E., Tatham, R.L. and Black, W.C. (1998) *Multivariate Data Analysis*, Fifth Edition, Englewood Cliffs, NJ: Prentice Hall.

Hayduk, L. (1987) *Structural Equation Modelling with LISREL: Essentials and Advances* Boston, John Hopkins University Press.

Hayduk, L. (1996) *LISREL: Issues, Debates and Strategies*, Boston, John Hopkins University Press.

Joreskog, K. (2002) *Structural Equation Modelling with Ordinal Variables Using LISREL*, http://www.ssicentral.com/lisrel/corner.htm

Joreskog, K. and Sorbom, D. (1989) *LISREL 7. A Guide to the Program and Applications*, Second Edition, SPSS Inc..

Joreskog, K. and Sorbom, D. (1996) *LISREL*, Scientific Software International.

Jorekog, K., Sorbom, D., du Toit, S, and du Toit, M. (2001) *LISREL 8: New Statistical Features*, Lincolnwood: SSI Scientific Software International.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998) 'Weighting for Unequal Selection Probabilities in Multi-level Models', *Journal of the Royal Statistical Society (Series B)*, vol. 60, pp. 23–40.

Raykov, T. (2003) *Scale Reliability with LISREL 8.50*, http://www.ssicentral.com/lisrel/mainlis.htm

Rowe, K. (2003) *Practical Multi-level Analysis with MLwiN and LISREL: An Integrated Course*, Camberwell, Victoria: Australian Council for Education Research.

Snijders, T. and Bosker, R. (1999) *Multi-level Analysis: An Introduction to Basic and Advanced Multi-level Modelling*, London: Sage Publications.

West S., Finch J. and Curran, P. (1995) 'Structural equation models with nonnormal variables: Problems and remedies' in Hoyle, R. (1995) *Structural Equation Modelling Concepts, Issues and Applications* California, Sage Publications Inc..

World Health Organisation (2004) *Mental Health: Overview*, January 15, 2004, http://www.afro.who.int/mentalhealth/

Zubrick, S.R., Lawrence, D.M., Silburn, S.R., Blair, E., Milroy, H., Wilkes, T., Eades, S., D'Antoine, H., Read, A., Ishiguchi, P. and Doyle, S. (2004) *The Western Australian Aboriginal Child Health Survey: The Health of Aboriginal Children and Young People*, Perth: Telethon Institute for Child Health Research http://www.ichr.uwa.edu.au/research/divisions/pop/projects/waachs/)

APPENDIXES

A. SCORING THE SDQ

A.1 Scoring the SDQ

SDQ Subscale	*No*	*Sometimes*	*Yes*	*Variable name**
Emotional symptoms scale				
Often complains of headaches, stomach aches or sickness	0	1	2	SOMATIC
Often seems worried	0	1	2	WORRIES
Often unhappy, sad or tearful	0	1	2	UNHAPPY
Nervous or clingy in new situations, easily lost confidence	0	1	2	CLINGY
Many fears, easily scared	0	1	2	AFRAID
Conduct problems scale				
Often has temper tantrums	0	1	2	TANTRUM
Usually done what adults told him/her to do	2	1	0	ROBEYS
Been in fights with other children or bullies them	0	1	2	FIGHTS
Often lies or cheats	0	1	2	LIES
Steals from home, school or elsewhere	0	1	2	STEALS
Hyperactivity scale				
Restless, overactive can not stay still for long	0	1	2	RESTLES
Constantly fidgeting or squirming	0	1	2	FIDGETY
Easily distracted, or poor concentration	0	1	2	DISTRAC
Able to stop and think things out before acting	2	1	0	RREFLECT
Good attention span and finished the things they start	2	1	0	RATTENDS
Peer problems scale				
Tends to play by themself	0	1	2	LONER
Has at least one good friend	2	1	0	RFRIEND
Generally liked by other children	2	1	0	RPOPULAR
Picked on or bullied by other children	0	1	2	BULLIED
Gets on better with adults than with other children	0	1	2	OLDBEST
Prosocial skills scale				
Considerate of other people's feelings	0	1	2	RCONSID
Readily shares with other children	0	1	2	RSHARES
Helpful if someone is hurt, upset or feeling ill	0	1	2	RCARING
Kind to younger children	0	1	2	RKIND
Often volunteers to help others	0	1	2	RHELPOUT

* Variable names used in structural equation models.

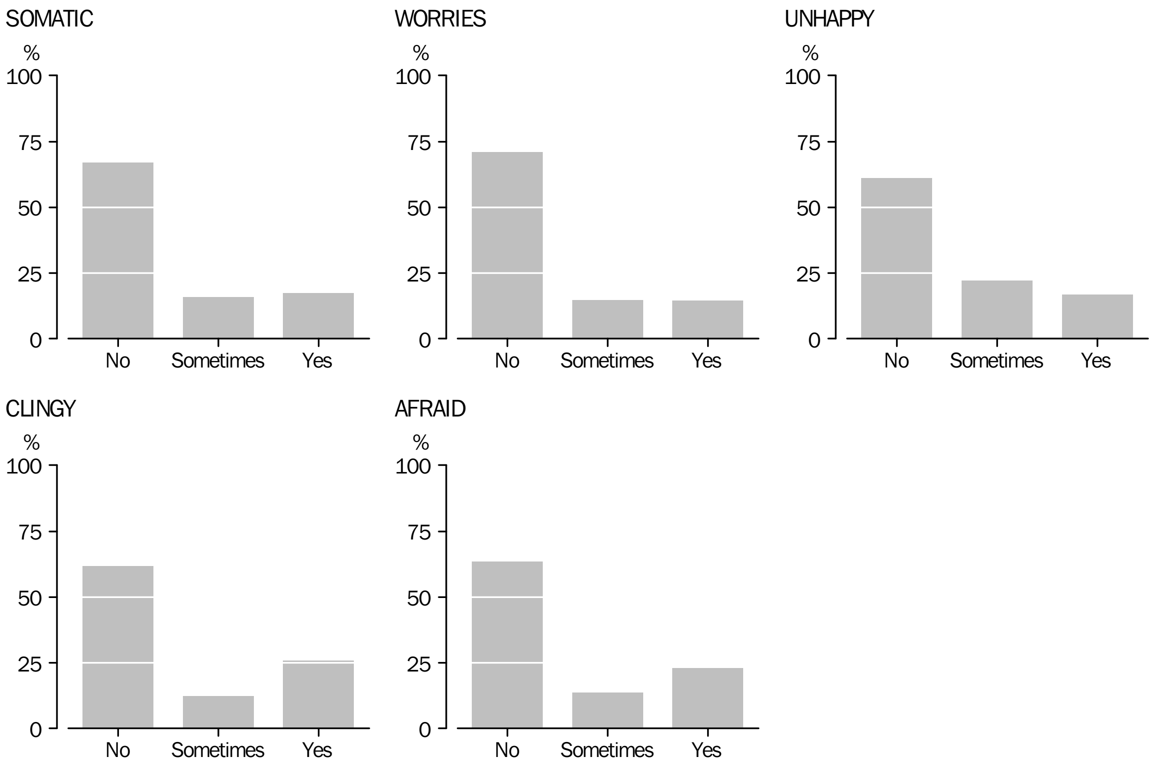
B. ANALYSIS OF SDQ ITEMS AND SDQ SCORES FOR CHILDREN AGED 4–17 YEARS.

B.1 Analysis of the 25 items

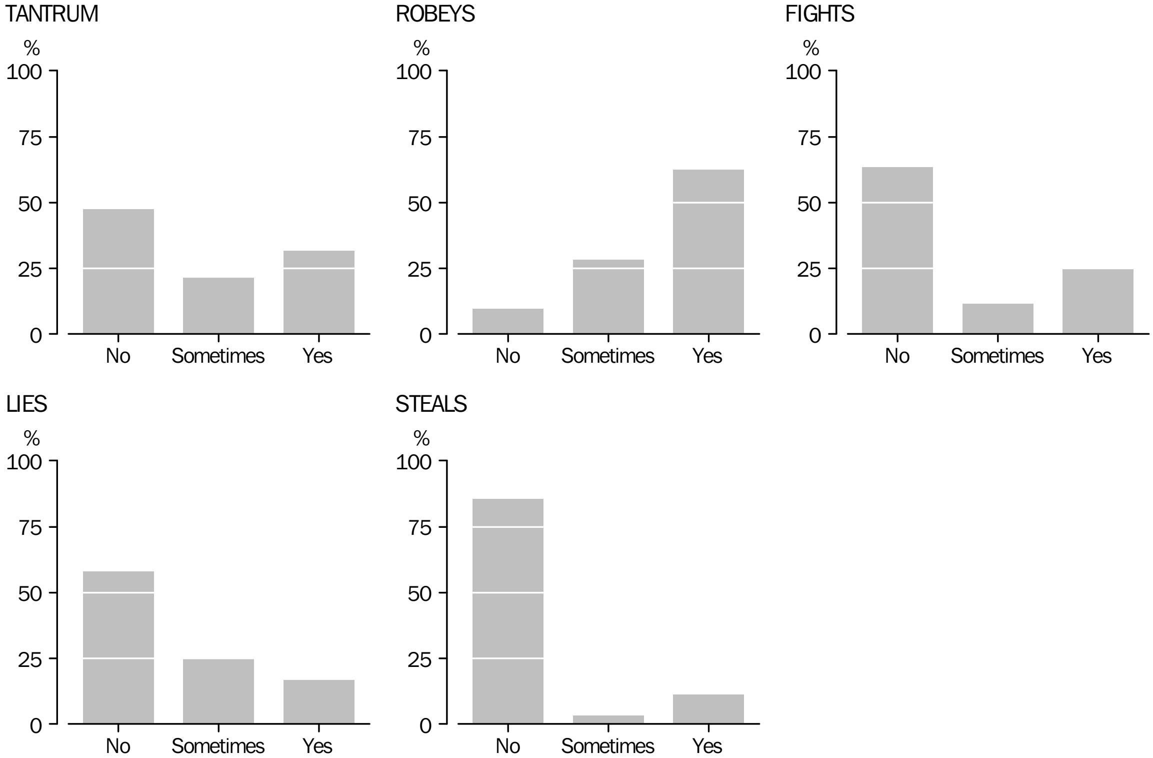
Univariate distributions for each of the 25 SDQ indicators are presented below.

Visual inspection of these charts indicate that the distributions are non-normal being skewed or U-shaped (and in some instances showing low < 5% response categories).

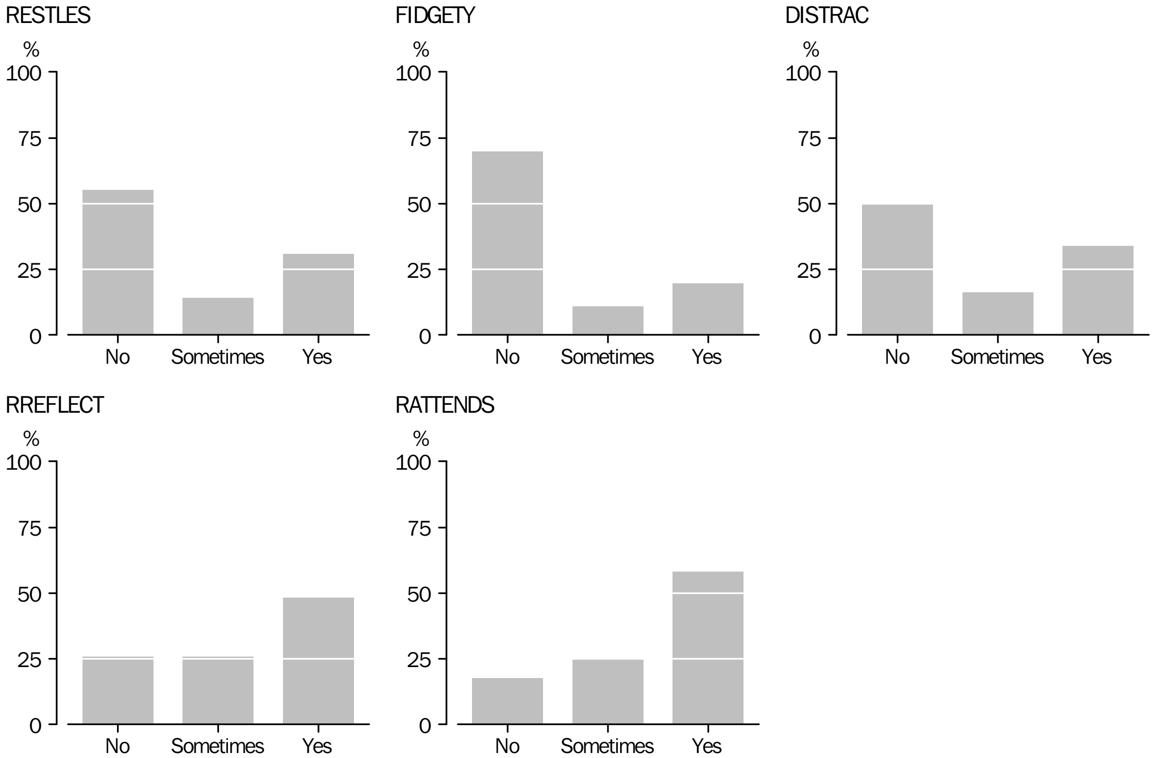
B.1 Emotional symptoms scale



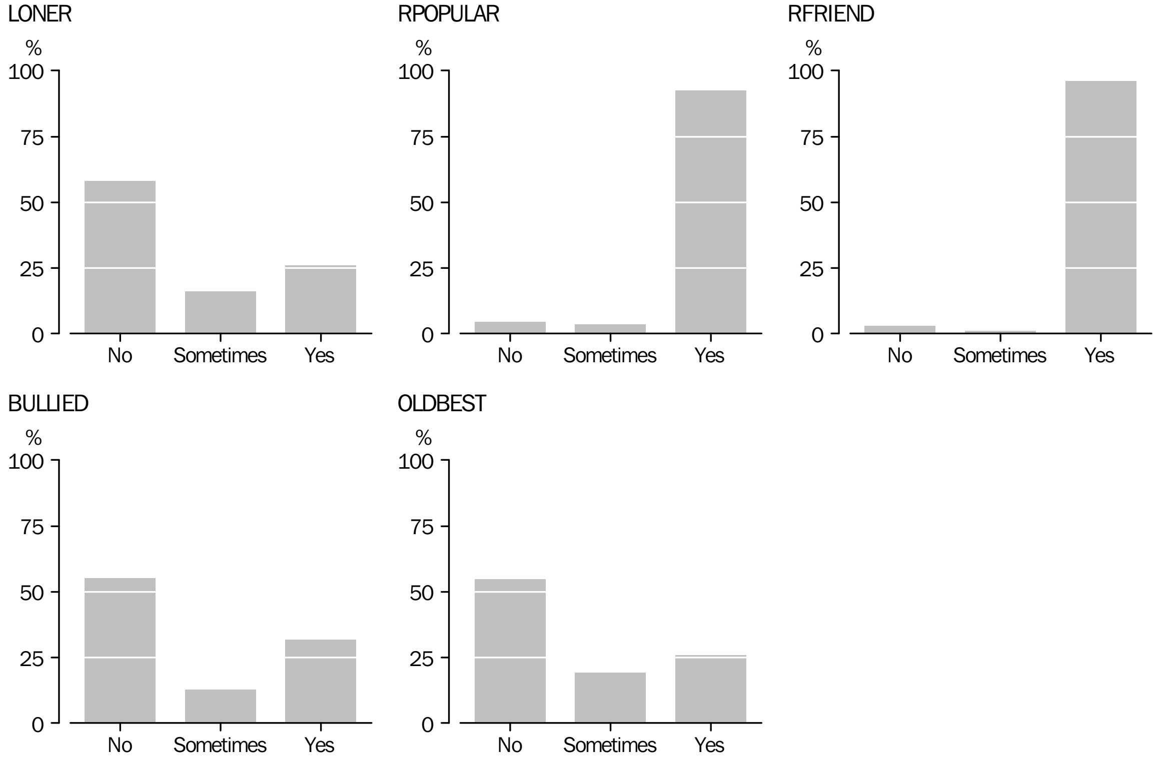
B.2 Conduct problems scale



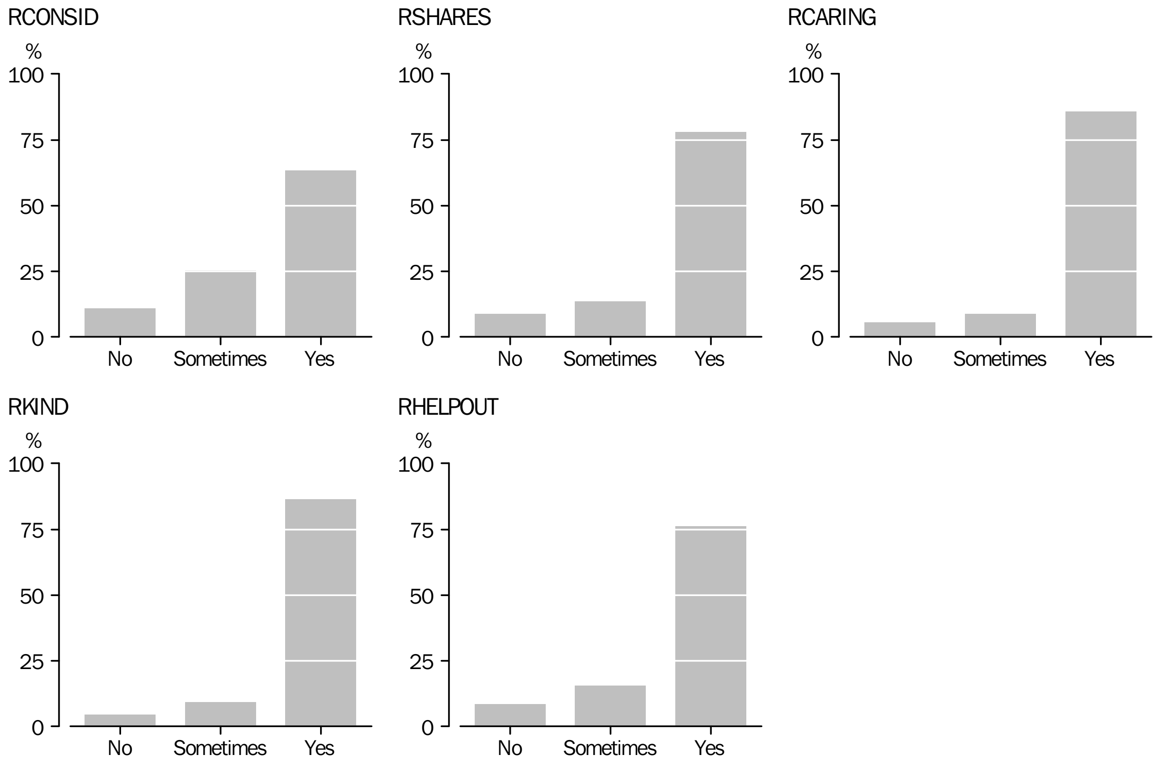
B.3 Hyperactivity scale



B.4 Peer problems scale



B.5 Prosocial skills scale



B.2 Analysis of SDQ scores

In this section, we provide frequency distributions of total SDQ scores for children aged 4 to 17 years by gender and age-group. Goodman, et al. (1998) suggest the following bandwidths to classify total SDQ score (for parent completed reports)

Normal:	0–13
Borderline:	14–16
Abnormal:	17–40

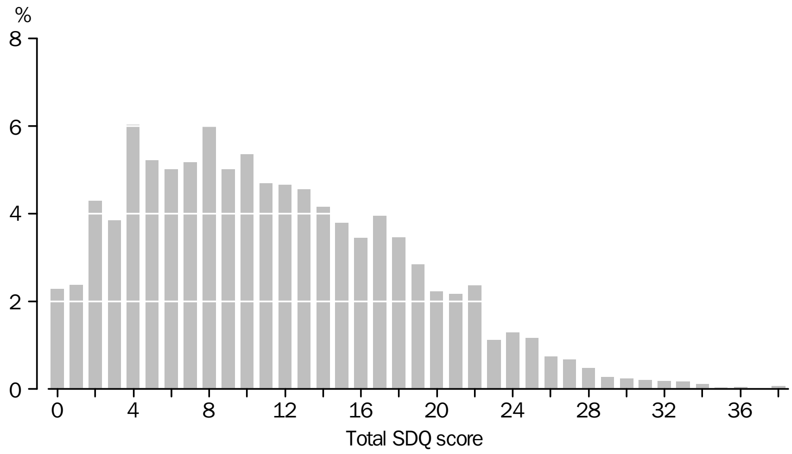
The average total SDQ for all children aged 4–17 years is 11.3, which falls into the normal classification (table B.6).

B.6 Average SDQ score by gender and age

		Total SDQ		
	Sample	*Average*	*Minimum*	*Maximum*
Males	2,014	11.9	0	8
Females	1,979	10.7	0	36
4–11 years	2,594	11.9	0	38
12–17 years	1,399	10.5	0	38
All	**3,993**	**11.3**	**0**	**38**

Figure B.7 below presents a frequency distribution of total SDQ scores for all children aged 4–17 years. We observe that nearly two-thirds of children have normal mental health. Nearly 12% of children are classified as borderline and about 24% are likely to have abnormal mental health (table B.8).

B.7 Frequency distribution of total SDQ scores – All

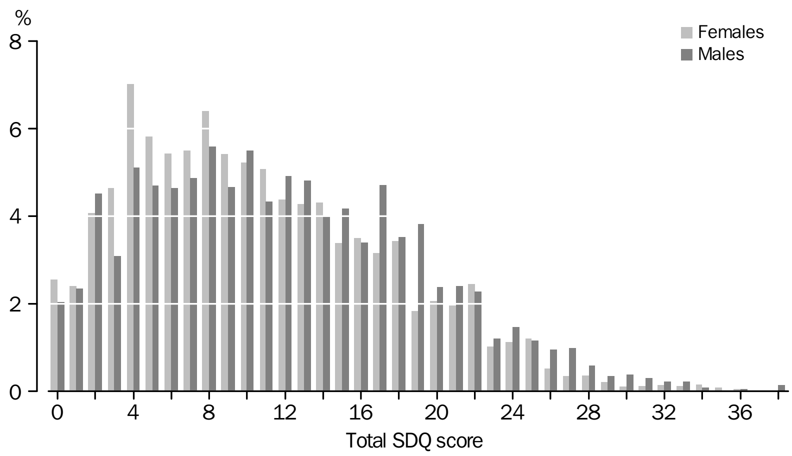


B.8 Classification of mental health by gender and age

	Normal	*Borderline*	*Abnormal*
Males	61.1%	11.6%	27.3%
Females	68.3%	11.2%	20.5%
4–11 years	61.0%	12.8%	26.3%
12–17 years	70.1%	9.4%	20.5%
All	64.6%	11.4%	24.0%

We also observe differences between males and females (figure B.9) with a higher proportion of males reported to have mental health disorders. The average total SDQ score for females is lower (10.7) compared to males (11.9). However, the average SDQ score for both sexes remains in the normal range.

B.9 Frequency distribution of total SDQ scores by gender

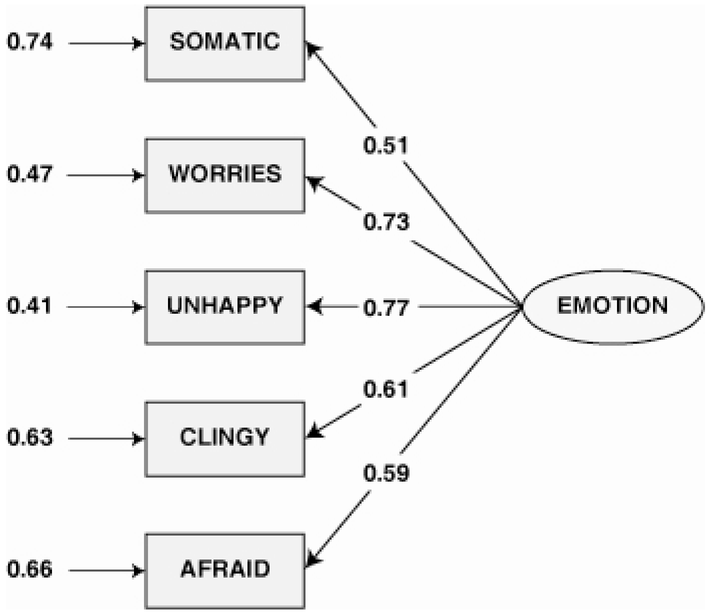


A higher proportion of females are in the normal range compared to males, similar proportions are in the borderline range and more males are in the abnormal range (table B.8). An analysis by age shows that younger age groups (4–11 years) are more likely to have mental disorders compared to older age groups (12–17 years). The average total SDQ score for both age groups remains in the normal range, however, it is higher for those aged 4–11 years (11.7) compared with those aged 12–17 years (10.5) (table B.6).

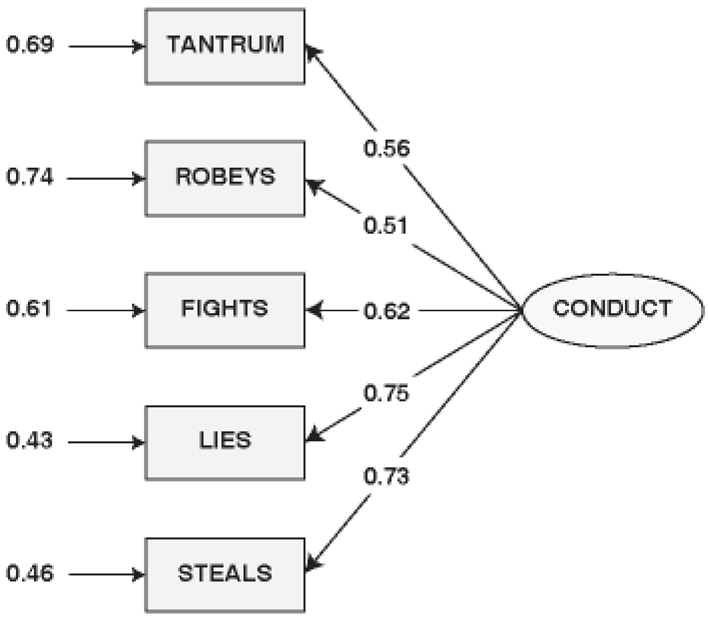
Nearly 70% of those aged 12–17 years are in the normal range compared to 61% of those aged 4–11 years. A slightly higher proportion of those aged 4–11 years are in the borderline range compared to those aged 12–17 years and also for the abnormal range.

C. SINGLE-LEVEL CONGENERIC MODEL – PATH DIAGRAMS OF THE FIVE SDQ SUBSCALES

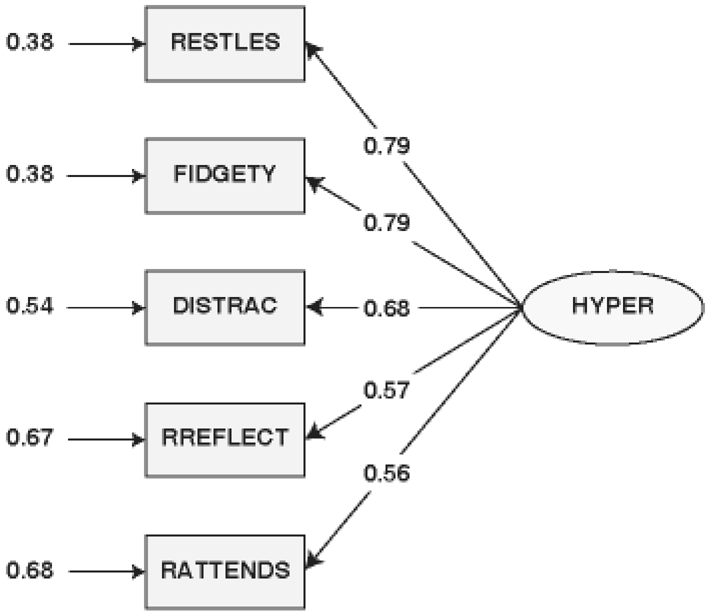
C.1 Model 1 – Emotional symptoms



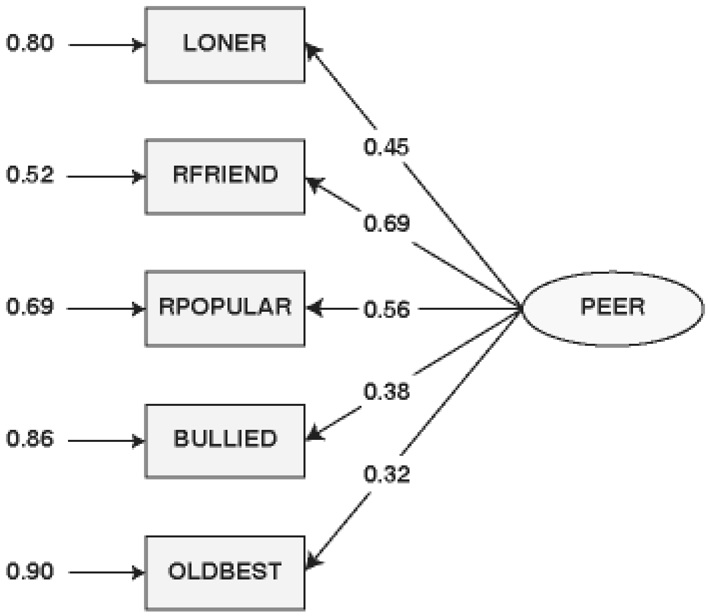
C.2 Model 2 – Conduct problems



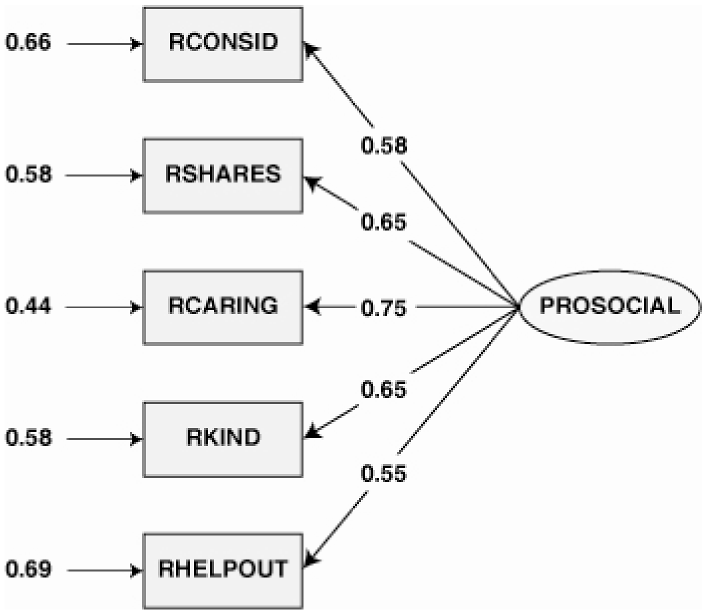
C.3 Model 3 – Hyperactivity



C.4 Model 4 – Peer problems



C.5 Model 5 – Prosocial skills



D. GOODNESS-OF-FIT STATISTICS

Chi-square test

The minimum fit function chi-square reported by LISREL is a goodness- (or badness-) of-fit measure in the sense that large χ^2 values correspond to bad model fit. The degrees of freedom serve as a standard to judge whether χ^2 is large or small.

This test measures the distance (difference, discrepancy, deviance) between the sample covariance (correlation) matrix and the fitted covariance (correlation) matrix.

Among others, Joreskog and Sorbom (1989) and Bearden, Sharma and Teel (1982) both note that the χ^2 measure is sensitive to sample size. Large sample sizes and departures from normality tend to increase χ^2 over and above that can be expected due to model specification error. Hair, et al. (1998) further state that the use of the χ^2 measure is only appropriate for sample sizes between 100 and 200. It has also been shown that this measure also varies based on the number of categories in the response variable.

As our models are estimated on large sample sizes (almost 4,000 observations), we choose to use additional goodness-of-fit statistics as described below.

Goodness-of-fit Index (GFI) / Adjusted Goodness-of-fit Index (AGFI)

The goodness-of-fit index (GFI) is another overall model fit measure. It gives the proportion of variance/covariance that is explained by the model.

Another way of interpreting the GFI is the proportion of variance in the unobservable variables that is explained by the observed indicators (see Fullarton, 2002, page 7). For example, a GFI value of 0.95 suggests that the observed indicators account for around 95 % of the variance in the latent factor.

The adjusted goodness-of-fit index (AGFI) is simply calculated as the GFI adjusted for the degrees of freedom in the model.

Fergusson, et al. (2003) suggest from their experience that an acceptable fitting model has an AGFI value in excess of 0.95.

Root Mean Square Residual (RMR)

The RMR is a measure of the average of the fitted residuals. It gives the proportion of variance in the data unaccounted for by the model. Lower values indicate ‘better model fit’.

Hair, et al. (1998) suggest that, as a rule of thumb, an RMR statistic less than 0.05 indicates a good model fit.

Root Mean Square Error of Approximation (RMSEA)

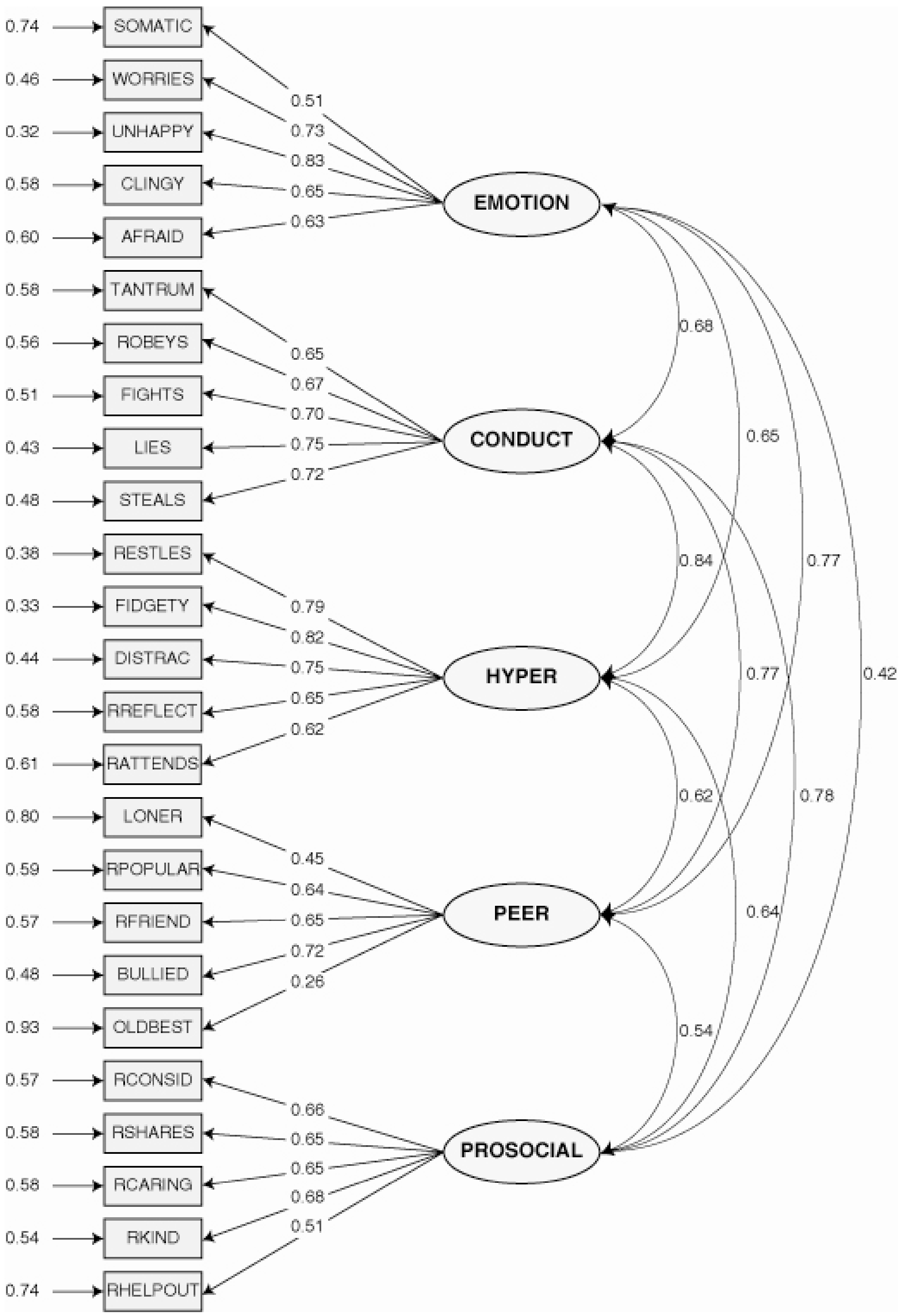
The use of chi-square as a central chi-square statistic is based on the assumption that the model holds exactly in the population. This may be an unreasonable assumption in most empirical research. A consequence is that models that hold approximately in the population will be rejected in large sample. Another fit measure that takes particular account of the error of approximation in the population is the RMSEA (LISREL help)

Brown and Cudeck (1993) suggest that a value of 0.05 indicates a close fit and values up to 0.08 represent reasonable errors of approximation in the population.

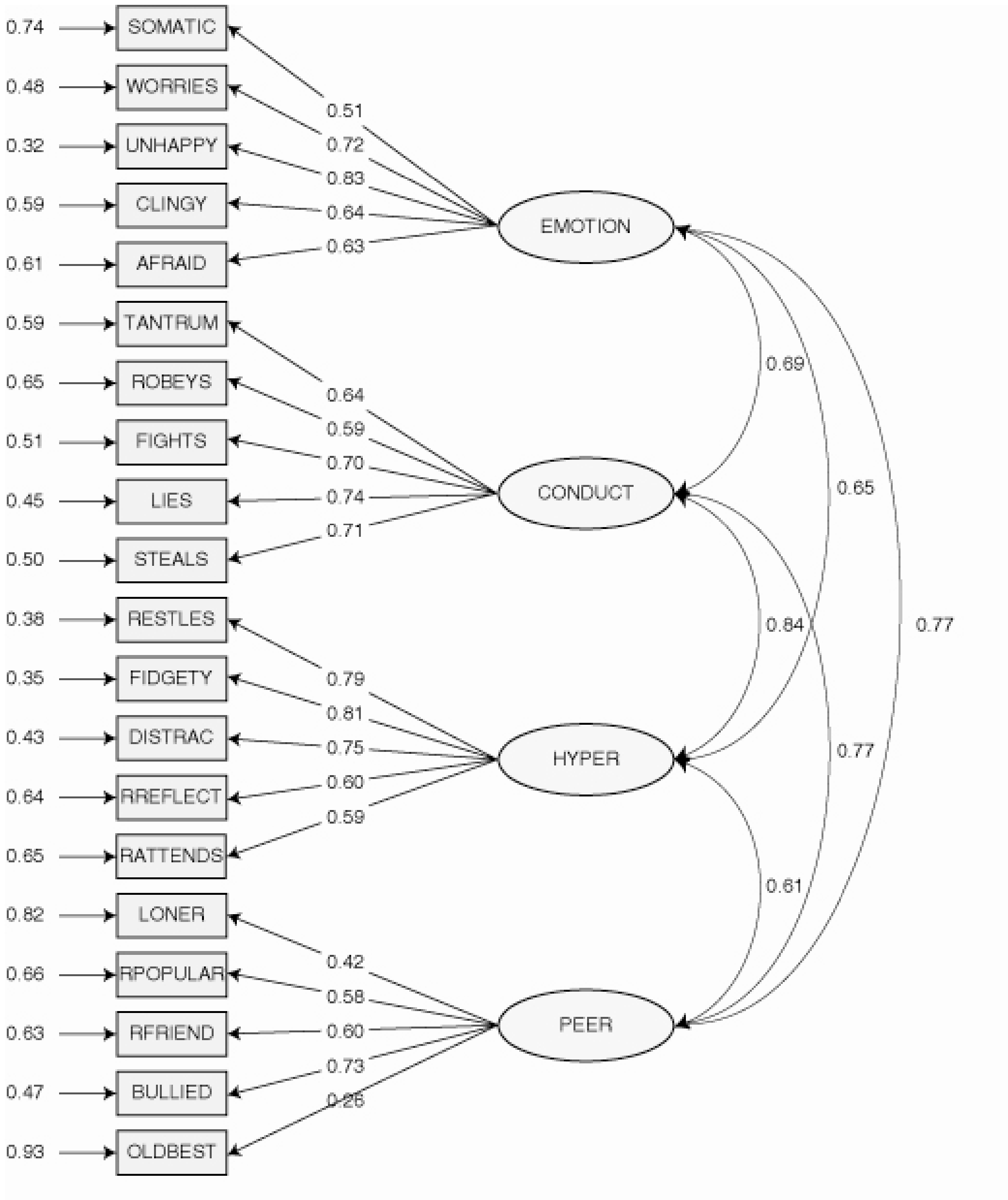
This measure also allows us to calculate the probability of obtaining the same results if a similar sample was taken from the 'super population'. For example, a RMSEA value equal to 0.0433 indicates that this 'probability' would be (100 – 4.33) ~ 96 %.

E. PATH DIAGRAMS FOR THREE MULTI-FACTOR CONGENERIC MODELS

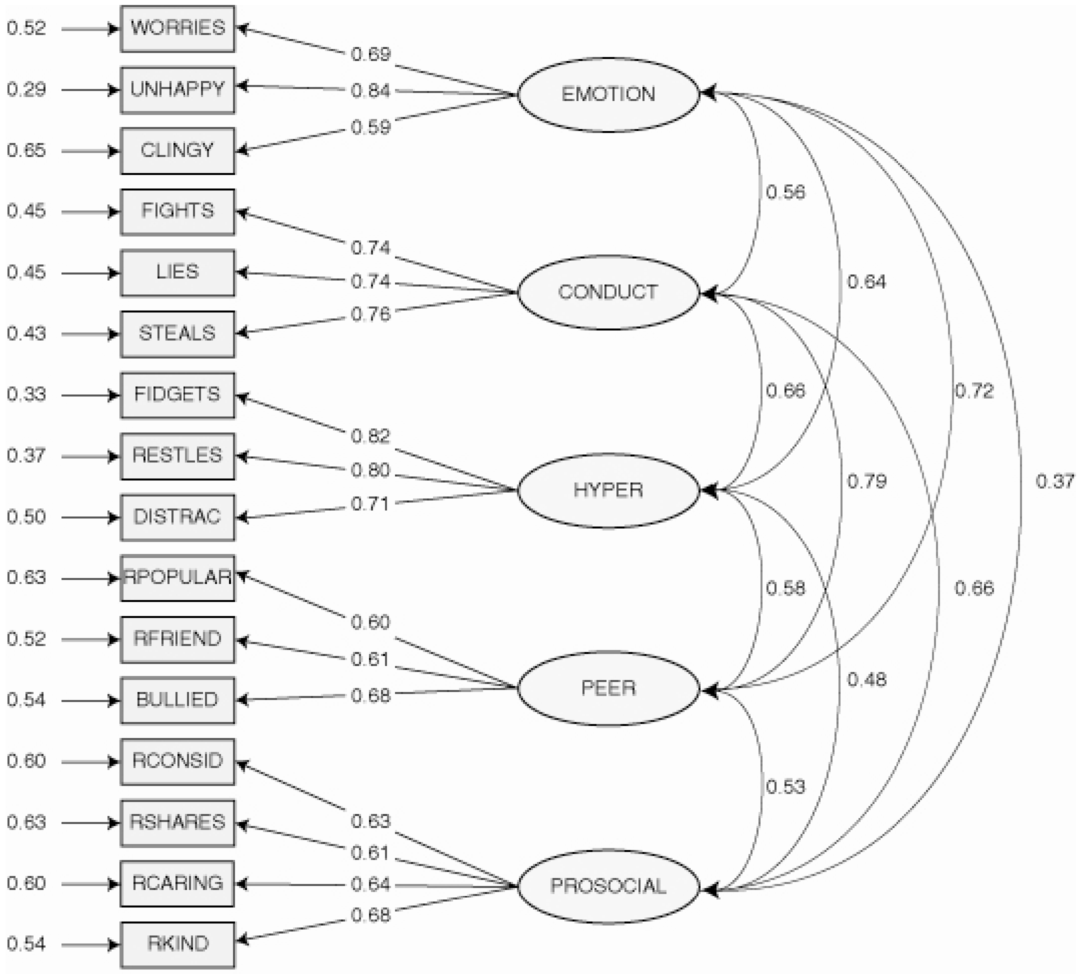
E.1 Model 1 – Five-factor congeneric model



E.2 Model 2 – Four-factor congeneric model



E.3 Model 3 – Best fit model – Five factors, 16 indicators



F. NORMALISATION OF THE COMPOSITE MEASURE

Before using the composite measure in model fitting, it is important to examine the distributional properties of the data. Rowe (2003) notes this is important as a key assumption of fitting linear models to data that contain continuous variables is that such variables are normally and independently distributed. If the normality assumption is violated, interpretations of parameter estimates and their standard errors are problematic and may be incorrect. Joreskog, et al. (1999) strongly recommend that non-normal continuous variables be normalised – especially in instances where there origins and units of measurement have no intrinsic meaning.

When we test for normality of the composite measure using the Kolmogorov–Smirnov, Cramer–von Mises and Anderson –Darling tests we conclude that the composite measure is non normal at all conventional levels of statistical significance. We then normalise the original composite scores using the NS command in PRELIS 2.50. After re-testing for normality, we conclude that the normalised composite score is normal based on the Cramer–Von Mises test (at 1% level). The normalised total composite is then used to fit the multi-level models discussed in Section 6.

For a full discussion of the normalisation procedure in PRELIS 2.50, please see Joreskog, et al. (2001, page 163). Here, we briefly describe the procedure:

Consider a data matrix of N cases on p variables. Consider any of these p variables to be normalized, and let

$$x_1, x_2, \ldots, x_N$$

be the sample values. Suppose there are k distinct values

$$x_1, x_2, \ldots, x_k$$

and let n_i be the frequency of the the occurrence of x_i, i.e. the number of times the value x_i occurs in the sample. Each $n_i \geq 1$ and $\sum_{i=1}^{k} n_i = N$.

The normal score z_i corresponding to variable x_i is calculated as

$$z_i = \frac{N}{n_i}\{\phi(\alpha_{i-1}) - \phi(\alpha_i)\} \qquad i=1, 2, \ldots, k$$

where $\alpha_0 = -\infty$, $\alpha_k = +\infty$, and

$$\alpha_i = \Phi^{-1}\left(\sum_{j=1}^{i} \frac{n_j}{N}\right) \qquad i=1, 2, \ldots, k-1$$

Here ϕ is the standard normal density function and Φ^{-1} is the inverse standard normal distribution function. PRELIS scales the normal scores so that they have the same sample mean and standard deviation as the original variable. Thus the normal score is a monotonic transformation of the original score with the same mean and standard deviation but with much reduced skewness and kurtosis.

FOR MORE INFORMATION...

INTERNET **www.abs.gov.au** the ABS web site is the best place for data from our publications and information about the ABS.

LIBRARY A range of ABS publications is available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our web site for a list of libraries..

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our web site, or purchase a hard copy publication. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice..

PHONE 1300 135 070

EMAIL client.services@abs.gov.au

FAX 1300 135 211

POST Client Services, ABS, GPO Box 796, Sydney 2001

FREE ACCESS TO PUBLICATIONS

All ABS statistics can be downloaded free of charge from the ABS web site.

WEB ADDRESS www.abs.gov.au

2000001527719
ISBN 0 642 48206 3

RRP $11.00

Produced by the Australian Bureau of Statistics