2013

Improving Evaluations of Anti-Bullying Programs in Schools

Thérèse Shaw
Edith Cowan University

2013

Improving Evaluations of Anti-Bullying Programs in Schools

Thérèse Shaw
Edith Cowan University

Edith Cowan University

Copyright Warning

Improving Evaluations of Anti-Bullying Programs in Schools

Thérèse Shaw

BSc, BSc (Hons), MSc (Statistics)

A thesis submitted for the degree of Doctor of Philosophy

at the Child Health Promotion Research Centre,

School of Exercise and Health Sciences,

Faculty of Computing, Health and Science,

Edith Cowan University

JUNE 2013

Statements

Declaration

I certify that this thesis does not, to the best of my knowledge and belief:

(i) Incorporate without acknowledgment any material previously submitted for a degree or diploma in any institution of higher education;
(ii) Contain any material previously published or written by another person except where due reference is made in the text of this thesis;
(iii) Contain any defamatory material; or
(iv) Contain any data that has not been collected in a matter consistent with ethics approval.

I also grant permission for the Library at Edith Cowan University to make duplicate copies of my thesis as required.

Statement of originality

This thesis is based on data collected as part of three studies conducted by the Child Health Promotion Research Centre (CHPRC) at Edith Cowan University, namely the Supportive Schools Project (SSP), Australian Covert Bullying Prevalence Study (ACBPS) and Cyber Friendly Schools Project (CFSP). The author of this thesis was a Principal Investigator on the ACBPS and an Associate Investigator on the SSP and the CFSP.

While the data utilised here were collected prior to the commencement of this PhD research, I declare that the work contained within this thesis is substantively different to the main objectives of the original three studies. Further, I was solely responsible for the development of the research questions and conceptual framework, preparation of the variables used and analyses conducted, and

preparation of the manuscripts published in peer review journals and of the thesis of this PhD research.

Statement of contribution to jointly-published work

Research rarely happens in isolation, and I would like to recognise my supervisors as co-authors in the development and review of each of the manuscripts published as part of this thesis, as well as the other co-authors of the publications. I am the first named author on each of the four main publications and as such was responsible for the theoretical conception, literature reviews, data analyses and discussions in each.

Statement of contribution by others

Professor Donna Cross

Professor Cross assisted in the conceptualisation of my research questions, provided guidance during the conduct of the research, and assisted in reviewing and commenting on the publications and content of this thesis. She is the chief investigator of the three projects, the data from which are analysed here and was involved in the development of the scale described in one of the publications.

Winthrop Professor Stephen Zubrick

Professor Zubrick provided expert advice during the conduct of the research and assisted in reviewing and commenting on the publications and content of this thesis.

Associate Professor Stacey Waters

Associate Professor Waters assisted in reviewing and commenting on one of the publications contained in this thesis. She was also involved in the development of the scale described in the publication.

Dr Laura Thomas

Dr Thomas contributed to the early conceptualisation and review of one of the publications contained in this thesis.

Dr Julian Dooley

Dr Dooley contributed to the conceptualisation and review of one of the publications contained in this thesis, and was involved in the development of the scale described in the publication. He was a chief investigator on one of the projects, the data from which are analysed here.

Thérèse Shaw

Professor Donna Cross

Winthrop Professor Stephen Zubrick

Acknowledgements

First of all I would like to express my sincerest thanks and appreciation to Professor Donna Cross, my principal supervisor. She has always believed in me and been a tremendous source of support and encouragement. This PhD would not have happened without her and I will be eternally grateful. She is a truly exceptional and wonderful person, and I am a better person for knowing her.

I would like to thank Professor Stephen Zubrick, my co-supervisor, for his comments and guidance during my doctoral studies. His contribution has been invaluable and greatly appreciated.

Thank you to Dr Kyrre Breivik for generously sharing his knowledge and time with me. In particular, his insightful comments on my data analysis chapter and advice with the analyses assessing measurement invariance, were most helpful.

A thank you also goes to my colleagues at the Child Health Promotion Research Centre for their encouragement and good will. Their generosity of spirit is inspiring.

My family deserves special recognition. I am deeply grateful to my husband, Peter. His belief in me is unshakeable and his support unwavering. Without him I would not have undertaken these studies. I am blessed to have two wonderful children, Lauren and Barry, who are proud and supportive of their mother.

A special thanks to my friend Jan Lewis. She was delighted when I enrolled and has remained willing to provide a sounding board ever since. Our discussions and her advice have been an enormous help to me. Thank you also goes to Margaret Hall, who helped me see this achievement was possible and has been a good friend and support throughout.

List of publications relevant to the thesis

Peer Reviewed Journal Articles

Shaw, T., Dooley, J. D., Cross, D., Zubrick, S. R., & Waters, S. (In press). The Forms of Bullying Scale (FBS): Validity and reliability estimates for a measure of bullying victimization and perpetration in early adolescence. *Psychological Assessment*.

Shaw, T., Cross, D., & Zubrick, S. R. (Under review). Testing for response shift bias in evaluations of school anti-bullying programs. *Evaluation Review*.

Shaw, T., Cross, D., Zubrick, S. R., & Thomas, L. T. (Under review). Bias in student survey findings from active parental consent procedures. *British Educational Research Journal*.

Shaw, T., & Cross, D. (2012). The clustering of bullying and cyberbullying behaviour within Australian schools. *Australian Journal of Education, 56*(2), 142-162.

List of publications relevant to, but not included in the thesis

Peer Reviewed Journal Articles

Perren, S., Dooley, J., **Shaw, T.**, & Cross, D. (2010). Bullying in school and cyberspace: Associations with depressive symptoms in Swiss and Australian adolescents. *Child and Adolescent Psychiatry and Mental Health, 4*(28), 1-10.

Runions, K., & **Shaw, T.** Teacher-child relationship, child aggression and withdrawal in the development of peer victimization. *Journal of Applied Developmental Psychology, (Under review)*.

Runions, K., Vitaro, F., **Shaw, T.**, Cross, D., Hall, M., & Boivin, M. Teacher-child relationship, parenting, and growth in likelihood and severity of physical aggression in the early school years. *Merrill-Palmer Quarterly, (In press)*.

Abstract

Bullying at school is associated with negative social, psychological and academic outcomes for both the victimised student and the perpetrator. As a consequence, national strategies to address bullying have been implemented in numerous countries, and education sectors and schools have increasingly directed resources to the problem. To ensure resources are allocated to programs and strategies that will prevent and successfully respond to bullying in schools, and therefore prevent harm to students, evidence of program effectiveness is required.

Evaluations of anti-bullying programs in schools have had mixed results and there is a lack of robust evidence as to their impact. Several factors have been proposed and investigated as explanations for the varying effectiveness of programs, including the lack of strong study designs and rigorous methods in many program evaluations.

The aim of this doctoral research was to investigate methodological challenges identified in the reviews of evaluations of school anti-bullying programs and to determine the implications of these challenges for such evaluation studies. Recommendations for future methodological research practice were also developed.

Threats to the validity of findings from evaluations of anti-bullying programs, as identified in the literature reviewing these evaluations, were investigated within a theoretical framework for assessing the methodological quality of program evaluations. This research utilised existing large data sets from three studies, one national cross-sectional survey and two group-randomised controlled trials. Self-report measures of bullying victimisation and perpetration comprised the key outcomes. Various advanced statistical techniques were applied to investigate specific threats to the construct, internal and statistical validity of findings from an

evaluation. Approval for research was obtained from the relevant authorities and conducted according to the appropriate ethical standards.

This research details how the validity of findings from a program evaluation can be threatened as a result of the instruments used to measure bullying behaviours; response shift bias or raised awareness of bullying within the intervention group; active only parental consent procedures; the use of data analysis methods unsuited to bullying outcomes; and insufficient sample sizes. Strategies for minimising these threats, such as suitable approaches to sample size determination and the use of novel statistical techniques, are described.

With careful planning and sufficient resources, evaluations of bullying programs in schools can provide valid evidence of their efficacy. Apart from the usual methodological considerations, the limitations of many evaluation studies can be addressed through the use of strong study designs; suitable methods to determine sufficient sample sizes; strategies to maximise response and consent rates when recruiting subjects; appropriate instruments to measure bullying behaviours; measures to assess and account for response shift; and appropriate data analysis methods.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1 : General introduction

1.1 Introduction

The evaluation of health promotion programs is essential to determine whether they meet their objectives to improve health and prevent harm. Evidence regarding program effectiveness assists policy makers and practitioners to determine the most effective policies and programs to implement to achieve health-related outcomes and thus ensure resources are spent appropriately and with the greatest potential gain. It is critical, therefore, that programs are rigorously evaluated to maximise the validity of the research findings. The negative short-term and long-term impacts of involvement in bullying behaviours highlight the need for effective school programs to prevent and reduce bullying, and hence robust evidence to assist schools and education systems when choosing policies and programs to implement. However, the evaluation of anti-bullying programs is predicated upon the use of rigorous methods, and it is here where there are potential threats to the validity of findings from such evaluation. This study aims to investigate major methodological challenges previously not fully addressed in the conduct of evaluations of anti-bullying programs in schools which may impact on the validity of the findings, and to provide recommendations to meet these challenges.

1.2 Bullying and the consequences of bullying

Bullying is defined as aggressive behaviour which meets the three criteria of being repeated over a period of time, characterised by a real or perceived imbalance of power and perpetrated with the intent to harm the target (Olweus, 1996). As young peoples' use of technology increases, cyberbullying or bullying using the Internet or mobile phone, has become of greater concern (P. K. Smith et al., 2008). An international study of adolescents aged 11, 13 and 15 years conducted in 40 countries found 11% had bullied others, 13% were bullied and 4% were involved both as a victimised student and a perpetrator (Craig et al., 2009). Based on data from the United States and Norway, mean rates amongst students in Grades 3-12 of being cyberbullied 2-3 times a month or more often, were 3% (Olweus,

2012). Findings from the Australian Covert Bullying Prevalence Study showed just over one quarter (27%) of school students aged 8 to 14 years reported being bullied and 9% reported bullying others every few weeks or more often (Cross et al., 2009). In Australia approximately 7% of students in Years 4 to 9 reported being cyberbullied every few weeks or more often in the last term at school (Cross et al., 2009).

Bullying behaviour can take many forms. While historically bullying was primarily seen as verbal and physical, other forms such as relational and social bullying are now also recognised (Crick & Grotpeter, 1995; Monks & Smith, 2006). Verbal bullying includes both name-calling and threatening behaviours, and physical bullying includes behaviours directed at the victimized person or their property (Slonje & Smith, 2008). Relational and social bullying differ in their intent. The aim of relational bullying is damage to a person's peer relationships, e.g. through exclusion or attempts to break up friendships (Monks & Smith, 2006). Social aggression aims to damage a person's social standing, usually through spreading nasty rumours or lies about the targeted person, often through a third party (Crick & Grotpeter, 1995; Monks & Smith, 2006; Underwood, 2002). Male students are more likely to report physical and verbal aggression toward others, while female students report higher rates of relational and social aggression (Archer, 2004; Crick & Grotpeter, 1995; Spears, Slee, Owens, Johnson, & Campbell, 2008).

Bullying between students at school has been associated with negative social, physical and psychological outcomes as well as lower academic performance for both the perpetrators and those who are victimised (Arseneault, Bowes, & Shakoor, 2010; Card & Hodges, 2008; Hawker & Boulton, 2000; Nansel et al., 2001; Rothon, Head, Klineberg, & Stansfeld, 2011; Wolke, Woods, Bloomfield, & Karstadt, 2001). In particular, victimised students have diminished self-concept and elevated levels of depression, anxiety and suicidal ideations (Arseneault et al., 2010; Card & Hodges, 2008; Hawker & Boulton, 2000; Rothon et al., 2011). Further, bullying is an independent contributor to mental health problems, over and above other factors such as prior mental health symptoms and family factors, and the effects on mental health extend beyond the immediate negative impacts into adulthood (Arseneault et al., 2010). The social health of students is also impacted – victimised students have lower social status and are socially isolated (Arseneault et al., 2010; Card & Hodges, 2008; Hawker & Boulton, 2000; Juvonen, Graham, & Schuster, 2003; Nansel et al., 2001). They are more likely to dislike school, have increased absenteeism, and may do less well academically (Card & Hodges, 2008; Rothon et al., 2011). Students who bully others

are more likely to be involved in other conduct problems and less likely to behave pro-socially (Juvonen et al., 2003; Nansel et al., 2001; Wolke et al., 2001). They also perform less well academically (Nansel et al., 2001) and are more likely to be involved in violent or aggressive behaviour in later life (Ttofi, Farrington, & Lösel, 2012). The serious negative impacts from involvement in bullying, both for the target and the perpetrator, highlight the need for effective school programs to address bullying behaviours and thereby improve the well-being of children and adolescents (Arseneault et al., 2010).

1.3 Evaluations of anti-bullying programs in schools

Concerns about the consequences of bullying have led education sectors and schools to increasingly direct resources to addressing the problem. National and regional authorities have developed strategies, such as the national strategies in Austria (Spiel & Strohmeier, 2011), Australia (Cross et al., 2011) and Finland (Salmivalli, Kärnä, & Poskiparta, 2011), or legislated policies and/or programs be implemented in schools to address bullying (Limber & Small, 2003). In some instances school funding is tied to the use of evidence-based programs (Ryan & Smith, 2009). It is critical then, that valid data from empirical investigations be available to assist educators choose policies and programs which will most benefit school communities and students.

The results describing program impact on bullying behaviours are mixed. Findings from individual program evaluations have ranged from positive (e.g. Kärnä et al., 2011; Olweus & Limber, 2010) to negative effects (e.g. Roland, 1993). Reviews of school-based programs targeting bullying reported a mix of positive, null and negative results in the included studies (Baldry & Farrington, 2007; Farrington & Ttofi, 2009; Ferguson, San Miguel, Kilburn, & Sanchez, 2007; Merrell, Gueldner, Ross, & Isava, 2008; J. D. Smith, Schneider, Smith, & Ananiadou, 2004; P. K. Smith, Ananiadou, & Cowie, 2003; Ttofi & Farrington, 2011; Vreeman & Carroll, 2007). Findings from two of the three meta-analyses conducted to date in which aggregated effects were calculated, estimated an average relative reduction of 20-23% in bullying and 17-20% in victimisation in experimental versus control schools (Farrington & Ttofi, 2009; Ttofi & Farrington, 2011). However, a significant but “very small” positive effect ($r = .12$) was calculated in the other meta-analysis (Ferguson et al., 2007). Considering the body of knowledge across studies and programs, evidence on the efficacy of anti-bullying programs in schools is considered inconclusive (Ryan & Smith, 2009) or

indicative of only overall small positive effects on behavioural outcomes (Ferguson et al., 2007; Merrell et al., 2008; J. D. Smith et al., 2004; P. K. Smith et al., 2003).

A range of factors have been investigated as possible reasons for the mixed findings from program evaluations (Hahn Fox, Farrington, & Ttofi, 2012; Jimerson & Huai, 2010; P. K. Smith, 2011). These can be broadly grouped as factors related to 1) the program which may directly influence program effectiveness and, 2) methodological issues that impact on the quality of the evaluations. The major program-related and methodological issues raised in reviews of anti-bullying program evaluations (Baldry & Farrington, 2007; Farrington & Ttofi, 2009; Hahn Fox et al., 2012; Jimerson & Huai, 2010; Merrell et al., 2008; Ryan & Smith, 2009; J. D. Smith et al., 2004; P. K. Smith et al., 2003; Vreeman & Carroll, 2007) are summarised in Table 1-1 and discussed below.

Table 1-1. Summary of major methodological issues raised in reviews as influencing the outcomes of program evaluations

Program-related	Program characteristics – rationale, content, intensity and duration Program monitoring and fidelity of implementation, context within which program is implemented
Evaluation methods	Study design (i.e. experimental or quasi-experimental) Bullying measures (e.g. data source, referent time period) Response shift (i.e. sensitisation or raised knowledge increasing reports of bullying post-program) Sample size (e.g. numbers of schools assigned to study group/treatment condition) Data analysis methods (e.g. inappropriate statistical models applied to school-based data)

With regard to the first group of factors, in the first instance, the rationale, content and delivery of the program is of importance – the intervention model used, the number of components to the program, its duration and intensity (Merrell et al., 2008; P. K. Smith et al., 2003; Ttofi & Farrington, 2011; Vreeman & Carroll, 2007). For example, whole-school

programs seemed more likely to have resulted in behaviour change than programs comprised of only classroom curriculum or social and behavioural skills group training (Vreeman & Carroll, 2007). Additionally, specific program components seem to be more effective than others (Ttofi & Farrington, 2011), although caution has been advised in terms of drawing definitive conclusions or making firm policy recommendations without further research (P. K. Smith, Salmivalli, & Cowie, 2012). Levels and integrity of program implementation also influence outcomes (Jimerson & Huai, 2010; J. D. Smith et al., 2004; P. K. Smith, 2011; P. K. Smith et al., 2003; Vreeman & Carroll, 2007), and the importance of rigorous monitoring of program implementation has been highlighted (Merrell et al., 2008; Ryan & Smith, 2009; J. D. Smith et al., 2004; P. K. Smith et al., 2003). The tension between modifying programs to be culturally relevant and suited to local conditions, while still maintaining fidelity to the original program (found to be effective within a certain context), has been noted (e.g. J. D. Smith et al., 2004). Other factors which may impact on program effectiveness include the age and gender of the children involved, teacher training, effort invested by schools, levels of support provided to schools and other contextual influences (Jimerson & Huai, 2010; P. K. Smith, 2011; P. K. Smith et al., 2003; Ttofi & Farrington, 2011).

With regard to the second group of factors, estimates of program impact also differ according to the study design and methods used in the evaluation (Farrington & Ttofi, 2009; Merrell et al., 2008; J. D. Smith et al., 2004). For example, positive results were more often found in uncontrolled than controlled studies (J. D. Smith et al., 2004) and the outcome measure used was found to influence the sizes of the estimated effects (Farrington & Ttofi, 2009; Merrell et al., 2008). More importantly, reviewers of evaluations of school-based anti-bullying programs highlighted the methodological limitations of many of the studies, which not only contributed to the varying results, but reduced confidence in the validity of the findings (Baldry & Farrington, 2007; Farrington & Ttofi, 2009; Merrell et al., 2008; J. D. Smith et al., 2004). Baldry & Farrington (2007) concluded that "future evaluations should use stronger research designs with greater attention to methodological quality" (p.201).

The concerns regarding the methods employed in anti-bullying program evaluations were confirmed in a study assessing the rigour of peer-reviewed reports of evaluations conducted between 1997 and 2007 (Ryan & Smith, 2009). Selecting features most relevant to anti-bullying programs from the Standards for Efficacy, Effectiveness and Dissemination produced by the Society for Prevention Research (Flay et al., 2005), the authors developed

three sets of criteria to code program evaluations as an efficacy, effectiveness and dissemination study respectively (Ryan & Smith, 2009). For example, the criteria for an efficacy study were: a detailed program description; measures of relevant behavioural outcomes; a control condition; adequate procedures such as randomisation for group assignment; psychometrically sound instruments, long-term follow-up on outcomes. To be coded as an effectiveness or dissemination trial, additional criteria to those of an efficacy study needed to be met. Many shortcomings were identified in the evaluations. Of the 31 evaluations reviewed, only one was coded as an efficacy study and two as effectiveness studies, while the remaining 28 evaluations were classified as pilot studies as they did not meet the criteria for any of the types of studies (Ryan & Smith, 2009).

The major methodological concerns commonly raised by reviewers were the research designs and outcome measures used, others related to sample size, data analysis methods, and sensitisation due to program exposure. In terms of study design, reviewers noted many studies did not employ an experimental design, i.e. did not include a control group or comparison data, and the units (usually schools) were often not randomly assigned to study groups (Baldry & Farrington, 2007; Merrell et al., 2008; Ryan & Smith, 2009; J. D. Smith et al., 2004; Ttofi & Farrington, 2011). These design weaknesses threaten the internal validity of the studies. Whilst randomised controlled trials, or in the context of school-based programs, group-randomised controlled trials are considered to provide the strongest evidence, this is only the case when sufficient numbers of schools are assigned to study groups (Hahn Fox et al., 2012). The strength of an experimental study design lies in the comparability of the intervention and control groups, this is difficult to achieve and assess when each group comprises only a small number of schools. Hence, apart from weaknesses in the designs employed in the evaluations, determination of an adequate sample size was lacking in many studies.

Given that the ultimate aim of anti-bullying programs is to reduce bullying perpetration and victimisation, involvement in these behaviours are the primary outcomes for evaluations of such programs. Although many of the reviews only included studies which analysed the most commonly used means of measuring bullying perpetration and victimisation, namely self-reported involvement, issues around the consistent and valid measurement of bullying were raised as requiring consensus in the research community (Jimerson & Huai, 2010; Ryan & Smith, 2009; J. D. Smith et al., 2004; Ttofi & Farrington, 2011). Ttofi & Farrington (2011) called for research into “the best methods of measuring bullying, on what time

periods to enquire about and on seasonal variations". The shortcomings of many of the measures used to evaluate program impact were also highlighted, with a mismatch seen between the wording of the items and the definition of bullying according to the criteria of repetition, intent and a power imbalance (Jimerson & Huai, 2010). Limited data are available on the psychometric properties of many bullying instruments (Cornell & Bandyopadhyay, 2010; Felix, Sharkey, Green, Furlong, & Tanigawa, 2011). A common recommendation made by reviewers was the use of multiple data sources and the supplementation of self-report with data from, for example, peers or teachers (Baldry & Farrington, 2007; Jimerson & Huai, 2010; Ryan & Smith, 2009; J. D. Smith et al., 2004; P. K. Smith et al., 2003).

Another issue in the estimation of program effects and a possible source put forward by some reviewers for small program effects, was the potential for the program to "sensitise" participants and thus impact on reporting of bullying behaviours (Merrell et al., 2008; P. K. Smith et al., 2003; Spiel & Strohmeier, 2011). Increased knowledge or ability to recognise bullying may actually cause subjects to report bullying behaviours *differently* before and after program implementation (Merrell et al., 2008). It is hypothesised that sensitisation or raised awareness would lead to higher rates of reporting after program implementation than may otherwise have occurred. This shift in the manner in which students may respond to measures of bullying behaviours post-program, as opposed to actual shifts in behaviours, is known as response shift bias. Such a response shift toward higher rates of reporting post-program would result in an underestimate of the program effect. While certain reviewers (Merrell et al., 2008; J. D. Smith et al., 2004; P. K. Smith et al., 2003), and indeed some evaluators (Nixon & Werner, 2010; Orpinas et al., 2000), raised the possibility that sensitisation may have biased the estimates of program impact, no studies that directly investigated the phenomenon of response shift in bullying-related measures could be identified to date. Thus, this potential source of bias in program evaluations has not as yet been explored or tested.

The final methodological issue to be considered here for anti-bullying program evaluations is the statistical methods used. Whilst rigorous designs and instruments with sound measurement properties are needed to validly estimate program effects, so too are statistical techniques which are appropriate to the properties and distribution of the collected data. Evaluations of school-based anti-bullying programs require advanced methods of analysis, such as hierarchical or multilevel models which appropriately deal

with the dependencies in the data flowing from the fact that the programs are implemented within a school context (Ryan & Smith, 2009). Ignoring the nesting of students within schools and classes can lead to too liberal significance tests and Type I error rates above the nominal level of 5% (Murray, 1998). Only five of the 31 studies reviewed by Ryan & Smith (2009) used multilevel modelling techniques in their analyses.

Robust evidence is required to assist schools and education systems choose effective policies and programs to implement. The evidence for school-based anti-bullying programs is mixed or indicative of small positive effects. The differences in findings regarding program impact from evaluations of these programs may be partly attributable to differences and shortcomings in methodology, rather than in program content and implementation. The lack of rigour is seen by some as precluding conclusive inferences regarding the effectiveness of anti-bullying programs in schools (Ryan & Smith, 2009).The methodological shortcomings of many of the evaluations of such programs have led to calls for high-quality studies, the findings from which truly reflect the program impacts (Baldry & Farrington, 2007; Merrell et al., 2008; Ryan & Smith, 2009). The need to develop an accreditation system for anti-bullying programs which will assist educators in programming decisions has been raised (Hahn Fox et al., 2012).

In light of the call for the use of more methodological rigour in evaluations, a key aim of this thesis is to add to the body of knowledge by investigating methodological issues identified in the reviews of school-based anti-bullying evaluations which are currently under debate or not as yet well addressed in the literature.

1.4 Framework for assessing validity of evaluation studies

Evaluations of health promotion programs, such as anti-bullying programs in schools, require the application of rigorous methods to ensure the validity of the findings, i.e. that the estimate of the program effect obtained in the study is an accurate reflection of the change (or lack of change) that has occurred as a consequence of the implementation of the program. Fortunately, there has been considerable work put forward to address this.

A framework for assessing the methodological quality of evaluation studies was given by Cook & Campbell (1979), modified by Shadish, Cook & Campbell (2002) and extended for evaluation research by Farrington (2003). Farrington & Ttofi (2009) applied the standards when conducting their detailed meta-analysis into evaluations of school-based anti-bullying

programs. Five types of validity or "methodological quality criteria" are referenced, namely construct, internal, statistical conclusion, external and descriptive validity (Farrington, 2003). The corresponding framework used in this thesis for the assessment of the methodological quality of evaluations of health promotion programs is summarised in Table 1-2.

Table 1-2. Framework for assessing the validity of findings from a program evaluation.

Construct validity: Refers to the adequacy of the operational definition and measurement of the theoretical constructs that underlie the program and the outcome variables.
Internal validity: Refers to the extent to which an observed effect on the outcome variables is free from bias and can be attributed to the program. Specific threats to internal validity include: • Selection bias • History effects • Regression to the mean • Attrition • Maturation • Testing effects • Instrumentation effects
Statistical conclusion validity: Refers to the appropriate use of statistical procedures to ensure the validity of inferences drawn regarding the association between the program and the outcome variables. Of particular relevance are: • Statistical power • Assumptions of statistical tests
External validity: Refers to the extent to which the results can be generalised beyond the conducted evaluation to different subjects, settings, times, and operational definitions of interventions and outcome variables.
Descriptive validity: Refers to the adequacy with which key features of the evaluation are presented in research reports.

Note. Based on Farrington (2003), Portney & Watkins (2000) & Windsor et al. (2003).

1.5 Types of validity relevant to anti-bullying program evaluations

Each of the types of study validity described in the framework above can be applied when evaluating anti-bullying programs in schools. The issues pertinent to program evaluations, identified as shortcomings of evaluations in the literature and those investigated in this doctoral research, are discussed below.

The reviewers of anti-bullying program evaluations recommend strong study designs are used in future studies (e.g. Baldry & Farrington, 2007; Ryan & Smith, 2009). Experimental studies, where sufficient numbers of sample units are randomly selected and randomly assigned to study groups, including a control group, with pre- and post-program data collections, provide the strongest evidence when evaluating programs, as they are best able to control for factors that may bias the results and therefore threaten the validity of the findings (Hahn Fox et al., 2012; Murray, 1998; Ryan & Smith, 2009; Windsor et al., 2003). In school-based research, programs are implemented at a school or class level necessitating randomisation of schools or classes to conditions. Group-randomised controlled trials are, therefore, preferred over other study designs when evaluating school-based programs (Murray, 1998).

The focus in this doctoral research is on threats to the validity of findings from experimental studies, specifically group-randomised controlled trials. The different types of validity are therefore, described below within the context of these study designs. Additionally, although more widely applicable, to present a more focussed examination of the issues, most of the discussion will relate to an evaluation conducted as an efficacy rather than an effectiveness or dissemination trial. Given the criteria for efficacy studies also apply to effectiveness and dissemination trials, this will not limit the generalisability of the findings to the broader study types.

Construct validity

Construct validity refers to the adequacy or validity of the generalisations from the operations used in the study, to the constructs they purport to measure or represent (Farrington, 2003; Shadish et al., 2002). In evaluations of school-based anti-bullying

programs this form of validity applies most to the dependent and independent variables utilised in the study, namely the measures of exposure to the program and bullying outcomes of victimisation and perpetration. In short, "whether the intervention really was an anti-bullying program and whether the outcome really was a measure of bullying" (Farrington & Ttofi, 2009).

In an evaluation of a specific anti-bullying program or intervention in a group-randomised controlled trial, the primary independent variable is the study group to which a student's school has been randomised, hence measurement of this variable is straightforward. Employing the preferred "intention-to-treat" approach to data analyses (Portney & Watkins, 2000), a student's group membership remains as assigned at the start of the study, even if they move to a school within a different study group during the course of the study. Of course, issues around the content of the program, and the fidelity and level of its implementation are critical factors in the interpretation of the impact of a program as estimated within an evaluation. However, exploration of these factors related to the program itself, is beyond the scope of this research.

Given bullying behaviours are the focus of the programs, these then are the theoretical constructs to be measured and analysed as dependent variables in the evaluations. While there is broad consensus on the theoretical definition of bullying, i.e. repeated aggression with intent to harm within a relationship of unequal power, the measurement of bullying is a topic under considerable debate (e.g. Cornell & Bandyopadhyay, 2010; Furlong, Sharkey, Felix, Tanigawa, & Greif-Green, 2010; Ortega et al., 2001). Thus, operational definitions of the level of involvement in bullying, as a student who is victimised or one who has bullied others, vary between studies and according to the approach seen as most valid by the research team. These discrepancies have led to calls for the development of a single instrument which produces responses which are valid measures of bullying for use in evaluation studies (Jimerson & Huai, 2010; Ryan & Smith, 2009; J. D. Smith et al., 2004; Ttofi & Farrington, 2011).

Relevant considerations are the informant used (e.g. self-, peer-, or teacher-report, observations), the referent time frame (e.g. last term, last year), whether a definition of bullying is provided or not and indeed whether the term "bullied" is used at all (Furlong et al., 2010; Ortega et al., 2001; Solberg & Olweus, 2003; Ybarra, Boyd, Korchmaros, & Oppenheim, 2012). Also relevant is whether a single global question or a multi-item scale of different types of bullying behaviours, is to be used. As an example, bullying may be

operationalised as self-reported frequency of involvement in bullying within the last term at school, in response to a global question preceded by a definition of bullying. At issue is whether the chosen operationalisation is a valid measure of the theoretical construct "bullying behaviour" referred to within the research question and hence the target of the program. Aspects of the measurement of bullying behaviours are investigated in this doctoral research.

Construct validity though, not only pertains to the adequacy of the operational definitions of these variables as measures of the theoretical constructs of interest, but also the time frames within those definitions (Portney & Watkins, 2000). The timing of the implementation of program components and collection of outcome data, influence the inferences that can be drawn from the study regarding, for example, short-term and long-term program impact.

Internal validity

Internal validity concerns the extent to which the estimated program effect is free from bias. As listed in Table 1-2, a number of sources can potentially bias the attribution of program impact. In addition, the threats may interact to create bias. The inclusion of a control group in an evaluation trial, however, can protect against many of these threats to validity. As long as changes due to effects which occur during the course of the trial, such as maturation, testing, instrumentation or attrition, occur in both the intervention and control groups to the same extent, i.e. are not differential, program impact can still largely be estimated without bias (Murray, 1998). Two potential sources of bias related to selection and history effects, to which evaluations including control groups may still be vulnerable, will be considered in this doctoral research and are detailed below.

Selection bias occurs when the studied sample is skewed and does not represent the population of interest. A common source of selection bias is non-response, but in studies of minors, low parental consent rates resulting from a requirement of active only consent procedures, may also lead to low participation rates and biased samples (e.g. Courser, Shamblen, Lavrakas, Collins, & Ditterline, 2009; Tigges, 2003; Unger et al., 2004). These biased samples in which particular groups of students may be under-represented, may in turn result in biased estimates of program effects. No studies of the impact of requiring active only parental consent procedures, i.e. where student participation is only permitted when their parent returns a signed form indicating their consent, on findings from evaluations of school-based anti-bullying programs were identified in the literature.

Historical biases can be a consequence of unplanned events, both internal and external events to the study, as well as planned events such as the program itself. Most relevant in this context is response shift bias or sensitisation, raised as a plausible alternate explanation of study findings in the reviews of program evaluations (Merrell et al., 2008; J. D. Smith et al., 2004; P. K. Smith et al., 2003). Response shift bias or shifts in intervention group students' conceptualisation of bullying behaviour as a direct consequence of the program, may change their reporting of involvement in bullying, irrespective of any actual behavioural change that may have occurred. In the absence of exposure to the program, a corresponding shift in reporting will not occur amongst the control group students. Any resultant differences between intervention and control groups may be misinterpreted as behavioural change. Thus, programs may be seen as more effective, less effective and possibly even harmful rather than ascertaining the actual effect on students' behaviour.

A further threat to internal validity, not addressed in detail in this doctoral research, is regression to the mean. Regression to the mean occurs when, usually by design, the sample comprises a select group in terms of the outcome variables (Farrington, 2003). For example, if a program was tested in a group of schools where the students were identified as at high risk of involvement in bullying behaviours compared to in the general student population. In studies where the schools are randomly selected (or the population is of a size that all schools are approached to participate) and participation rates are high, regression to the mean is unlikely to be an issue. However, this effect can also occur in controlled trials if the schools are not randomised to study groups or insufficient numbers are randomised, and as a consequence, one of the intervention or control groups comprises a biased group of students e.g. those at higher risk. Regression to the mean must be considered as an alternate explanation for effects found in statistical analyses, under the above conditions.

Statistical conclusion validity

Statistical conclusion validity is the validity of the inferences drawn from the statistical analyses conducted in the evaluation. Statistical considerations most relevant to evaluation studies are the sample size and the application of appropriate statistical methods for the data analyses (Farrington, 2003). Both of these issues will be investigated in detail in this doctoral research.

In evaluations of school-based programs implemented at the school level, insufficient numbers of schools lead to an inability to assess program effects, and insufficient numbers

of participants to reduced power to detect program effects. In school-based studies which are clustered samples, power calculations need to account for the design effects or the studies will be underpowered (Heeringa, West, & Berglund, 2010; Murray & Short, 1997). These calculations require reliable estimates of the intraclass correlations from which design effects are derived. Whilst intraclass correlations are available for a range of health outcomes (e.g. Murray, Phillips, Birnbaum, & Lytle, 2001; Murray et al., 2006; Scheier, Griffin, Doyle, & Botvin, 2002), no studies reporting these correlations for bullying outcomes for school level clustering have been published to date.

The validity of results from statistical analyses depends on the extent to which the assumptions underlying the methods hold. In the first instance, the dependencies in the data through the clustering of students in schools need to be accounted for in the analyses through the use of appropriate statistical methods, such as multilevel models or robust estimation methods. The requirement for specialised methods to deal with this violation of one of the assumptions of traditional statistical techniques (e.g. Murray, 1998; Murray, Varnell, & Blitstein, 2004), is well known but not always implemented in evaluations of anti-bullying programs (Ryan & Smith, 2009). Less well known is the need to apply specialised statistical methods to account for the distinctly non-normal distributions of bullying outcomes. Since many students are not involved in bullying, either as a victimised student or a perpetrator, measures of the frequency of bullying victimisation and perpetration are highly skewed with a high percentage of values at the minimum. Assumptions of normality and homoscedasticity do not hold for these data and the application of traditional methods may lead to invalid inferences (e.g. McClendon, 1994; Muthén & Asparouhov, 2011; Osgood, Finken, & McMorris, 2002; Vittinghoff, Glidden, Shiboski, & McCulloch, 2011). The use of statistical methods appropriate to the distribution of the data is of particular importance when the effects of interest are small (Osgood et al., 2002). As this is the case for many anti-bullying programs, the imperative to use suitable methods to avoid erroneous inferences, is pertinent.

External validity

External validity of the findings from an evaluation study is the extent to which they apply or may be generalised beyond the particular group of subjects included in the evaluation, the setting within which and the time at which the evaluation is conducted (Farrington, 2003; Portney & Watkins, 2000). In school trials, this refers to whether the observed program effect is generalisable beyond the studied group of schools and students. In

addition, external validity applies to the generalisability of the findings to other interventions and outcomes (Farrington, 2003; Shadish et al., 2002). External validity will not be considered in this thesis for the following reasons.

To conclude a program is associated with a specific outcome, usually an improvement in health behaviours, alternate plausible explanations for the observed effect need to be discounted, i.e. the study needs to have internal validity. Thus, internal rather than external validity is of primary concern to evaluators (Windsor et al., 2003). Further, whilst the ability to generalise the results beyond the studied sample is important, without internal validity, the findings are questionable and therefore, their generalisability is of secondary importance.

It is usually beyond the scope of an individual trial, particularly an efficacy study, to show generalisability – the wide generalisability of findings is demonstrated in large multi-site trials (Farrington & Ttofi, 2009; Windsor et al., 2003). For example, to be able to generalise to Australian students a program would need to be trialled in multiple Australian states, school sectors and geographic settings. In addition, in evaluations of anti-bullying programs, the primary research question to be addressed is whether the particular program reduced bullying behaviours. Whilst a broader issue may be whether anti-bullying programs in general are an effective means of preventing and reducing bullying in schools, this is not the aim of an individual trial. Hence generalisability of the findings beyond the given subjects, setting, intervention and outcome variables is a secondary concern to establishing the effect of the particular program within the given context. The focus in this doctoral research is on the determination of a valid measure of the program impact, rather than the generalisability of the observed association.

Descriptive validity

To evaluate the validity of the findings from an evaluation, information regarding a number of key features of the study is required (Farrington, 2003). A list of the minimum items to be included in a research report describing an evaluation is given in Farrington (2003). The Consolidated Standards of Reporting Trials (CONSORT) statement for the reporting of group or cluster randomised trials, including a checklist and suggested diagram for inclusion in research reports, is also an excellent guide evaluators may use for reporting the methods applied in their studies (Campbell, Piaggio, Elbourne, & Altman, 2012). However, the restricted word counts for many research journals, render the reporting of all of the features of school-based health program evaluations, the program and its implementation,

as recommended, extremely difficult. This form of validity relates to reporting standards rather than the planning and conduct of an evaluation, and as such is not discussed in detail in this thesis.

1.6 Research aims

The aim of this doctoral research was to investigate methodological challenges identified in the reviews of evaluations of school anti-bullying programs and to determine the implications of these challenges for future research practice. The objectives were to explore factors, either unaddressed or with sparse information available or under scientific debate in the bullying research literature, which influence the validity of findings from evaluations of anti-bullying programs, thereby adding to the body of knowledge in this field. Specific threats to study validity to be considered were 1) the validity of measures of bullying behaviours; 2) the presence of response shift bias in self-report measures resulting from program exposure; 3) biased samples resulting from the use of active only parental consent procedures; 4) the clustering effects resulting from surveying students in schools; 5) and the use of inappropriate statistical methods for analysis of bullying outcomes which typically have highly skewed distributions.

The specific research questions addressed were:

1. What are the psychometric properties of the Forms of Bullying Scale (FBS) measuring bullying victimisation and perpetration?
2. To what extent do shifts in perceptions of bullying occur differentially in intervention and control groups in anti-bullying program trials?
3. What are the implications for bullying-related research of requiring active only parental consent versus active-passive parental consent for student participation?
4. What are the sizes of the clustering effects present in school-based studies of bullying outcomes and how should these be accounted for when designing such studies?
5. Which multivariable statistical methods are appropriate for analysing the frequency of bullying behaviours?

6. What are the implications for research practice of these major methodological challenges for research evaluating school bullying prevention and reduction programs?

The research findings for the above methodological issues are synthesised and discussed with regard to their potential impact on the validity of inference from program evaluations. Recommendations for future methodological research practice are also developed. In answering these research questions this thesis attempts to address the gaps in the current literature on the evaluation of anti-bullying school programs.

1.7 Summary of methods

Research strategy

A review of the literature on evaluations of anti-bullying programs in schools was conducted as well as reviews of the literature relevant to each of the methodological issues considered in this doctoral thesis. Data from existing projects were utilised and a range of advanced statistical methods applied to explore the research questions. A suitable framework for the assessment of the methodological rigour of program evaluations was also identified in a review of the literature, and this framework was utilised in the synthesis of the findings from the research conducted to address the first five research questions. The implications for research practice on methods for improving the validity of inference from program evaluations are discussed, together with recommendations for future practice.

Data analysed

The research questions were assessed using data from the Australian Covert Bullying Prevalence Study (ACBPS; Cross et al., 2009), the Cyber Friendly Schools Project (CFSP; CHPRC, 2010) and the Supportive Schools Project (SSP; Waters, Epstein, Cross, & Shaw, 2008) conducted by the Child Health Promotion Research Centre (CHPRC) at Edith Cowan University.

The CFSP was a group-randomised intervention trial conducted in Western Australia. Schools were recruited from the pool of all secondary non-government schools in the Perth metropolitan area with a total Year 8 student population greater than 90 students. In total, 36 of the 53 eligible schools were recruited to the study (68%). Schools were randomised to

Phase 1 (intervention in 2010/2011) and Phase 2 (intervention in 2013) groups. Pre-program data were collected in 2010 from 3,496 Year 8 students in 36 schools (87% of all the Year 8 students). Data from one school assigned to the control or Phase 2 group, which specified receipt of the program as a condition of participation, was excluded from the impact analyses. Thus, for the purposes of determining program effects, pre-program data from 3,382 Year 8 students from 35 schools collected in 2010 and post-program data from 2,813 (83%) of these students in 2011 were analysed. The questionnaires were administered primarily electronically in classrooms/computer laboratories by trained personnel from the CHPRC, using strict procedural and verbal protocols during class time.

The SSP was a cluster-randomised intervention trial conducted in 21 Catholic secondary schools (72% of those eligible). Schools were randomised to intervention and control conditions. Longitudinal data used from this project were collected from 2,739 students tracked from the beginning of Year 8 to the end of Year 9 (2006-2007). This constituted 81% of the baseline Year 8 cohort, losses were due to parental non-consent and drop-out. Trained CHPRC personnel administered the student questionnaires according to strict procedural and verbal protocols and students completed the hard copy questionnaires during class time.

The ACBPS constitutes cross-sectional data collected from 106 schools (46% response rate) and 7,418 students (84% of those with parental consent) in Years 4 to 9 (typically 9-15 years of age) in 2007 across all the States and Territories of Australia. The sampling population included all schools in Australia other than non-mainstream schools, those in remote areas and schools with less than 30 students in 2007 in each of the sampled year levels. Schools were sampled according to a stratified cluster sampling scheme. Students completed hard copy questionnaires during class time, the questionnaires were administered by school staff according to a strict verbal and procedural protocol provided by the CHPRC.

The pre-program data from 36 schools in the CFSP trial was utilised to answer Research Question 1 (testing the Forms of Bullying Scale) and Question 3 (exploring impacts of active parental consent procedures), and the pre- and post-program data from 35 schools for Research Question 2 (concerning response shift bias). Due to the statistical requirement for a large number of schools to address Research Question 4, the ACBPS data were analysed for this research question regarding intraclass correlations and sample size determination.

Research Question 5 was addressed through varied analyses of the longitudinal data from the SSP.

Measures

The primary outcome measures were single item questions and multi-item scales measuring bullying victimisation and perpetration. These items were developed by the CHPRC, based on the Olweus Bully/Victim Questionnaire (Olweus, 1996) and the Peer Relations Questionnaire (Rigby, 1998). These measures are described in Chapter 2.

Predictor variables considered in this research included student gender and as appropriate to the specific research questions:

- mental health outcomes and peer support (Research Question 1);
- study group (intervention vs. control) (Research Question 2);
- parental consent status (active only vs. active-passive), demographic variables, mental health and social outcomes, as well as school sector and school type (Research Question 3);
- school characteristics such as school sector, geographic location, school level and school size (Research Question 4); and
- study group (intervention vs. control), and school characteristics (Research Question 5).

Details of each of these measures are given in the relevant chapter.

Data analyses

Statistical methods included the following (listed in order according to the first five research questions):

- Tests of measurement validity and reliability of the bullying victimisation and perpetration scales, such as categorical confirmatory factor analyses (including tests of measurement invariance), nonparametric tests and correlations, Cronbach's alpha's (Research Question 1, Chapter 2).
- Categorical factor analyses models to assess measurement invariance across intervention and control groups on the bullying-related scales to test for response shift bias. Additionally, logistic regression analyses to compare intervention and control students' responses to traditional and retrospective pre-test questions on bullying victimisation and perpetration. (Research Question 2, Chapter 3)

- Multivariable logistic regression analyses to test for differences according to consent status, i.e. between students with active and active-passive parental consent, on a range of outcome variables. Multivariable tobit regression analyses to test for the impact of consent status on the associations between bullying outcomes and several social-emotional variables, controlling for potential confounding variables. (Research Question 3, Chapter 4)
- The calculation of intraclass correlations (ICC's) and confidence intervals for bullying-related outcomes, as measures of school-level clustering as well as demonstrations of the use of ICC values in sample sizes calculations. (Research Question 4, Chapter 5)
- Tobit regression and two-part growth models, in addition to the more conventional binary logistic regression models, logistic and linear growth models, to test for program effects. (Research Question 5, Chapter 6)

The advanced statistical analyses were conducted in Mplus 6.0 and Stata 12.0. All models accounted for the clustering of students in schools, e.g. using robust estimates of standard errors or random effects models, as well as the non-normal nature of the bullying outcomes, e.g. using the weighted least square mean variance (WLSMV) estimator in the categorical factor analyses as appropriate for ordinal non-normal data. Missing data were dealt with using appropriate methods, such as full information maximum likelihood (FIML), or the potential for bias in the results as a consequence of the missing data was assessed. Details of the statistical analyses conducted to investigate each research question are given in the corresponding chapter.

Ethics

In all instances parental and student informed consent were obtained prior to survey completion, and ethical approval from the Human Research Ethics Committee at Edith Cowan University and the relevant school authorities. All efforts were taken to ensure the confidentiality of student responses. Students were given the option to withdraw or not complete the survey or questions on the survey, without prejudice. Non-participating students completed alternate activities as assigned by their classroom teacher. At the completion of the survey, all participating students received information regarding help they could access confidentially should the survey have raised any issues of concern for them.

1.8 Contents of the thesis

A number of limitations of evaluations of school-based anti-bullying programs were identified by reviewers of the evaluations. Certain of these methodological issues as detailed in the research questions will be investigated in this thesis, using the framework outlined in **Table 1-2** and within the context of group-randomised controlled trials.

This thesis comprises an explanatory introduction chapter; a series of four research papers and a chapter addressing five of the research questions of this doctoral research; and a general discussion and conclusions chapter addressing the final research question and synthesizing the findings. In the course of preparing this thesis, two of the four research papers have been published and two are under review. **Error! Reference source not found.** shows the relationship between each of the manuscripts and chapters to the study's research questions.

Table 1-3. Contents of the thesis relative to the research questions.

Chapter	Chapter/Publication Title	Research Question
1	General introduction	
2	The Forms of Bullying Scale (FBS): Validity and reliability estimates for a measure of bullying victimization and perpetration in adolescence [Published in *Psychological Assessment*]	1
3	Testing for response shift bias in evaluations of school anti-bullying programs [Under review in *Evaluation Review* – 2nd round]	2
4	Bias in student survey findings from active parental consent procedures [Under review in *British Educational Research Journal* – 1st round]	3
5	The clustering of bullying and cyberbullying behaviour within Australian schools [Published in *Australian Journal of Education*]	4
6	Data analysis of bullying outcomes	5
7	General discussion	6

Chapter 1: This general introduction to the thesis sets the context within which this doctoral research was conducted. The chapter includes a review of the literature on school-based anti-bullying evaluations and presents the framework used to evaluate the validity of findings from such evaluations. The criteria for assessing the quality of evaluations as they apply to anti-bullying programs are discussed, while indicating the specific threats to study validity investigated in this thesis, and the corresponding research questions are outlined. The chapter also presents a summary of the methods used in the doctoral research.

Chapter 2: Debate around the valid measurement of bullying behaviours is ongoing and the journal article included in this chapter aimed to add to the body of knowledge around the measurement of bullying. While many bullying-related instruments are available, data on their psychometric properties is scant and the characteristics of some make them difficult to implement in evaluations. The journal article describes the development and testing of a multi-item scale to measure bullying victimisation and perpetration. This journal article has been published in *Psychological Assessment*.

Chapter 3: Although several reviewers of evaluations and evaluators themselves raised sensitisation or response shift as a possible explanation for study results, no studies which specifically explored the phenomenon of response shift in bullying outcomes or tested for its presence were identified in the bullying intervention literature. Hence this potential source of bias has been largely unexplored in this context and was the focus of the research paper in Chapter 3. The scale described and tested in Chapter 2 was utilised to test for response shift. At the time of submitting this thesis, this journal article was in its second round of reviews in the journal *Evaluation Review*. The article was resubmitted with responses to the reviewers on 10 April 2013. Unfortunately the editor of the journal, who was also one of the reviewers, was unwell and passed away recently. Hence, a final decision on the article has been delayed and is pending.

Chapter 4: While the issue of sampling and selection bias was not identified by the reviewers of evaluations as an issue, the research experience in the Child Health Promotion Research Centre suggests that a requirement of active parental consent can severely impact on student participation rates. The potential for biased samples to produce biased estimates of program effects was identified and included as one of the methodological issues to investigate in this research. The findings from this exploration are presented in the research paper in this chapter. At the time of submitting this thesis, this journal article has been submitted to the *British Educational Research Journal*.

Chapter 5: Further to the issue of sampling methods, this chapter explored the determination of an adequate sample size for a school-based anti-bullying program evaluation. The sample sizes of the studies included in the reviews of anti-bullying program evaluations were often inadequate, with few schools assigned to study groups in experimental studies. The determination of the required sample size for clustered samples where the outcomes are measures of bullying perpetration and victimisation depends on reliable estimates of intraclass correlations. The journal article comprising this chapter provides estimates of these intraclass correlations and describes how sample size calculations incorporating these values, to determine an adequate sample size for a school-based evaluation study, can be conducted. The magnitude of the samples, in terms of the numbers of schools required for group-randomised controlled trials of health promotion programs to have sufficient power to accurately estimate program effects, is also explored. This journal article has been published in the *Australian Journal of Education*.

Chapter 6: The statistical techniques most appropriate for bullying outcome measures, which typically have severely skewed distributions with large numbers of observations at the minimum value, have been largely unexplored. Two methods more appropriate to these distributions, namely tobit regression and two-part growth models are described and the results from analyses utilising these methods contrasted with those from more traditional statistical techniques. While the statistical methods are not new, what is new is their application to bullying outcomes. This expository chapter describes the use and advantages of these methods using empirical data.

Chapter 7: In the last chapter the findings from the previous five chapters are incorporated into a discussion on methods to improve evaluations of anti-bullying programs in schools, with the aim of addressing the limitations of evaluations conducted to date and reducing threats to the different forms of validity outlined in the framework presented in Section 1.4.

1.9 Significance of this doctoral research

Evidence on program efficacy is reliant on the conduct of rigorous program evaluation research. The shortcomings of the evaluations of anti-bullying programs in schools conducted thus far have been highlighted in the literature, with several specific methodological issues commonly raised by the reviewers of program evaluations as

contributing to the lack of confidence in the findings of the evaluations. This doctoral research investigated aspects of these methodological challenges, most of which have not previously been explored in the context of anti-bullying programs. Application of the principles and methods discussed in this thesis will improve research practice and the methodological rigour of evaluation studies. Following the recommendations will assist in enhancing the validity of the findings of evaluation studies. The availability of sound evidence will enable the appropriate expenditure of resources and assist educators to choose programs which will be of most benefit to their school communities and students. The implementation of evidence-based policy and practice found to be effective in the management and prevention of bullying will ultimately lead to improved academic, social and mental health outcomes for young people.

Chapter 2 : The Forms of Bullying Scale (FBS): Validity and reliability estimates for a measure of bullying victimization and perpetration in early adolescence

Citation

Shaw, T., Dooley, J. D., Cross, D., Zubrick, S. R., & Waters, S. (In press). The Forms of Bullying Scale (FBS): Validity and reliability estimates for a measure of bullying victimization and perpetration in early adolescence. *Psychological Assessment*.

Date submitted: 14 June 2012

Date accepted: 15 March 2013

Contribution of authors

The candidate was responsible for the literature review, conceptual framework, data analyses and discussion for this publication. Professors Cross and Zubrick, and Dr Dooley provided expert advice during the preparation of the manuscript as well as assisted in reviewing the manuscript. Assoc Professor Waters assisted in reviewing and commenting on the manuscript. Professor Cross, Dr Dooley and Assoc Professor Waters contributed to the development of the FBS scale described and tested in the manuscript.

Relevance to the thesis

This chapter presents discussion and analyses addressing Research Question 1 of this thesis. The valid and reliable measurement of bullying is explored, in particular the psychometric properties of the Forms of Bullying Scale (FBS). The outcomes of this chapter inform the discussion on the choice of appropriate instruments to use in anti-bullying program evaluations, to enhance the construct validity of such studies. The findings also inform the other research papers presented in this thesis, in particular the scale is utilised to explore the research questions addressed in Chapters 3 and 4.

Shaw, T., Dooley, J. D., Cross, D., Zubrick, S. R., & Waters, S. The Forms of Bullying Scale (FBS): Validity and reliability estimates for a measure of bullying victimization and perpetration in early adolescence. *Psychological Assessment*. Advance online publication. doi: 10.1037/a0032955

Abstract

The study of bullying behaviour and its consequences for young people depends on valid and reliable measurement of bullying victimisation and perpetration. Whilst numerous self-report bullying-related measures have been developed, robust evidence of their psychometric properties is scant and several limitations inhibit their applicability. The Forms of Bullying Scale (FBS), with versions to measure bullying victimisation (FBS-V) and perpetration (FBS-P), was developed based on existing instruments, for use with 12-15 year old adolescents to economically yet comprehensively measure both bullying perpetration and victimisation. Measurement properties were estimated. Scale validity was tested using data from two independent studies of 3,496 Grade 8 and 783 Grade 8-10 students respectively. Construct validity of scores on the FBS was shown in confirmatory factor analysis. The factor structure was not invariant across gender. Strong associations between the FBS-V and FBS-P and separate single item bullying items demonstrated adequate concurrent validity. Correlations, in directions as expected with social-emotional outcomes (i.e., depression, anxiety, conduct problems and peer support), provided robust evidence of convergent and discriminant validity. Responses to the FBS items were found to be valid and concurrently reliable measures of self-reported frequency of bullying victimisation and perpetration, as well as being useful to measure involvement in the different forms of bullying behaviours.

Keywords

bullying, victimisation, measurement, psychometrics, adolescents

Acknowledgements

This research was supported by grants for Study 1 from the Western Australian Health Promotion Foundation (Healthway) and for Study 2 from Edith Cowan University. We thank the students and staff in participating study schools as well as colleagues within our research group who contributed to earlier versions of the scales. We also thank the researchers who provided feedback on our questionnaire: Peter Smith, Ersilia Menesini, Marilyn Campbell, and Maritta Välimäki.

2.1 Introduction

Bullying between students at school has been shown to be associated with poorer social, physical, psychological, and academic outcomes for both the perpetrators and targeted students (Arseneault, Bowes, & Shakoor, 2010; Card & Hodges, 2008; Nansel et al., 2001; Wolke, Woods, Bloomfield, & Karstadt, 2001). The study of bullying behaviour and its negative consequences for children and adolescents is reliant on the valid measurement of bullying victimisation and perpetration. While there are several instruments currently available, there are limited data available on their psychometric properties (Cornell & Bandyopadhyay, 2010; Felix, Sharkey, Green, Furlong, & Tanigawa, 2011) and their characteristics may limit their use in certain contexts. This paper describes our reasoning behind the adaptation of existing scales to develop the Forms of Bullying Scale and testing of the psychometric properties of the scale.

Bullying has been defined as intentional aggressive behaviour repeated over a period of time, where there is a power imbalance between the person being bullied and the perpetrator (Olweus, 1996). Bullying behaviour can take many forms. Historically it was seen as only repeated verbal and physical acts. Verbal bullying includes both name-calling and threatening behaviours, and physical bullying, behaviours typically directed at the victimized person and/or their property (Slonje & Smith, 2008). Other forms of bullying, such as relational and social bullying are now also recognized (Monks & Smith, 2006). Relational bullying aims to damage a person's peer relationships through exclusion or attempts to break up friendships (Monks & Smith, 2006). Similarly, social bullying aims to damage a person's social standing, usually through spreading nasty rumours or lies about the targeted person, activities often carried out by a third party (Crick & Grotpeter, 1995; Monks & Smith, 2006; Underwood, 2002).

Students involved in bullying have worse mental health outcomes than non-involved students (Kaltiala-Heino, Rimpelä, Rantanen, & Rimpelä, 2000; Wolke et al., 2001). Victimized students report higher levels of internalizing problems (Arseneault et al., 2010; Hawker & Boulton, 2000; Juvonen, Graham, & Schuster, 2003) and lowered social status. Importantly, acceptance in the peer group, having more friends, and friends able to assist and protect have been shown to be protective of victimisation (Card & Hodges, 2008; Juvonen et al., 2003; Kendrick, Jutengren, & Stattin, 2012). In comparison, those who bully others are more likely to be involved in other problem behaviours, such as externalizing

conduct problems, and less likely to engage in pro-social behaviours (Juvonen et al., 2003; Nansel et al., 2001; Wolke et al., 2001).

Bullying measurement

Although some consensus has been reached on the characteristics of bullying behaviour as outlined in the given definition and the forms bullying takes (Bovaird, 2009; Smith, del Barrio, & Tokunaga, In press), the approach to measurement of bullying is still under considerable debate. The consequences of this debate can be seen in the development of a range of instruments which use different methods to measure bullying (Furlong, Sharkey, Felix, Tanigawa, & Greif-Green, 2010). The choice of an appropriate format for an instrument to measure involvement in bullying behaviours is guided by the aim of the study and therefore the purpose of the measurement (Felix et al., 2011; Greif & Furlong, 2006). If the research aims to estimate and compare the prevalence of bullying victimisation and perpetration in general, single global questions are often utilized to categorize students as having been bullied or bullied others (Solberg & Olweus, 2003). Suitable questions and cut-off points for dichotomization of responses in accordance with students' frequency of involvement in bullying behaviours have been suggested by Solberg and Olweus (2003).

A second aim may be to estimate the prevalence of different forms of bullying behaviours (e.g., verbal, relational, etc.) and track changes over time, where a scale comprising items describing different behaviours would be more relevant. To determine associations between involvement in bullying behaviours and other variables (e.g., mental health) a multi-item scale measuring involvement is more appropriate than a single global question designed to measure prevalence (Felix et al., 2011). Composite scores, for example, mean scores, on such a scale would have greater sensitivity and variability than a binary outcome and reflect the continuous nature of the latent victimisation or perpetration variable. Although some authors have described the choice as being of one or the other, a researcher may opt to use both a global question as well as a multi-item scale in their study.

As discussed by others (e.g., Furlong et al., 2010; Solberg & Olweus, 2003; Ybarra, Boyd, Korchmaros, & Oppenheim, 2012) further considerations in instrument choice are whether self-report or report from another informant will be used; a definition of bullying will be provided and/or the term *bullied* used. In addition, choices around question wording, such as a suitable referent time period and appropriate response options, are pertinent.

Informant

Self-report assessments are most commonly used to measure bullying behaviours (Felix et al., 2011; Swearer, Siebecker, Johnsen-Frerichs, & Wang, 2010).The relative merits of self-report versus other types of assessments such as peer- and teacher-nomination have been comprehensively discussed (Cornell & Bandyopadhyay, 2010; Furlong et al., 2010; Ortega et al., 2001; Solberg & Olweus, 2003). Self-report taps into the student's perspective and is thus more likely to reflect intentionality and power imbalance (Furlong et al., 2010). It provides the opportunity for those victimized to report bullying that may not be known other than to the student victimized and the perpetrator. Peer and teacher report is problematic within secondary schools where students change classes and teachers throughout the school day, and such reports would have limited ability to accurately reflect individual students' bullying involvement (Bovaird, 2009; Espelage & Swearer, 2003). Report by third parties, particularly by teachers may also be limited to more overt than covert forms of bullying (Griffin & Gross, 2004). Among the practical advantages to self-report are the ability to quickly obtain data from large numbers of students (Ortega et al., 2001) at relatively low cost and without the ethical and consent issues related to peer-nominations and observational studies (Espelage & Swearer, 2003; Griffin & Gross, 2004).

Use of definition and term *bullied*

The use of the word *bullied* together with a definition of the construct has been endorsed by several researchers (Ortega et al., 2001; Smith, Cowie, Olafsson, & Liefooghe, 2002; Solberg & Olweus, 2003). It has been questioned by others, primarily as labelling the behaviour may lead to under-reporting (Greif & Furlong, 2006; Kert, Codding, Tryon, & Shiyko, 2010) as well as reporting based on different individual understandings of the term (Smith et al., 2002). Providing a definition aims to ensure some degree of common understanding of the phenomenon and increase the comparability of responses (Griffin & Gross, 2004; Solberg & Olweus, 2003). It also enables the researcher to illustrate the three characteristics of bullying (i.e., intention, repetition, power imbalance), and distinguish bullying from aggression between equals and playful teasing (Ortega et al., 2001; Solberg & Olweus, 2003). Ybarra and colleagues (Ybarra et al., 2012) recently assessed the impact of the use of the term and provision of a written bullying definition. They recommended the inclusion of the word *bully* in question wording for English-speaking samples in the USA, as this had resulted in the lowest rate of misclassification of students. Furthermore, providing

a definition did not appear to impact on prevalence rates as similar levels were obtained with and without its use (Ybarra et al., 2012). However, only a written definition, without pictorial examples of different types of bullying as recommended by Ortega et al. (2001), was provided and it is unclear whether providing a definition of bullying which includes illustrations would impact on prevalence rates.

Time frame and response options

The referent time frame chosen (e.g., past week, past month, last three months, last year, ever) within which students are asked to report their bullying involvement affects responses and prevalence rates (Cook, Williams, Guerra, & Kim, 2009) and there is little consistency in the periods used across studies. Ortega et al. (2001) refer to the use of the last three months whilst the widely used Olweus Bully/Victim Questionnaire (OBVQ) refers to the "past couple of months" as a period of time relevant to the school year that is less likely to be affected by memory recall (Olweus, 1996). When, as is often the case, the items measure frequency of bullying experiences, the response options used reflect the referent period (e.g., once or twice in the last couple of months for the OBVQ). If the responses are to be categorized, for example to calculate prevalence rates, the response options are chosen so they can be grouped according to the cut-off points that will be used to categorize students.

Existing scales and our research purposes

Whichever of the above approaches are chosen, the study of bullying behaviour is reliant on the valid and reliable measurement of both bullying victimisation and perpetration. Although a number of instruments have been developed to measure self-reported involvement in bullying victimisation and perpetration (Furlong et al., 2010; Hamburger, Basile, & Vivolo, 2011), none is universally recognized as the instrument of choice. Furthermore, the existing self-report multi-item scales, which we identified for use with adolescents and which specifically measure bullying behaviour (Bond, Wolfe, Tollit, Butler, & Patton, 2007; Espelage & Holt, 2001; Felix et al., 2011; Hunt, Peters, & Rapee, 2012; Mynard & Joseph, 2000; Olweus, 1996; Reynolds, 2003; Rigby, 1998), were not sufficient for the purposes of our study for several reasons. Some measure only victimisation (Felix et al., 2011; Hunt et al., 2012; Mynard & Joseph, 2000; Reynolds, 2003) or the perpetration and victimisation items differ (Espelage & Holt, 2001). Some have a large number of items (e.g., 20-30) (Hunt et al., 2012; Reynolds, 2003) making their administration difficult within a broader questionnaire such as ours, whereas others with relatively few items may be

limited in their representation of the different forms of bullying (Bond et al., 2007; Rigby, 1998). Further, some are skewed towards verbal and physical forms of bullying with fewer items measuring more indirect forms (Espelage & Holt, 2001; Olweus, 1996; Rigby, 1998). Intentionality of the bullying act is not always conveyed by the wording of the items, thus some items and scales may not adequately differentiate instances of bullying from, for example, playful interactions or fights between students (Bond et al., 2007; Espelage & Holt, 2001; Mynard & Joseph, 2000). Lastly, in contrast to the approach we take, some measures intentionally do not provide a definition of bullying and/or use the word "bullied" (Espelage & Holt, 2001; Felix et al., 2011; Hunt et al., 2012). Whilst the concurrent and predictive validity of scores on certain self-report measures, based on categorizations of the responses to the measures which identify students as "bullies" and "victims", has been demonstrated (Bond et al., 2007; Felix et al., 2011; Solberg & Olweus, 2003), until recently relatively little evidence was available regarding the construct validity of scores from multi-item scales designed to measure the continuous latent constructs of bullying victimisation and perpetration. Exceptions include the Olweus Bully/Victim Questionnaire (OBVQ; Breivik & Olweus, Under review) and the Personal Experiences Checklist (PECK; Hunt et al., 2012).

Present study

The limitations of existing self-report instruments for the succinct measurement of the different forms of bullying amongst adolescents and the general absence of robust estimates of item and scale validity for bullying measures led us to adapt scales from one of the most widely used questionnaires, the revised version of the OBVQ (Olweus, 1996), and to a lesser extent, the Peer Relations Questionnaire (PRQ; Rigby, 1998) into the Forms of Bullying Scale (FBS). The FBS, which has versions to measure victimisation (FBS-V) and perpetration (FBS-P), is a multi-item scale for use with adolescents (12-15 years), of sufficient length to comprehensively assess self-reported involvement in different forms of both bullying victimisation and perpetration without being too long to administer within a broader questionnaire.

The aim of this paper is to describe the Forms of Bullying Scale (FBS) and assess the validity of the item responses to the FBS as measures of the continuous latent variables of bullying victimisation and perpetration. Since male students are more likely to report direct forms of bullying (Archer, 2004), we were interested to also test the invariance of the factor structure of the scale across gender groups.

2.2 Method

Construction of the FBS

The items in the FBS were based on the revised version of the OBVQ (Olweus, 1996) and the PRQ (Rigby, 1998). Based on pilot work with students and teachers, items on the OBVQ and PRQ were reworded to ensure the FBS was appropriate for secondary school students, and existing items split or additional ones added to measure the different forms of bullying in a more detailed way. We also aimed to ensure the wording of each of the items reflected intent to harm (see Appendix 2-B for the two versions of the FBS).

Bullying behaviours were measured in a general sense, that is, possibly occurring both online and offline, as these two means of bullying may co-occur making it difficult for young people to report on these behaviours separately and as the wording of most items does not imply the bullying occurred offline. Consistent with others (Crick & Grotpeter, 1995; Felix et al., 2011; Monks & Smith, 2006; Solberg & Olweus, 2003; Underwood, 2002), five broad forms of bullying were defined:

- Verbal – nasty teasing and name-calling;
- Threatened – made afraid, intimidated, or made to do what others want;
- Physical – physically hurt, property damaged, or stolen;
- Relational – damage to social relationships through exclusion or having friendships broken; and
- Social – lies told, false rumours spread to damage social standing.

Verbal name-calling and teasing was distinguished from threatening behaviours to reflect the more distinct nature of the latter, namely to intimidate or manipulate another. Threatening behaviour was also seen as different from actual physical actions taken against another person or their property. Similarly, relational and social bullying were viewed as separate forms given the first is aimed at damaging relationships, whereas the second targets a person's reputation (Monks & Smith, 2006; Underwood, 2002). Bullying on the basis of gender, race or sexuality was considered within the more general forms of bullying, for example, verbal. The FBS was constructed to include two items for each of five forms to enhance the content validity of the scale, recognizing the diversity of behaviours within

each category and, since such scales are often analysed as mean scores, to give equal representation of the different forms of bullying in such mean scores. The order of the items in the scale was assigned randomly in two parts so that one item from each form appears in each half of the scale. Since bullying victimisation and perpetration can take the same forms, the two versions of the scale comprise the same items, with wording changed to reflect victimisation and perpetration as appropriate.

The FBS-V and FBS-P were placed within a broader bullying questionnaire which also included global questions measuring the bullying behaviours. As recommended by Ortega (2001), the behavioural items and global questions were preceded by a definition of bullying (see Appendix 2-A) based on that of Olweus (1996), where examples of different forms of bullying with pictures illustrating each form were included (with representations of non-cyber as well as cyber methods), as were two examples of behaviours that are not bullying. A definition of cyberbullying similar to Smith et al. (2008) was also provided (see Appendix 2-A). The series of questions on bullying victimisation preceded those on perpetration.

The term *bullied* was used in the question stem and a definition provided to achieve some commonality of understanding of the phenomenon and to emphasize the distinctive characteristics, for example, power imbalance, of bullying which distinguish these behaviours from aggression in general. The word *bullied* was also used for pragmatic reasons. The broader questionnaire included questions measuring details of students' bullying experiences (e.g., duration, identity of perpetrator, response to being bullied) and use of the word enabled simple and clear wording for these questions. Further, use of the term in the stem of the scales aimed to achieve greater consistency between responses to the bullying scales (behavioural items) and the global questions that followed them. Apart from the use of the term *bullied* to elicit responses in line with the definition presented, terms such as *deliberately*, *to hurt*, and *nasty* were also included within the items to indicate intent. Repetition was incorporated in terms of the frequency of bullying experiences as indicated by the response options of the items.

All questions referred to the respondents' experiences in the previous school term, a period of about ten weeks. The questions were phrased to reflect that students may have been victimized by one or more perpetrators and similarly, that the bullying may have been perpetrated as an individual or within a group. The referent period of *last term* was chosen in accordance with the recommendations of others of eight (Solberg & Olweus, 2003) or 12

weeks (Ortega et al., 2001) and forms a natural period of time within which students can recall their experiences. The use of a specific period of time (one term) also enabled the comparison of changes in students' bullying experiences over time and the testing of the impact of the intervention. The response options were similar to those used by others (Felix et al., 2011; Solberg & Olweus, 2003) and were the same for all four scales: "This did not happen to me/I did not do this"; "Once or twice"; "Every few weeks"; "About once a week"; and "Several times a week or more".

The testing of the FBS followed four stages – feedback was sought from international bullying research experts; the questionnaire was piloted and focus groups conducted with students following its administration; the questionnaire was administered to 3,496 students in an initial study from non-government schools, and subsequently in a second study with 783 students from government schools.

Expert Feedback and Piloting

Feedback on our questionnaire was obtained from four international bullying research experts. Pilot testing with 50 students aged 12-14 years in two schools (not part of either sample described below) was also conducted and following completion of the online survey, focus groups were conducted with the students to assess social relevance, clarity and face validity of the items. Both consultations resulted in minimal changes to the scale.

Students in the focus groups endorsed the use of a definition of bullying together with pictographs – the examples illustrating behaviours not considered bullying were described by the students as particularly helpful. Several also mentioned the importance of the inclusion of nonphysical forms of bullying, such as exclusion, in the definition as they did not instinctively include indirect forms when thinking of bullying behaviours.

Testing of the FBS – Study One

Validity was assessed using data from 3,496 Grade 8 students from 36 schools. These data were the baseline measures for a large group-randomized controlled cyberbullying intervention trial. All non-government schools in Perth, Western Australia were approached to participate, with a response rate of 68% (reasons cited by schools for not participating were competing priorities within the school or participation in other on-going research projects). Consent was sought from parents and all Grade 8 students in each school (non-consent rate 7%) and the combined consent/response rate was 87% (non-response was due to students being absent from school on the day of survey

administration and failing to complete the survey on their return to school). Data collection was conducted in weeks 6-11 of Term 2, 2010 (8 to 13 weeks after the end of Term 1, the term students were asked to report on). Note that in Australia the school year runs from late January to December, and is divided into four terms of about 10 weeks length, separated by two week holiday periods. The mean age was 12.9 years (*SD*=.38), and 51.6% (*n*=1,798) were girls. Students completed online surveys administered by trained research staff during normal class periods, some in school computing laboratories and others on laptops in their classrooms. The surveys were not anonymous as student responses were to be tracked over time. Each student was provided a unique numeric login and assured of the confidentiality of their responses. The majority of students completed the full survey in about 25 to 35 minutes.

Testing of the FBS – Study Two

Participants were 783 students from seven government schools in Perth, Western Australia surveyed in 2011 as part of a cross-sectional study assessing students' use of technology (consent/response rate=43%, the fairly low rate was due to the prohibitive consent procedures required in government schools). Approximately half of the students were in Grade 8 (53%), 17% in Grade 9 and 30% in Grade 10. One half of the students were male (*n* = 401, 51.2%) and ranged in age from 12 to 16 years (*M* = 13.9, *SD* = .88). Similar sampling and data collection methods were utilized, and the definition with pictorial representations of bullying and the bullying scales were the same, as those for the first study.

Ethics

In all instances we obtained parental and student informed consent, and ethical approval from the Human Research Ethics Committee at Edith Cowan University and the relevant school authorities.

Analytic Plan

Nonparametric tests and methods that accounted for the highly skewed distributions of the item and scale mean scores and their ordinal nature, were used throughout the analyses.

In recognition of the manner in which scales are often analysed in practice, a mean score was used for each of the FBS-V and FBS-P when calculating the descriptive statistics and correlations. These mean scores represent a sum of the frequency and the number of different ways in which a student was bullied or bullied others, with higher scores representing greater exposure to or involvement in bullying. As is the case with almost all

bullying scales, the composite score for the victimisation version is not a measure of the level of harm or impact on the bullied student. To measure harm, each individual person needs to indicate the effect of the bullying for him or her, as two students can experience different harm from the same behaviour.

Construct validity was assessed through confirmatory factor analyses within the framework of a reflective rather than a formative measurement model (Bollen & Bauldry, 2011; Bovaird, 2009). However, we caution against the conceptualization of the latent constructs, in particular victimisation, as a trait of the targeted person, which would seem to imply something inherent in the person which is the cause of his or her victimisation. Rather, as argued by Edwards (2011), we see the causation as occurring at the time of completion of the questionnaire, where the person responds to the items on the basis of his or her experiences and self-perception of his or her status or position on the latent continuous scale. Further justification for the appropriateness of a reflective framework, was the relative invariance we found in the factor loadings when various items were dropped from the scales, indicating stability in the latent victimisation and perpetration constructs being measured.

The factor analyses were conducted in MPlus Version 6 using the weighted least square mean variance (WLSMV) estimator as appropriate for ordinal non-normal data (Muthen & Muthen, 1998-2009). Since bullying victimisation and perpetration are correlated with some students both being bullied as well as bullying others, two-factor models were fitted to the combined items from the two scales. Measures of goodness of fit used were the Root Mean Square Error of Approximation (RMSEA) and Comparative Fit Index (CFI). Recommended values for the RMSEA are less than .06 (Hu & Bentler, 1998) and .08 (Brown & Cudeck, 1993), and .95 or higher for the CFI (Hu & Bentler, 1998). The chi-square test is also reported although this test is sensitive to sample size and departures from normality.

The invariance of the factor loadings and thresholds across gender groups was assessed with a comparison of the fit of the models assuming and not assuming invariance using the CFI. Differences larger than .002 indicate a lack of measurement invariance (Meade, Johnson, & Braddy, 2008). The Satorra-Bentler scaled chi-square difference test is also reported (MPlus, 2012; Muthen & Muthen, 1998-2009), although chi-square tests are known to be too liberal with large sample sizes such as this (Joreskog & Sorbom, 1993).

Concurrent validity was assessed through Mann-Whitney tests comparing the mean scores for the FBS-V items between the groups of students categorized as having been victimized or not, based on the global single item question. A similar test was conducted for the mean of the FBS-P items.

Tests of *convergent and discriminant validity* were based on hypothesized correlations between the FBS-V and FBS-P and six measures of mental and social health outcomes described below, namely depression, anxiety, emotional symptoms, conduct problems, peer problems and peer support. Spearman correlations were obtained due to the extreme skew in the data, particularly the FBS-P mean scores.

Symptoms of depression and anxiety were assessed using the Depression Anxiety Stress Scale (S. Lovibond & P. Lovibond, 1995) which comprised seven items for each of depression, anxiety and stress respectively (the stress scale was not utilized here). The validity of DASS scores have been shown previously – DASS depression scale and Beck Depression Inventory (BDI) r = .74; DASS anxiety scale and Beck Anxiety Inventory (BAI) r = .81 (P. Lovibond & S. Lovibond, 1995), and the three factor structure confirmed in a sample of Grade 7-9 students (RMSEA = .052, CFI = .946) (Szabó, 2010). Item responses ranged from 0 (*Does not apply to me*) to 3 (*Most of the time*). For these data, good model fit was found in confirmatory factor analyses for one factor models and good internal consistency, for depression (RMSEA = .043, CFI = .993, α = .92) and for anxiety (RMSEA = .033, CFI = .978, α = .82). Mean scores were calculated, with higher scores reflecting greater symptoms of depression and anxiety.

The emotional symptoms (e.g., "I get a lot of headaches, stomach-aches, or illness"), conduct problems (e.g., "I fight a lot"; "I can make other people do what I want"), peer problems (e.g., "Other young people pick on me or bully me" and reverse coded: "Other people my age generally like me"), and pro-social behaviour (e.g., "I usually share with others") subscales of the Strengths and Difficulties Questionnaire (SDQ) were utilized to measure these four constructs (Goodman, 1997). Each subscale comprises five items with response options 1 (*Not true*), 2 (*Somewhat true*), and 3 (*Certainly true*). Means were obtained for each of the subscales, with higher scores representing greater levels of emotional symptoms (α = .73), pro-social behaviours (α = .72), conduct problems (α = .55), and peer problems (α = .54). Although the Cronbach's alpha values of the first two scales were acceptable, the values for the conduct problems and peer problems scales were below the standard of .7 in our sample. Unidimensionality of the subscales was

demonstrated in confirmatory factor analyses, for emotional symptoms (RMSEA = .046, CFI = .983), conduct problems (RMSEA = .029, CFI = .984), pro-social behaviours (RMSEA = .009, CFI = .999), and to a lesser extent, for peer problems (RMSEA = .037, CFI = .982, allowing the two positively worded items to covary). Results related to the peer problems scale, therefore, need to be interpreted in light of the less than optimal performance of the scale in this sample.

Peer support was measured using an eleven item scale adapted from the 24 item Perceptions of Peer Social Support Scale (Ladd, Kochenderfer, & Coleman, 1996) (e.g., "How often would other students invite you to do things with them"). Response options were 1 (*Lots of times*), 2 (*Sometimes*), and 3 (*Never*). The items were reverse coded and a mean score was calculated, with higher scores reflecting greater perceptions of support by the respondent's peers (α = .86). A one factor model fitted the data well (RMSEA = .058, CFI = .933).

At most, 4% of cases were excluded from the analyses due to missing data and, due to the pairwise deletion approach taken in MPlus when the WLSMV estimator is used, less than 1% in the confirmatory factor analyses.

2.3 Results

Results from the larger study, Study One, are presented first, followed by those of the second study.

Descriptive statistics

Summary statistics for the FBS-V and FBS-P indicate that students reported low levels of involvement in bullying behaviours (see **Table 2-1**). Mean scores are close to the minimum value of one and large percentages of students have mean scores at this value, particularly for perpetration. Relatively more boys than girls (about a third versus a fifth) reported no victimisation for all the different forms, but fewer boys reported not being involved as perpetrators of bullying behaviours.

The correlation between the FBS-V and FBS-P was moderate at .38. The items within each version of the FBS were sufficiently but not extremely highly correlated, the bivariate

correlations varied between .15 and .62 for the victimisation items and .16 and .54 for the bullying items.

Both the FBS-V and FBS-P displayed high internal consistency reliability, with Cronbach's alpha values of .87 (item-to-total correlations .48 - .71) and .85 (item-to-total correlations .44 - .67) respectively.

Table 2-1. Descriptive Statistics for the FBS-V and FBS-P (Study One)

	Forms of Bullying Scale - Victimisation (FBS-V)			Forms of Bullying Scale - Perpetration (FBS-P)		
Statistic	Females	Males	Total	Females	Males	Total
n	1,779	1,666	3,453	1,767	1,654	3,430
M[a]	1.41	1.40	1.41	1.10	1.14	1.12
(*SD*)	(0.529)	(0.572)	(0.550)	(0.248)	(0.275)	(0.262)
% non-involved (score = 1)	24.7%	32.4%	28.4%	61.3%	55.9%	58.7%

[a] Individual mean scores range from 1-5

Construct Validity

Since bullying victimisation and perpetration are not independent constructs, construct validity was assessed through confirmatory factor analyses fitting two-factor models to the combined items of the FBS-V and FBS-P. The fit of the two-factor model was confirmed for the overall sample and for female and male students (see **Table 2-2**). All the factor loadings were above .7 (see **Table 2-3**) apart from one item for each of the versions of the scale, indicating each of these two items were relatively less characteristic of the latent constructs of victimisation and bullying respectively.

Table 2-2. Fit Indices for the Two-Factor Models (Study One)

Group	$\chi 2$	RMSEA 90% CI	CFI	Correlation between the factors
All observations (n = 3,484)	1449.7***	.047 [.044, .049]	.960	.463
Females (n = 1,796)	587.2***	.037 [.034, .040]	.975	.513
Males (n = 1,679)	548.8***	.037 [.033, .040]	.979	.488

Note. RMSEA = Root Mean Square Error of Approximation; CI = Confidence Interval; CFI = Comparative Fit Index
*** p<.001

The factor structure was not invariant across gender groups, difference in CFI = .019 > .002 and Satorra-Bentler scaled chi-square difference test TRd = 1215.0, df = 96, p < .001; implying that care needs to be taken when comparing composite scores of these scales for boys and girls. Overall, the differences in the standardized factor loadings were .11 or less in magnitude, .03 on average for the victimisation and .01 for the bullying items (see **Table 2-3**). On inspection, differences between the gender groups were more evident for the victimisation, difference in CFI = .014 > .002 and TRd = 869.1, df = 49, p < .001, than the perpetration items, difference in CFI = .004 > .002 and TRd = 353.7, df = 49, p < .001. In further testing of the factor loadings and the corresponding thresholds for each of the items respectively, based on the difference in CFI criteria, items *c, e* and *i* in the FBS-V and item *e* in the FBS-P were not invariant for males and females. Items *c* and *i* measure relational and item *e* physical bullying (see Appendix 2-B for item wording). These differences between the gender groups are in concordance with the higher likelihood of males engaging in more direct and girls in relational forms of bullying (Archer, 2004; Crick & Grotpeter, 1995), and are similar to differences found in testing the OBVQ (Breivik & Olweus, Under review).

To gauge the substantive impact of assuming scalar invariance, we calculated factor scores for the students in the sample based on assuming and not assuming equal factor loadings

(and thresholds) for the gender groups. The median differences in the factor scores (equal minus unequal assumption scores) for the FBS-V were -.020 and .019 for males and females respectively, and for the FBS-P were .045 and -.038 respectively (range of differences -0.22 to 0.15 between values on a four point scale of 1 to 5). Assuming invariance, therefore, would result in a small average underestimation of the extent of the victimisation for males and overestimation for females, and the opposite would be the case for perpetration.

Concurrent Validity

The survey instrument included global single item questions on frequency of bullying victimisation and perpetration. As recommended (Solberg & Olweus, 2003), the global questions were used to dichotomize students as being victimized (experience every few weeks or more often) or not (having experienced bullying behaviours 1-2 a term or less often) and similarly, as being perpetrators or not. Validity of the multi-item scale scores were further assessed by comparing the mean scores on the FBS-V and FBS-P within the groups defined by the global questions. Comparisons were conducted for the total sample and for each gender group. The victimized/perpetrating groups scored significantly higher on the corresponding versions of the scale in each instance (see **Table 2-4**). These findings were replicated in each of the gender groups (data not shown as the gender results are similar to those of the total sample, available from the first author).

Table 2-3. Standardized Factor Loadings in Two-Factor Model (All Observations and per Gender, Study One)

Item	All observations		Females		Males	
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 1	Factor 2
Teased	.82	.82	.80	.83	.84	.80
Secrets told	.83	.79	.85	.80	.81	.79
Friendship broken	.66	.72	.69	.74	.72	.75
Made to feel afraid	.72	.83	.69	.85	.75	.81
Physically hurt	.73	.78	.70	.79	.78	.78
Called names	.81	.82	.79	.77	.83	.85
Made to do things	.71	.82	.72	.83	.71	.80
Property damage	.64	.79	.64	.83	.64	.77
Left out	.74	.66	.70	.70	.81	.68
Lies told / Rumours spread	.85	.85	.87	.86	.83	.84

Note. Factor 1= Bullying victimisation; Factor 2= Bullying perpetration

Table 2-4. Mann-Whitney Tests of FBS-V and FBS-P Mean Scores by Groups Based on Global Questions (Study One)

Global question and grouping	*M* (*SD*)	Mann-Whitney Z test
Victimisation (*n* = 3,435)		
1-2 a term	1.27 (0.327)	-29.8***
Every few weeks or more often	2.29 (0.786)	
Perpetration (*n* = 3,400)		
1-2 a term or less	1.10 (0.186)	-14.7***
Every few weeks or more often	1.85 (0.791)	

Convergent and Discriminant Validity

As evidence of convergent validity, we hypothesized that the FBS-V mean score, as a measure of level of victimisation, would correlate positively with measures of depression, anxiety, and emotional symptoms, as well as peer problems and correlate negatively with level of peer support. We also hypothesized that scores on the FBS-P for perpetration would be positively correlated with conduct problems, but negatively with pro-social behaviours. Furthermore, as evidence of discriminant validity, we expected the FBS-V to correlate more strongly than the FBS-P with mental health symptoms, and hypothesized stronger correlations between conduct problems and the FBS-P than would be the case for the victimisation version of the scale.

Descriptive statistics for the measures used to assess convergent and discriminant validity are provided in **Table 2-5**. On average the sampled students reported low levels of depression, anxiety, emotional symptoms, conduct and peer problems, whilst high levels of pro-social behaviour and peer support were found.

Table 2-5. Descriptive Statistics Mental and Social Health Outcomes (Study One & Two)

	DASS Depression	DASS Anxiety	SDQ Emotional symptoms	SDQ Conduct problems	SDQ Peer problems	SDQ Pro-social behav's	Peer support
Study One							
n	3381	3381	3344	3344	3349	3354	3444
M	1.33	1.23	1.50	1.38	1.32	2.54	2.47
(SD)	(.562)	(.380)	(.457)	(.327)	(.328)	(.394)	(.382)
Range	1-4	1-4	1-3	1-3	1-3	1-3	1-3
Study Two							
n	778	778	779	779	779	779	
M	1.46	1.36	1.53	1.42	1.54	2.46	
(SD)	.632	.449	.472	.354	.281	.376	
Range	1-4	1-4	1-3	1-3	1-3	1-3	

Note. DASS = Depression Anxiety Stress Scale; SDQ = Strengths and Difficulties Questionnaire

As hypothesized, higher scores on the FBS-V were associated with increased mental health problems and greater problems with peers (see

Table 2-6). Similar correlations have been found in other studies, namely between victimisation and depression of .45 (Hawker & Boulton, 2000) and .47 (Hunt et al., 2012) and for victimisation and anxiety of .25 (Hawker & Boulton, 2000) and .36 (Hunt et al., 2012). Additionally, higher scores on the FBS-P significantly correlated with increased conduct problems and less reported engagement in pro-social behaviours (see

Table 2-6). Further evidence of the validity of the two versions of the scale scores was found with the FBS-V correlating more highly with the mental health outcomes than the FBS-P. Furthermore, the association between conduct problems and the FBS-P was marginally higher than for the FBS-V (see

Table 2-6). The high correlation between the FBS-V and conduct problems is not entirely unexpected given students with behavioural problems are more likely to be bullied, particularly students who both bully others and are bullied (Arseneault et al., 2010; Juvonen et al., 2003; Nansel et al., 2001).

Table 2-6. Spearman Correlations between FBS-V, FBS-P and Mental and Social Health Outcomes (Study One & Two)

	DASS Depression	DASS Anxiety	SDQ Emotional symptoms	SDQ Conduct problems	SDQ Peer problems	SDQ Pro-social behaviours	Peer support
Study One							
FBS-V	.415***	.373***	.340***	.304***	.351***	-.031	-.231***
FBS-P	.232***	.214***	.136***	.346***	.151***	-.213***	-.112***
Study Two							
FBS-V	.412***	.407***	.384***	.347***	.321***	.069	
FBS-P	.295***	.255***	.179***	.367***	.092*	-.121***	

Note. DASS = Depression Anxiety Stress Scale; SDQ = Strengths and Difficulties Questionnaire

*$p < .05$, two-tailed, ** $p < .01$, two-tailed, *** $p < .001$, two-tailed

These same associations were evident in each of the gender groups, the correlations were of the same order of magnitude as those in

Table 2-6 for the male and female students respectively (data not shown as the gender results are similar to those of the total sample, available from the first author).

Study Two Results

Findings based on the data from the first study were replicated using those from the second. Construct validity was confirmed in a two-factor CFA based on the data from the second study, $\chi^2(169) = 477.2$, $p < .001$; RSMEA = .048, 90% CI [.043, .053]; CFI = .970; correlation between the factors = .513; factor loadings for victimisation items ≥ .72; factor loadings for bullying items ≥ .75. The reliability of the FBS-V scores, $\alpha = .92$ and the FBS-P scores, $\alpha = .91$, were also demonstrated, as were convergent and discriminant validity (see

Table 2-6) in a similar manner to that based on the larger first study of Grade 8 students. (A measure of peer support was not included in the second study.) These findings provide some evidence for the applicability of the scales with Australian students in Grades 8 to 10 (13-15 years of age) and in both government and non-government schools.

2.4 Discussion

Various self-report instruments have been developed for the measurement of bullying, however evidence of validity and reliability is limited (Cornell & Bandyopadhyay, 2010; Felix et al., 2011). The FBS-V and FBS-P were designed to measure frequency of involvement in different forms of bullying victimisation and perpetration. The two versions of the scale are for use with adolescents aged 12-15 years and can be administered within the context of a broader questionnaire.

Results from these two studies involving 12-15 year old students support the validity of the item responses to the FBS-V and FBS-P within this age group. Factor analysis confirmed the construct validity. The factor structures were not found to be invariant across gender groups, however, particularly for the FBS-V and the more direct forms of bullying for the FBS-P. Although the effects of these differences seem not to be substantive, researchers need to be cognizant of them when using the FBS, and test for measurement invariance and account for lack of invariance should it exist, prior to evaluating gender differences in their study. Associations as expected with global questions demonstrated concurrent validity of the mean scores on the FBS-V and FBS-P, and associations as expected with conceptually related variables, robust evidence of convergent and discriminant validity. Additionally, scores on the two versions of the scale were found to have good internal consistency. These psychometric properties were demonstrated in the context of online administration of the scales and providing a definition of bullying, together with pictorial representations of the different forms of bullying, and the use of the term bullying.

Advantages of the FBS-V and FBS-P are that both bullying victimisation and perpetration are assessed, and with equivalent items. The items can be used to measure prevalence of involvement in different forms of bullying behaviours and compare this involvement over time or between groups. Composite scores from the FBS-V and FBS-P can be utilized to test individual-level associations, for example correlations between level of victimisation and

mental health outcomes. The scales are comprehensive of the different forms of bullying whilst not being too long to include in larger questionnaires measuring other factors of importance in studies of bullying in schools.

Bullying behaviour is typically characterized by repetition, intent, and a power imbalance and valid measurement of involvement in bullying behaviours requires incorporation of these characteristics (Greif & Furlong, 2006). The wording of the items in the FBS-V and FBS-P aimed to convey intent and the response categories measure frequency of involvement. Unlike the California Bullying Victimization Scale (Felix et al., 2011), the FBS-V and FBS-P items do not specifically refer to an imbalance in power. The nature of a power imbalance may not be easily defined, for example sources may be differences in physical strength (or the targeted person may be outnumbered), social status, intelligence, technological expertise (Greif & Furlong, 2006; Rigby, 2002; Smith, 2012). The power differential is perhaps more easily specified in terms of its consequences, namely that the targeted person has difficulty in stopping the behaviour, and this approach is taken in the wording of the definition of bullying which precedes the items. In addition to the definition, we believe that a disparity in power is further implied through the use of the word *bully* in the stem of the question. As recommended by others (Ortega et al., 2001; Solberg & Olweus, 2003), provision of the definition and use of the term *bully*, helps separate bullying from other acts which are not characterized by these three factors. Additionally, as suggested by Felix et al. (2011), our use of coloured visual depictions of bullying and non-bullying behaviours may have enhanced students' ability to take into account the characteristics of bullying when responding to the scales. Whilst the quality of any pictures is of relevance, those included in our questionnaire were well received by the students during the piloting phase and we believe their use is likely to have enhanced students' understanding of bullying as defined in the questionnaire. For these reasons, we would recommend the definition (together with illustrations if possible) be included with the items in a survey. As acknowledged by Felix et al. (2011), the terms used in the CBVS to describe the power imbalance i.e.: more popular, smart, strong, are not exhaustive nor necessarily ideal. Further research with young people is required regarding the appropriate specification of a power difference.

The validity of the FBS-V and FBS-P scores demonstrated here are as measures of self-perception. Similar to Juvonen et al. (2001), we view self-report as a measure of a person's self-concept, whereas peer and teacher report are to a large extent measures of social

reputation. As such, there is no expectation for high levels of agreement between responses from different sources. The conceptualization of self-, peer- and teacher-report as different constructs implies careful consideration needs to be made in relation to the construct that is of most importance within each research study and therefore, when framing research questions (Griffin & Gross, 2004). For victimisation, the targeted person is arguably the best placed to report on the occurrence of the victimisation. Self-report delivers the perspective of the person experiencing the bullying and is the most useful measure when researching the "victim's plight" or consequences of the victimisation such as depression (Felix et al., 2011; Juvonen et al., 2001). On the other hand, self-report may be limited as a means to measure perpetration of bullying, where one may wish to be less reliant on self-concept which may lead to under-reporting of such behaviours (Cornell & Bandyopadhyay, 2010). These differences have resulted in several authors recommending multiple informants be utilized (Bovaird, 2009; Cornell & Bandyopadhyay, 2010; Griffin & Gross, 2004).

The FBS was developed to measure the major forms of bullying utilizing relatively few items. In constructing the items to be included in the scale, we used wording which was general enough to include a range of behaviours within the form of bullying, without being so broad that young people were not able to relate their experiences to the items. For example, the item "I was hurt by someone trying to break up a friendship" could include a wide range of specific behaviours or actions, both online and offline. Furthermore, whilst not necessarily detailing specific actions, we strengthened items by making the intent clear, e.g. "lies were told and false rumours spread about me by someone, *to make my friends or others not like me*". Nevertheless, young people may not have found a "fit" for their experiences in the listed items and the scale may therefore not capture all instances of bullying.

In the FBS bullying behaviours were measured in a general sense, that is, possibly occurring both online and offline, for three reasons. First, we conceptualize cyberbullying primarily as bullying behaviour (e.g. name-calling, relational, social damage, threatening) delivered through the use of technology, that is, through different modes rather than as a different form of bullying (Dooley, Pyżalski, & Cross, 2009; Felix et al., 2011; Smith, 2012; Smith et al., 2008; Varjas, Henrich, & Meyers, 2009). Whilst the three criteria that distinguish bullying from aggression, repetition, power imbalance and intent, may be operationalized differently for cyberbullying, they are still able to be applied (Dooley et al., 2009; Smith,

2012). For example, the inability to remove a message or image from circulation by the targeted person can be seen as a differential in power. Second, as is the case with many bullying scales, the wording of the majority of the items (apart from physical bullying and damage to property) does not indicate the behaviours are necessarily offline. For example, having rumours or secrets spread about someone could be achieved with and without the use of technology. Should a researcher therefore, wish to measure only offline instances of rumour spreading, this would need to be specified in the item. It cannot be assumed that items from existing scales measure offline behaviour simply because they have historically done so. Third, the bullying behaviour may occur both online and offline. For example, what may start as face-to-face name-calling may be continued on social networking sites or vice versa. Young peoples' increasing use of technology in their social interactions, will only serve to increase the overlap between offline and online bullying. Therefore, we presented bullying behaviours as possibly occurring offline and/or online as we fundamentally see the separation of these modes as problematic when measuring adolescents' involvement in different forms of bullying as defined in this paper.

Thus the FBS measures, what some refer to as "traditional" forms of bullying, but we see more globally as forms of bullying which may occur off- or online. It is not intended, however, to be a comprehensive measure of cyberbullying and as mentioned above, not all instances of bullying, or more specifically cyberbullying, behaviours may be captured by the scale. Further, we would recommend that separate cyber specific scales be included in surveys to enable the exploration of unique aspects of this means of bullying, such as the differing motivations of the perpetrator and impacts on the targeted person (Menesini, 2012; Smith, 2012).

Limitations and Future Directions

While use of the FBS is supported by the results of this study, there are some limitations to the findings. Although the large sample size is a strength, only West Australian metropolitan secondary schools were included and the generalizability of the results beyond this population is unknown. Whilst the FBS can be administered in hard copy, the extent to which the results are applicable to modes other than online administration of questionnaires, is also unclear. The data were collected in a term subsequent from the one that the students were reporting about, separated by a two week holiday period. Students' experiences in the current term may, therefore, have influenced their responses to the questions on bullying involvement in the previous one. Due to the large sample, it took four

to five weeks to collect data from all study schools, which may also have contributed to some variation in student responses.

Correlations, as measures of validity, can be the result of shared method variance. This is the variance shared by variables measured by the same method, in this case self-report, which can artificially inflate the correlations between such variables. The low cross-correlations between internalizing problems and perpetration on the one hand and externalizing problems and victimisation on the other hand, provide evidence of a lack of shared variance in these data and therefore, the validity of the results. Zero cross-correlations are not expected because of the *bully-victim* subgroup of students who both experience and perpetrate bullying behaviours and so would be expected to experience both internalizing as well as externalizing problems.

While scores on the FBS were valid and reliable measures of the frequency of bullying victimisation and perpetration, further testing of the scale with other age groups and in other contexts is required. Furthermore, it will be of value to explore additions to the FBS, in particular the victimisation version, to measure the severity of different forms of bullying in terms of the impact or harm as perceived and reported by the targeted person.

Chapter 3 : Testing for response shift bias in evaluations of school anti-bullying programs

Citation

Shaw, T., Cross, D., & Zubrick, S. R. (Under review). Testing for response shift bias in evaluations of school anti-bullying programs. *Evaluation Review*.

Date submitted: **4 December 2012**

Date resubmitted: **10 April 2013**

Contribution of authors

The candidate was responsible for the literature review, conceptual framework, data analyses and discussion for this research paper. Professors Cross and Zubrick provided expert advice during the preparation of the manuscript as well as assisting in reviewing the manuscript.

Relevance to the thesis

This chapter presents discussion and analyses central to Research Question 2 of this thesis. Methods for assessing response shift bias are described and their application illustrated. The outcomes of this chapter inform the discussion on assessing this potential bias in program evaluations.

Note: This article was resubmitted with responses to the reviewers on 10 April 2013. Unfortunately the editor of the journal, who was also one of the reviewers, was unwell and passed away recently. Hence, a final decision on the article has been delayed and is pending.

Abstract

Background: Involvement in bullying at school is detrimental to students' mental and physical health; however, school anti-bullying programs have not been found to be uniformly successful. Self-reported frequency of involvement in bullying victimisation and perpetration, often used as outcome measures in program impact evaluation studies, may be subject to response shift, particularly in intervention conditions. Such differential shifts could lead to biased estimates of program effects.

Objectives: This study investigated the presence of reconceptualization, reprioritization, and recalibration response shift, resulting from intervention implementation.

Subjects: Grade 8 students (n = 3,382) in the 35 schools participating in the Cyber Friendly Schools Project, a longitudinal group-randomized intervention trial.

Analyses: Response shift was assessed by comparing traditional and retrospective pre-test measures of bullying involvement, as well as testing for measurement invariance over time in the Forms of Bullying Scale (FBS) using confirmatory factor analyses.

Results: No evidence of response shift was found, indicating students' understandings of bullying behaviour remained stable over time. These findings also demonstrate the applicability of the FBS in longitudinal studies involving adolescents.

Conclusion: Whilst response shift was not present in our study, researchers conducting program evaluations in other contexts are advised to consider testing for this potential source of bias in their studies.

Keywords

Response shift, measurement invariance, bullying measurement

Acknowledgements

This research was made possible through funding from the Western Australian Health Promotion Foundation (Healthway). We thank the students and staff in participating study schools, and colleagues within our research group who worked on the Cyber Friendly Schools Project, in particular Patricia Cardoso.

3.1 Introduction

For children and young people, bullying victimisation is associated with concurrent and lasting negative psychological and social outcomes (Arseneault, Bowes, & Shakoor, 2010; Card & Hodges, 2008) and for males, bullying perpetration is predictive of engagement in problem behaviours in early adulthood (Renda, Vassallo, & Edwards, 2011; Sourander et al., 2011). Whilst some school bullying prevention programs have been shown to be effective, others have resulted in small, null, or even negative effects (Farrington & Ttofi, 2009; J. D. Smith, Schneider, Smith, & Ananiadou, 2004). The reasons for these differing results are not fully understood, but include differences in the methods used to evaluate the programs, lack of program implementation, insufficient staff training, and differences in program components and approaches (Cross, Epstein, et al., 2011; Farrington & Ttofi, 2009; Merrell, Gueldner, Ross, & Isava, 2008; J. D. Smith et al., 2004). One further explanation may be response shift bias. In this paper we explore the phenomenon of response shift as it relates to the measurement of bullying behaviour and its potential impact on evaluation studies of anti-bullying programs in schools, and describe and apply two methods of testing for response shift in this context. To our knowledge, this is the first study to investigate response shift in the bullying context.

Response Shift

Response shift is a change over time in the meaning an individual places on a construct. It has been characterized as reconceptualisation (a redefinition of the construct), reprioritisation (a shift in values or the importance of the various domains of the construct), and/or recalibration (a shift in internal standards of measurement) of the construct being measured (Sprangers & Schwartz, 1999). Initial research examined the impact of the phenomenon on the evaluation of educational training programs (Howard, 1980) and in management science research (Golembiewski, Billingsley, & Yeager, 1976). Response shift has also been assessed in evaluations of training programs in leisure studies (Sibthorp, Paisley, Gookin, & Ward, 2007) and parenting programs (Hill & Betz, 2005). Researchers assessing health-related quality of life (HRQL) lead the field (e.g., Barclay-Goddard, Epstein, & Mayo, 2009; Nolte, Elsworth, Sinclair, & Osborne, 2012; Schwartz, Sprangers, & Fayers, 2005).

The different types of response shift can be illustrated by considering the measurement of HRQL in terminally ill patients (Barclay-Goddard et al., 2009). As a patient's health status worsens, his or her conceptualisation of the meaning and quality of life may change,

resulting in changes in the content of the domains of HRQL. Additionally, as his or her physical condition deteriorates, the priority or value placed on different domains of HRQL (e.g., physical functioning, bodily pain, social functioning, and vitality) may shift (reprioritisation); as may his or her perception of pain, resulting in recalibration of pain scores. The types of response shift are inter-related, with changes of one type able to lead to shifts of another type (Schwartz & Sprangers, 1999). Critically, studies aimed at obtaining valid measures of changes in self-reported quality of life must account for possible reconceptualisation, reprioritisation, and recalibration shifts that may occur as a consequence of changes in the health status of respondents, as even small shifts can lead to biased estimates of true change (Schwartz et al., 2006). Although primarily investigated in terms of self-report, response shift can potentially occur in any subjective measurement and has particular relevance in program evaluation studies.

Golembiewski (1975) described alpha, beta, and gamma change, where alpha change referred to change in the level of a construct measured without response shift or "true" change, beta change to recalibration, and gamma change to redefinition or reconceptualisation response shift. When evaluating program impact, the aim is usually to measure alpha change. Assessment of program effects in terms of change in pre-and post-program scores, relies on stable, consistent measurement of the construct over time (Cronbach & Furby, 1970) – a concept closely aligned to that of measurement invariance (Vandenberg & Lance, 2000; Wu, Li, & Zumbo, 2007). In the same way as response shifts may occur in HRQL as a consequence of changes in health status, conceptualisation, prioritisation, and calibration changes may also occur as a result of a program or intervention (e.g., Hill & Betz, 2005; Howard, 1980; Nolte et al., 2012; Sibthorp et al., 2007). In fact, such change may be an aim of the intervention. For example, a first step in an anti-bullying program may be to achieve a common understanding amongst participants of bullying behaviour, inclusive of the different forms of bullying (e.g., Cross, Monks, et al., 2011). As another example, chronic disease self-management interventions may target participants' perspectives in order to assist them to lead active lives despite their condition (Nolte et al., 2012). Response shift manifested as systematically different responses to outcome measures pre- and post-program can, therefore, impact on the validity of estimates of program effects based on those measures.

Response shift is a threat to not only the valid comparison of pre-, post-program scores within the group participating in the program, but also the comparison of such changes

over time between study conditions in pre-post evaluation studies. If the response shift is not equivalent between an intervention and control group, then the extent to which the estimated differences between the groups are a consequence of true change in the levels of the construct or differences in the reporting of the construct will be unclear. Similar to differential history, maturation, testing or instrumentation effects (Murray, 1998), differential response shift, e.g., when it occurs to a greater extent in an intervention than a control group, will introduce bias into group comparisons. Thus, if unaccounted for, response shift will likely lead to biased estimates of program effects (e.g., Hill & Betz, 2005; Howard, 1980; Sibthorp et al., 2007) and has been raised as a possible explanation for null or negative effects of interventions targeting aggression and bullying in schools (Nixon & Werner, 2010; Orpinas et al., 2000; J. D. Smith et al., 2004; P. K Smith, Ananiadou, & Cowie, 2003).

Bullying and Response Shift

A commonly used definition of bullying is intentional aggressive behaviour repeatedly perpetrated over a period of time against an individual who finds it difficult to defend him/herself (Olweus, 1996). The repetition and power imbalance specified in the definition, differentiate bullying from other forms of aggression. Bullying may take different forms, namely verbal (e.g., nasty name-calling), threatening (e.g. making others feel afraid), physical (e.g., hitting), relational (e.g. exclusion from friendship groups), and social (e.g., spreading false rumours) (Crick & Grotpeter, 1995; Shaw, Dooley, Cross, Zubrick, & Waters; Solberg & Olweus, 2003). In recent years, technology has become a tool used by perpetrators. The resultant ease with which messages and/or images can be rapidly distributed to a wide group and resultant publicity, as well as the extension of the bullying beyond the school environs and hours, and the ability in certain circumstances for the perpetrator to remain anonymous, mean that cyberbullying incidents can potentially be more harmful than bullying perpetrated through more traditional means (Campbell, 2005; Slonje, Smith, & Frisén, 2013; P.K. Smith et al., 2008).

Self-reported frequency of involvement in bullying victimisation and perpetration are often used as outcome measures in program impact evaluation studies (Merrell et al., 2008). As individuals may have different understandings of the term bullying (P.K. Smith, Cowie, Olafsson, & Liefooghe, 2002), researchers who explicitly use this term in their question wording often provide respondents with a definition of bullying behaviour in an attempt to achieve a common understanding of the construct according to the stated definition

(Griffin & Gross, 2004; Solberg & Olweus, 2003). A common understanding is further enhanced by including in the definition examples of the different forms of bullying behaviours as well as of behaviours, such as playful teasing, which are not bullying (Griffin & Gross, 2004; Shaw et al.; Solberg & Olweus, 2003). Disadvantages of the use of the term bullying are that using an emotionally laden label for the behaviour, may lead to under-reporting (Greif & Furlong, 2006; Kert, Codding, Tryon, & Shiyko, 2010).

We were unable to find studies that have directly investigated the phenomenon of response shift in self-report measures of bullying victimisation and perpetration. Authors of some studies have suggested raised awareness and sensitisation, as a consequence of participation in an intervention, as a possible explanation for greater reporting of bullying, aggression, and violence (Nixon & Werner, 2010; Orpinas et al., 2000; J. D. Smith et al., 2004; P. K Smith et al., 2003), and others postulated that an intervention may lead to less reporting of undesirable behaviours (Cornell & Bandyopadhyay, 2010). In light of this lack of direct evidence, but in line with the suggestions of these authors, we postulate that an anti-bullying intervention in schools may induce the three types of response shift defined by Sprangers & Schwartz (1999), as described below.

An intervention which describes bullying according to the three characteristics of repetition, power imbalance, and intent detailed in the definition, may lead participants to reconceptualise the construct. Qualitative studies have shown that young people may define bullying differently from researchers' definitions, for example not include a power differential or repetition (deLara, 2012; Varjas et al., 2008). Another change in understanding following an intervention may involve the recognition of certain behaviours as bullying, not seen as such previously e.g., social exclusion (Nixon & Werner, 2010; Orpinas et al., 2000; P. K Smith et al., 2003), resulting in a broader understanding of the construct.

Reprioritisation involves a change in values, e.g., that the different forms of bullying would be perceived as more or less important or severe than before. For example, exposure to the impact that a form of bullying had on a victimized student, of which the intervention participant was previously unaware, may lead to changes in his or her perception of the relative severity of different bullying behaviours. It is unclear, however, how this may impact on responses to the commonly used measures which assess the frequency of behaviours.

Recalibration implies a shift in scale, in this case a shift in report of bullying frequency. Raised awareness of bullying may increase the frequency of reporting (Orpinas et al., 2000; J. D. Smith et al., 2004), although a shift in norms may decrease the frequency (Cornell & Bandyopadhyay, 2010). The types of response shift are inter-related and one type could lead to another. For example, recognition of exclusion as a form of bullying (reconceptualisation), could lead to a change in the understanding of the frequency in which such incidences occur (recalibration).

Methods of Assessing Response Shift

Many methods of testing for and dealing with response shift have been proposed (Barclay-Goddard et al., 2009). Two principally used methods, also applied in this study, will be described: the use of retrospective pre-test (or then-test) questions (Hill & Betz, 2005; Howard, 1980; Schwartz & Sprangers, 2010) and testing measurement invariance in factor analyses or structural equation models (Millsap & Yun-Tein, 2004; Oort, 2005; Vandenberg & Lance, 2000). The former would apply in studies where single-item questions are utilized to measure bullying involvement, often with the aim of categorizing students into groups e.g. as “bullies” or “victims”, to measure prevalence (Solberg & Olweus, 2003). The latter is appropriate when the bullying construct is measured using a multi-item scale, e.g. to investigate individual-level associations between bullying outcomes and other measures such as mental health outcomes. Both forms of questions are used to assess the impact of anti-bullying programs (Farrington & Ttofi, 2009).

Retrospective pre-test questions included in post-program surveys in longitudinal studies ask the respondent to retrospectively report on his or her status prior to participation in the program. For example, after a training program targeting certain skills, participants may be asked to report on their competence to perform a specific task as it was prior to (“then”) and as it is after the training (“currently”). The initial motivation for using these questions was to avoid bias that may be present in scores obtained pre-program due to, e.g., an individual’s lack of understanding or awareness of the construct being measured. More generally the aim is to measure change unaffected by response shift that may occur as a result of program participation or changes in health status (Barclay-Goddard et al., 2009; Heller et al., 2011; Howard, 1980; Lam & Bengo, 2003). Because both are reported at the same time and so presumably, made from the same perspective and on the same metric, for certain constructs, retrospective pre-test and post-test scores are seen as more

comparable than are traditional pre-test, post-test scores (Heller et al., 2011; Howard, 1980; Kumpfer, Xie, & O'Driscoll, 2012; Sibthorp et al., 2007).

Retrospective pre-tests may also be subject to a range of biases, however, such as recall (where responses are subject to memory distortions or failures), effort justification (where the respondent reports change as a justification for the effort they have taken to participate in a program), implicit theories of change (where the respondent expects change to occur as a result of attendance), self-enhancement (where a respondent wishes to present themselves in the best light), and social desirability bias (where change is reported as this is seen to be what is expected by others) (Hill & Betz, 2005; Schwartz & Sprangers, 2010; Taylor, Russ-Eft, & Taylor, 2009). These may lead respondents to ensure retrospective pre-test and post-test scores reflect change, especially if the two sets of questions are placed adjacent to each other in the survey, such that respondents can compare their responses to each, regardless of any change or lack thereof that may have occurred. The inclusion of retrospective pre-tests in post-test surveys has also been found to bias reports of post-test status (Nolte et al., 2012).

A vast body of literature on testing measurement invariance exists (e.g.,Millsap & Yun-Tein, 2004; Vandenberg & Lance, 2000; Wu et al., 2007). Oort (2005) attached meaning to the usual tests of measurement invariance by linking each form of invariance to a type of response shift, describing the use of a structural equation framework to detect response shift over two time periods. Fitting linear factor analysis models, reconceptualisation is indicated by a change in the pattern of the factor loadings (Oort, 2005). As a simple example, an indicator(s) loading on one factor at one time point and not at the second, would indicate the meaning of the construct changed over time. Assessment of model fit and comparison of the patterns in the factor loadings in confirmatory factor analyses, fitting the same model to the data from the two occasions, are methods of detecting this type of response shift. Differences over time in the magnitudes of the factor loadings within a factor, are seen as indicative of reprioritisation (Oort, 2005) i.e., changes in the loading for a particular indicator variable (lack of factor invariance) would signify changes in the "importance" of that indicator with regard to the construct being measured. Clearly these two forms of response shift are inter-related for one factor models. Recalibration was described as being uniform and/or non-uniform. Within the framework of linear factor analyses, uniform recalibration was seen as a constant shift across the range of the indicator variables resulting in a shift in the means of the observed indicator variables (lack

of invariance of the intercepts) (Oort, 2005). Oort (2005) proposed that shifts of different magnitude across the variable ranges would result in non-uniform recalibration, manifested in differences in the covariance structure of the observed indicators (invariance of residual factor variances). Recalibration tests are simplified for categorical data, where multiple thresholds (one less than the number of categories) are modelled for each indicator. Recalibration would result in shifts in the thresholds - uniform recalibration as a constant shift across all thresholds, and non-uniform as a differential shift depending on the particular category/threshold.

The Current Study

The Cyber Friendly Schools Project (CFSP) was an intervention trial conducted in 35 schools, randomized to an intervention and control group. The aim of this paper is to assess whether response shift occurred in the self-report measures of bullying from prior to after intervention implementation and if so, whether this shift was differential and introduced bias into the comparisons of bullying victimisation and perpetration outcomes between the study conditions. We hypothesized a response shift occurred in the intervention group in the reporting of bullying behaviours, either as a result of changes in perceptions of bullying and/or raised awareness, which did not similarly occur amongst the control students. We were also interested to determine whether the magnitude of any such shifts, should they be present, differed for bullying perpetration and victimisation reports.

3.2 Methods

Study Design and Participants

Data presented here were collected from the cohort of participating students prior to and after implementation of the Cyber Friendly Schools Project (CFSP) intervention in 2010 and 2011. The intervention was designed for students in Grades 8 and 9 (mostly 13-14 years of age), immediately following the transition to secondary school, a time of increased prevalence of bullying. In addition, students' increased use of technology for social interaction as they move into their teenage years, places them at greater risk of being cyberbullied or using technology to bully others (Slonje et al., 2013; P.K. Smith et al., 2008). The 2010 Grade 8 student cohort were, therefore, the focus of the study.

A cluster sampling scheme was employed to recruit schools and students. All non-government schools in Perth, Western Australia were approached to participate (school response rate 67%, reasons given by schools for non-participation were competing priorities within the school and participation in other on-going research projects). Government schools were not approached due to restrictive consent procedures required in this school sector which in our previous studies had resulted in student recruitment rates below 20%. Schools were randomized to intervention (n = 16) and control groups (n = 19). Consent was sought from parents and all Grade 8 students in each school (non-consent rate=7%). The combined consent/response rate was 87% for the pre-test data collection in 2010 (n = 3,382), of these 83% completed usable surveys at the follow-up data collection in 2011 (n = 2,813). Non-response was mainly due to students being absent from school on the day of survey administration and failing to complete the survey on their return to school, and attrition due to the movement of students out of study schools. The mean age in Grade 8 was 12.9 years (SD = 0.42), and 53% (n = 1,804) were girls. Information on students' ethnicity was not collected, however, students in metropolitan schools in Perth come from a range of ethnic backgrounds, and typically a small percentage (less than 5%) are Indigenous. Based on a socio-economic index of the Australian Bureau of Statistics, 75% of the students lived in areas of higher than average socio-economic status.

Data Collection and Ethics

Participating students completed online surveys administered by trained research staff in controlled conditions during normal class periods. Students were assigned unique login numbers as their responses were to be tracked over time, and assured of the confidentiality of their responses. Non-participating students completed alternate activities as assigned by their classroom teacher. At the completion of the survey, the students received information regarding help they could access confidentially should the survey have raised any issues of concern to them.

In all instances we obtained parental and student informed consent, and ethical approval from the Human Research Ethics Committee at Edith Cowan University and the relevant school authorities.

CFSP Intervention

The whole-school CFSP intervention was based on a socio-ecological approach (Bronfenbrenner, 1995) targeting the many online contexts in which 13 to15 year old students interact and the responses they receive from their actions in each context.

Intervention schools were encouraged to integrate the CFSP resources into their current approaches addressing inappropriate social behaviour, including traditional bullying. The intervention's conceptual framework recognized the seamless online/offline social context of students' lives and the means by which they engage with others in these environments. It addressed the proximal and distal ecological, cognitive and psychosocial risk and protective factors that can be regulated or mediated at the school, classroom, family and individual levels to reduce cyberbullying. Accordingly the intervention targeted the pastoral care staff and student cyber-leaders to implement the whole-school activities, and also students via their classroom teachers and parents.

Whole-school intervention – pastoral care staff and student cyber-leaders

The whole-school strategies were implemented in each year of the study by Grade 10 (age 15 years) student cyber-leaders, in collaboration with their school's pastoral care staff. A six-hour training was provided to student leaders and pastoral care staff in each intervention school, using hard copy and online support materials provided by the project. Approximately four students in each intervention school were trained to act as a catalyst in their school to lead positive cyber safety action and to enliven the cyberbullying-related strategies implemented by the pastoral care staff. These strategies included reviewing related school policies, increasing student awareness of their rights and responsibilities online (especially as a bystander), using restorative approaches to deal with bullying incidents, and providing parent cyberbullying prevention training.

During the training the cyber-leaders used a purpose designed website to plan and implement at least three major whole-school activities to encourage students' positive use of technology in their relationships with other users. The pastoral care team completed a 'map the gap' tool to assess their current whole-school practices to reduce bullying. This tool was based on principles for successful practice developed by the research team [citation deleted for blind review]. School-level findings from this assessment were used by staff to determine which whole school actions were missing or poorly addressed by their school. School teams were provided with resources to respond to these gaps in practice, including reviewing how their school policies addressed cyberbullying.

Student cohort intervention - classroom teachers and parents

During the first year of the study, when the student cohort were in Grade 8, the intervention teachers were provided with hardcopy classroom teaching and learning materials comprising eight modules addressing the nature and prevalence of cyberbullying,

social problem solving/positive conflict resolution, social skill and empathy development and school policy information. In the second year the teaching and learning was provided online through the CFSP website. The self-directed online program extended the learning in year one, via eight modules which addressed social rules, "netiquette" development, legal issues and reporting of cyberbullying, negative influences on online behaviour, safe use of ICT including internet privacy and protection, preventative action, and online bystander education.

All intervention teachers of the student cohort in each of years one and two of the study participated in a two-hour training to consolidate common understandings and their role in teaching the learning activities.

Online and hard copy self-help parent resources were also provided to develop parents' knowledge of young people's behaviour online, while building their self-efficacy to help their children in this environment. Schools were also encouraged, and given prepared powerpoint presentations, to offer face to face family awareness raising seminars through their student leaders. Skills-based newsletter items were also provided to encourage parents to proactively support their children's online behaviour and to visit the CFSP website for further support and ideas.

Measures

Bullying victimisation and perpetration were measured using multi-item bullying scales as well as global single-item questions.

Bullying scales: Involvement in different forms of bullying victimisation and perpetration were measured using the respective versions of the Forms of Bullying Scale (FBS) [citation deleted for blind review]. In the FBS bullying behaviours are measured in a general sense, that is, possibly occurring both online and offline. Reasons are that we see cyberbullying primarily as bullying behaviour delivered through the use of technology, that the wording of the majority of the items in the bullying scales do not indicate the behaviours are necessarily offline and that the bullying behaviour may occur both online and offline, making it difficult for young people (who may not distinguish between the modes in the same way as adults do) to report on each separately.

Each version of the FBS comprises ten items measuring the different forms of bullying. For example, the victimisation version (FBS-V) asks: "Last term, how often were you bullied (including cyberbullying) by one or more young people in the following ways?" with items

such as “I was called names in nasty ways”, “Secrets were told about me to others to hurt me”, “I was hurt by someone trying to break up a friendship”. The perpetration version of the scale (FBS-P) is made up of the same items, with wording changes to reflect perpetration as appropriate. The response options assess the frequency bullying occurred within the previous term (a period of approximately ten weeks): “This did not happen to me/I did not do this”; “Once or twice”; “Every few weeks”; “About once a week”; and “Several times a week or more”. The FBS-V and FBS-P were preceded in the survey by a definition of bullying based on that of Olweus (1996), which included examples and pictures illustrating the different forms of bullying (with representations of bullying by technology as well as more traditional methods), as well as two examples of behaviours that are not bullying.

Construct validity (two-factor model fit RMSEA = .047; CFI = .960) and concurrent reliability (FBS-V: α = .87 and FBS-P: α = .85) have been demonstrated for young adolescents (primarily based on the Grade 8 data utilized in this study). Additionally, significant correlations with social-emotional outcomes, such as depression, anxiety, conduct problems, and peer support, provided robust evidence of convergent and discriminant validity [citation deleted for blind review].

Traditional pre-test questions: The Grade 8 survey, administered in Term 2 of Grade 8, included global single-item questions measuring the extent to which students were bullied / had bullied other students at their school in Term 1 of Grade 8. These questions were placed after the respective bullying scales and the response options were as for the bullying scales.

Retrospective pre-test questions: The Grade 9 survey, administered at the end of Grade 9, included global single-item questions measuring the extent to which students were bullied/had bullied another student when they were in Term 1 in Grade 8. These questions were placed after the bullying definition but prior to the scales assessing Grade 9 involvement. The response options were as for the bullying scales.

Analytic Approach

Both the use of retrospective pretest scores and confirmatory factor analyses (CFA) were utilized to test for response shift. The different types of response shift were assessed using CFA while analyses of pretest scores focused on identifying response shifts within subgroups of individuals.

Traditional and retrospective pre-test questions

To determine whether a shift had occurred over time we compared the students' responses regarding bullying involvement in Term 1 Grade 8 as reported in Term 2 of that year (traditional pre-test) with their reports obtained in Term 3 or 4 of Grade 9 (retrospective pre-test) using descriptive statistics. Students were categorized as having been bullied or having bullied others if this occurred every few weeks or more often (Solberg & Olweus, 2003). To assess whether the shifts were differential between the groups, we determined the percentages of students whose Grade 9 reports were consistent with their Grade 8 responses and those for whom the Grade 9 reports were inconsistent. Using logistic regression we tested for differences in the odds between the intervention and control groups of retrospectively reporting involvement in bullying, for those who reported involvement in Grade 8. Random intercepts were included in the models to account for the clustering of students within schools.

Confirmatory factor analyses

The proposed approach to testing for the presence of response shift bias was to initially test for response shift over time within each study condition and if found to be evident, to compare the magnitude of the shift in the groups to determine whether it was differential i.e., whether a larger shift was present in the intervention than the control group. Separate analyses were conducted for bullying victimisation and perpetration. We recoded the items to three categories by combining the top three due to the small numbers in the two highest frequency response options. In testing for response shift, we followed the testing procedure proposed by Oort (2005), but adapted for categorical factor analyses, which we applied given the ordinal, highly skewed nature of the items in the FBS-V and FBS-P. We conducted the analyses in Mplus 6.0, applying the WLSMV weighted least squares estimator and delta parameterisation (Muthen & Muthen, 1998-2009). Robust standard error estimation was used to account for school-level clustering.

Use of pairwise deletion strategies in MPlus, ensures minimal exclusion of cases due to missing data (between 4 and 19 cases with missing data on all the items in the scales). However, the analyses assume the data lost through drop-outs from the first to the second data collection (n = 569, 17%) are missing completely at random (MCAR). Comparisons of the cases lost and not lost to follow-up using random coefficients logistic regression showed no significant differences on study condition (p = .681), gender (p = .477), school type (p = .390), or on measures of socio-economic status such as being a child in a single-

parent family (p = .153) or living in a below average socio-economic area (p = .180). Importantly the two groups did not differ on the mean scores for the victimisation (p = .075) and perpetration scales (p = .089) in Grade 8. Thus, this missing data assumption seems reasonable.

As a first step, the presence of reconceptualisation was evaluated by assessing the fit of longitudinal one-factor CFA models – adequate model fit and factor loadings above .6 on both occasions, were seen as indicative of a stable conceptualisation of the victimisation / perpetration construct within the group. In a second step, as recommended (Muthen & Muthen, 1998-2009) the invariance of the factor loadings and thresholds were tested simultaneously, because these jointly determine the probability curves of the items. Models with the parameters constrained and not constrained to be invariant were compared, and invariance of factor loadings across time was seen as evidence that reprioritisation had not occurred and invariance of the threshold parameters as evidence against recalibration (both uniform and/or non-uniform).

3.3 Results

Traditional and retrospective pre-test questions

Table 3-1 summarizes the results for bullying victimisation and perpetration between students. In general, across both groups, students who were categorized as not involved in bullying in their Grade 8 survey, tended to report the same level of involvement retrospectively in Grade 9, i.e., 93% for victimisation and 96% for perpetration respectively. In contrast, many of those who were categorized as involved in Grade 8, reversed their responses to non-involvement in their Grade 9 report – 64% and 66% for victimisation and perpetration respectively. So, the responses were largely consistent from Grade 8 to Grade 9 for those who initially reported non-involvement, but largely inconsistent for those who initially reported involvement.

Table 3-1. Grade 8 & 9 report of bullying involvement in Term 1 Grade 8, by or toward another student

Traditional pre-test (Grade 8 report)		Retrospective pre-test (Grade 9 report)			
		n	%	*n*	%
Victimisation (*n* = 2,764)		Not bullied in Gr8		Was bullied in Gr8	
Not bullied in Gr8	Intervention	1,249	93	91	7
	Control	1,045	93	79	7
	Total	2,294	93	170	7
Was bullied in Gr8	Intervention	113	62	70	38
	Control	78	67	39	33
	Total	191	64	109	36
Perpetration (*n* = 2,763)		Did not bully in Gr8		Bullied others in Gr8	
Did not bully in Gr8	Intervention	1,437	96	55	4
	Control	1,179	97	42	3
	Total	2,616	96	97	4
Bullied others in Gr8	Intervention	14	54	12	46
	Control	19	79	5	21
	Total	33	66	17	34

When comparing the relative shifts in the two study conditions, it appears as though the reversal from Grade 8 to 9 was larger for the control than the intervention students, particularly for perpetration (79% versus 54%) and less so for victimisation (67% versus 62%). Hence, the intervention students seem to be more likely than those in the control condition to report bullying involvement in the post-intervention survey. Logistic regression analyses were conducted testing intervention to control differences for each of these two subgroups of students that were categorized as having bullied others (*n* = 50) or having been bullied (*n* = 300) respectively, based on their Grade 8 survey. Comparisons of the control and intervention groups within each of these subgroups defined by their Year 8

responses, were non-significant indicating consistency of reporting perpetration in the Grade 9 survey in the first subgroup (*OR* = 0.23, 95% CI [0.05, 1.03], *p* = .055) and of victimisation in the second subgroup (*OR* = 0.83, 95% CI [0.41, 1.72], *p* = .622). Thus, although the percentages in each of the study conditions appeared to differ post-intervention with regard to their reporting of bullying behaviours, particularly for perpetration (with an odds ratio of 0.23 corresponding to a large effect size of 0.8), the statistical evidence based on the retrospective pre-test questions does not support the presence of response shift.

Confirmatory Factor Analyses

First, the presence of reconceptualisation was evaluated by fitting unconstrained longitudinal one-factor CFA models for the combined data over the two time points for victimisation and perpetration respectively, for each of the study conditions. As can be seen in Table 3-2, the fit of each of the four unconstrained models was adequate, RMSEA < .06 and CFI > .95 (Hu & Bentler, 1999). Additionally, the factor loadings were above .6 in value for each time point for each of victimisation and perpetration and each of the two groups. This consistency of factor loading patterns supported the stability of the measurement of the constructs over time. No evidence was found, therefore, that the bullying victimisation or perpetration constructs were conceptualized differently at the two time points.

Second, models wherein the factor loadings and thresholds were constrained to be equal were fitted and compared to the above unconstrained models to assess measurement invariance and hence, the presence of reprioritisation and/or recalibration forms of response shift. Results from simulation studies have shown that alternate fit indices, such as the CFI, are less sensitive to sample size and more sensitive to a lack of invariance (Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008). The use of these indices are, therefore, recommended above chi-square tests to assess measurement invariance in large samples such as in this study. Thus, we are reporting changes in the CFI together with the usual chi-square difference tests. In addition, we inspected differences in the RMSEA as an alternate measure of model-fit. A lack of measurement invariance would be indicated by significantly worse fit in the constrained models than that of the unconstrained models. As is evident in Table 3-2, contrary to expectations, the CFI indicated either the same or improved fit under the constrained model. The chi-square difference test was not significant for perpetration in either group or victimisation in the control group, but was for

victimisation in the intervention group (p = .045). Given the large sample size (n > 1500) and the known sensitivity of chi-square tests to sample size (Joreskog & Sorbom, 1993; Meade et al., 2008), we are hesitant to interpret this latter result as indicative of a lack of measurement invariance, especially as, for example, the differences in the standardized factor loadings are .007 or smaller in size. Additionally, in each instance the RMSEA of the constrained model indicated better fit than that of the unconstrained, hence measurement invariance is also indicated by this measure of model fit. Our overall conclusion is that the data support the invariance of the thresholds and factor loadings over time for both bullying victimisation and perpetration in each of the groups, and hence reprioritisation and recalibration response shifts are not evident based on these data. Furthermore, the lack of response shift in each study condition rules out the possibility of differential response shift in the two groups.

Table 3-2. Fit indices, longitudinal one-factor CFA models

	CFI	RMSEA	$\chi 2$	*df*	*p*
Victimisation					
Control group (n = 1,581)					
Unconstrained model[†]	.977	.031			
Constrained model[‡]	.978	.029	28.1	18	.060
Intervention group (n = 1,919)					
Unconstrained model[†]	.960	.035			
Constrained model[‡]	.962	.032	29.3	18	.045
Perpetration					
Control group (n = 1,567)					
Unconstrained model[†]	.983	.020			
Constrained model[‡]	.984	.018	20.5	18	.303
Intervention group (n = 1,914)					
Unconstrained model[†]	.981	.020			
Constrained model[‡]	.981	.019	22.0	18	.231

Note. CFI = Comparative fit index; RSMEA= Root-mean-square error of approximation.

[†] Thresholds and factor loadings free at both time points, (both factor means fixed at zero and all scale factors fixed at one as identification constraints).

[‡] Thresholds and factor loadings specified as equal over time points, (time 1 factor mean zero and time 2 free, time 1 scale factors set to one and free for time 2).

3.4 Discussion

Although not tested, response shift has been suggested as one possible explanation for the lack of positive findings in evaluations of programs targeting aggression (Nixon & Werner, 2010; Orpinas et al., 2000; J. D. Smith et al., 2004). Based on analyses utilizing responses to retrospective pre-test questions and confirmatory factor analyses, however, little evidence of response shift – reconceptualisation, reprioritisation and/or recalibration – was found with regard to reports of bullying victimisation and perpetration in our study.

The lack of response shift in the intervention group is of particular interest, as this was the group where changes may have been expected. Explanations for the stability of the construct over time are likely varied and complex. It may be that by the age of 12-13 years, young people have an established understanding of bullying inclusive of the different forms (Varjas et al., 2008), or perceptions that are not conducive to change by a universal multi-faceted intervention such as the CFSP. Alternatively, our intervention may not have contained relevant or enough information to change the students' understandings of bullying in general (given the focus on cyberbullying), or the relevant components of the intervention may have been insufficiently implemented across schools to have had an impact on the participants' understandings of bullying. Preliminary analyses of process data show that implementation of the curriculum component of the intervention was largely limited to the first few modules.

The results from the factor analyses are encouraging as they demonstrate that scores on both versions of our FBS scale were time invariant. We believe a major contributing factor to this invariance is the format of the survey, where respondents are provided with a detailed definition which includes pictorial illustrations, of bullying behaviours. This approach would seem to have achieved its aim, namely as a means of calibrating the construct (Griffin & Gross, 2004; Solberg & Olweus, 2003), not only between individuals but also in achieving a degree of stability in understanding over time. This adds to the evidence of the sound psychometric properties of the Forms of Bullying Scale or FBS [citation deleted for blind review].

Our analyses based on the retrospective pre-test questions may, however, warrant further examination in future research studies of the potential for an intervention to impact on student reports of bullying. Whilst not significant, there was a four-fold difference in odds for bullying perpetration (equivalent to an effect size of 0.8), where intervention students

seemed more likely to consistently report having bullied others. This result may simply be an anomaly due to the small numbers of students reporting perpetration, however, the small sample size (*n*=50) also implies reduced power to achieve statistical significance for anything other than extremely large effect sizes. It is also unlikely to be subject to the biases associated with retrospective pre-test questions, such as effort justification or implicit theories of change bias. The retrospective and post-test questions were separated in the survey, thus the students are unlikely to have completed the two sets of questions in tandem with responses to one influenced by the other.

Comparisons of the traditional and retrospective pre-test reports are insightful. First, the stability of the reports for non-involved students from Grade 8 to Grade 9 (93%-96% in both study conditions "retained" their non-involved status), suggests that the intervention did not result in a group of "new" reporters of bullying involvement, as may have been expected (Nixon & Werner, 2010; Orpinas et al., 2000; J. D. Smith et al., 2004; P. K Smith et al., 2003). Our data do not therefore support the proposition that raised awareness and sensitisation may lead students to reconsider their behaviour or the behaviour of others and consequently report a greater frequency of bullying others/ being bullied. Second, the inconsistency, for those who initially reported involvement, towards retrospectively reporting no involvement in bullying victimisation and perpetration is interesting. Averaging across the study conditions, only 36% and 34% of those initially categorized as having been bullied / bullying others reported victimisation / perpetration in the retrospective tests. One possible explanation for the reversals is regression to the mean, i.e., those above average (where the average in this case is no involvement) at an initial measurement tend to revert back to more average values at a subsequent measurement. Another issue may be related to memory recall. Students completed the retrospective pre-test questions at the end of Grade 9, 15-18 months after the period on which they were reporting (Term 1 Grade 8). The phenomenon of satisficing may have been operating (Krosnick, 1991; Lam & Bengo, 2003) – when faced with a task which is cognitively demanding, respondents tend to choose the response which will suffice rather than exerting the effort to optimize their response, in this case the students may simply have chosen the most socially appropriate response. The magnitude of the drop-off, about 65%, raises questions as to the validity of retrospective pre-test questions as comparisons of post-test bullying involvement in determining change over time and assessing program impact.

The contrast in results from the CFA analyses applied to multi-item scales and the retrospective pre-test single-item questions highlights the advantages and disadvantages of each, and the likely merits of their combined use. Advantages of the CFA analyses are the ability to assess the different types of response shift and importantly, account for these in the model so that alpha or "true" change over time can easily be estimated, i.e., as the difference in mean of the latent variable at the two time points (Oort, 2005). A potential limitation of this approach is that differences at group level are tested and thus, a large number of individuals may need to experience a response shift for it to be detected (Ahmed et al., 2005; Oort, 2005). This may be a problem in evaluations of programs targeting problem behaviours such as bullying perpetration, as a majority of individuals remain uninvolved in the behaviours over time and the potential for response shift may be greater in the relatively small groups of frequently involved individuals who are the primary focus of the program. For example, if as a result of the intervention a response shift occurs in the subgroup of frequent perpetrators within the intervention group, without a similar change being present amongst the frequent perpetrators in the control group. This shift in reporting in a relatively small group may not be detected in CFA analyses, but may be sufficient to bias the estimates of the program effects on actual behaviours. The advantage of using single item traditional and retrospective pre-test questions is that responses to these can be categorized and shifts in subgroups of individuals identified, as was the case here. The responses of the non-involved students remained fairly stable over time, in contrast the responses of the students who initially reported involvement changed. In this study, the magnitude of the shifts within this subgroup of involved students did not differ significantly between the intervention and control groups. Had they differed, steps would need to be taken to account for the differential response shift to avoid bias in the estimation of the program impact. A summary of approaches to addressing response shift is given in Barclay-Goddard et al (2009).

Response shift may not be isolated to self-report and such changes could occur in peer-, teacher-, and parent-reports (P. K Smith et al., 2003). For example, teacher training and curriculum are common components of school anti-bullying programs. Training and delivery of program content may alter teachers' conceptualisation and / or awareness of bullying behaviour and thus change the standards by which they report on their students' behaviour. Similarly, peer reports can be subject to response shift due to changes in understandings of bullying behaviour or norms within the class or school as a consequence of an intervention. The shifts may also not be confined to an intervention condition, but

may also occur in a control group if schools in this study condition choose to implement anti-bullying strategies of their own.

Our findings are subject to certain limitations. Only non-government metropolitan secondary school students were included, and the sample was skewed toward families with higher socio-economic status – the generalisability of the results beyond this population is unknown. The extent, to which the results are applicable to administration modes other than online administration of questionnaires, is also unclear.

The results presented here must be interpreted within the context of our study. The CFSP intervention was targeted at cyberbullying prevention and so did not focus extensively on bullying behaviours more broadly. Programs targeting all modes of bullying may be more likely to induce response shift in scales measuring bullying in general. Whilst, the sample size in the project was large, the numbers of students who reported bullying perpetration were small and this impacted on the power of statistical tests within this group. Thus, our null findings should not be seen as evidence that it is unnecessary to consider this potential source of bias. Researchers should consider testing for response shift in their studies to ensure unbiased estimates of program effects are being obtained. This can be achieved through the application of statistical methods in the analysis of multi-item bullying scales such as conducted here and the inclusion of single-item retrospective pre-test questions in post-program surveys. In light of the different insights that may be gained from each approach, researchers could consider using more than one means of assessing response shift.

Chapter 4 : Bias in student survey findings from active parental consent procedures

Citation

Shaw, T., Cross, D., Thomas, L.T. & Zubrick, S. R. (Under review). Bias in student survey findings from active parental consent procedures. *British Educational Research Journal*.

Date submitted: 24 June 2013

Contribution of authors

The candidate was responsible for the literature review, conceptual framework, data analyses and discussion for this research paper. Professors Cross and Zubrick assisted in reviewing and commenting on the manuscript. Dr Thomas contributed to the early conceptualisation of the paper and review of the manuscript.

Relevance to the thesis

This chapter presents discussion and analyses central to Research Question 3 of this thesis. The potential for stringent parental consent procedures to result in biased samples and parameter estimates was explored, and the analyses and discussion informed the broader discussion in this thesis on threats to the validity of evaluations of school-based programs.

Abstract

Increasingly, researchers are required to obtain active (written) parental consent prior to surveying children and adolescents in schools. This study assessed the potential bias present in a sample of actively consented students, and in the estimates of associations between variables obtained from this sample. Students (n = 3,496) from 36 non-government metropolitan schools in Western Australia completed an online survey in 2010 as part of the Cyber Friendly Schools Project. Students with active (35%) and passive (65%) parental consent were compared on a range of variables including demographic, bullying and social-emotional outcomes. The moderating effects of consent status were also tested. Comparisons of the two consent groups showed that older students and students involved in problem behaviours such as bullying others, with lower pro-social scores, who lived with one parent and reported doing less well academically than their peers, were underrepresented in the sample with active parental consent. Additionally, consent status was a significant moderator of the associations between bullying victimisation and certain social-emotional variables. Active only parental consent leads to biased samples and biased estimates of associations between outcomes of interest, which could lead to miss-targeted health policies and interventions. Strategies to boost response rates to levels sufficient to warrant the conduct of the research are labour-intensive and costly, and the obtained samples are still likely to be biased. For low risk research, such as bullying surveys, rigorous active-passive consent procedures which result in higher participation rates, lower costs and reduced burden on teachers and schools, are recommended.

Keywords

Parental consent, bias, ethics, student survey

Acknowledgements

This research was made possible through funding from the Western Australian Health Promotion Foundation (Healthway). We thank the students, parents and staff in participating study schools as well as our colleagues at the Child Health Promotion Research Centre (CHPRC) at Edith Cowan University for their contributions to the research project.

4.1 Introduction

In 2013 in Australia all state government education departments and several non-government education sectors require the implementation of active parental consent prior to student participation in studies conducted by researchers external to the school. A similar requirement is mandated in education sectors in several other countries. Researchers conducting surveys in schools under these jurisdictions are therefore required to obtain signed consent from parents prior to their children's participation in a study. This is equivalent to an "opt-in" system. In contrast, procedures which include a passive consent phase, where parents are notified of the study intention and required to "opt-out" (hence non-response is deemed as implied consent), do not comply with these regulations. It is important to note that in most school-based evaluations of health programs, such as those addressing bullying or tobacco and drug use, consent from parents is only required for data collection as schools have the authority to implement programs deemed to be beneficial to students. This paper will explore the impact of required active parental consent on study findings when administering health-related surveys to students, particularly those aged between 10 and 16 years.

Educators and health practitioners are reliant on research evidence to determine the most effective policies and practices to enhance health-related behaviours in young people. Many research studies utilise student surveys to collect health outcome data and evaluate program impact, and the participation of young people in such research is indispensable. In Australia research involving humans is conducted within the guidelines of the National Health and Medical Research Council or NHMRC National Statement on Ethical Conduct in Human Research (hereon referred to as the "National Statement"). The statement specifies that, in addition to obtaining consent from the minor whenever he or she has the capacity to grant it, parental consent is needed when surveying minors, particularly young people of "developing maturity" and children (National Health and Medical Research Council, Australian Research Council, & Australian Vice-Chancellors' Committee, 2007). The National Statement does not, however, distinguish between active and passive forms of parental consent.

Consent procedures vary across studies from passive only to active only procedures including combinations of the two, with no or different degrees of follow-up. For example, typically in active-passive procedures parents are approached once or twice for active consent, followed by one to two further contacts seeking passive consent from non-

respondents. Thus, parents are first asked to opt-in (or opt-out), and then if they do not respond, are informed that consent will be assumed unless they opt-out indicating their non-consent. Response rates vary according to the form of consent required, the extent and mode of contacts with parents and as well as other factors such as the use of incentives (Wolfenden, Kypri, Freund, & Hodder, 2009).

Participation rates in health-related studies under passive consent conditions are often above 90% (Ellwood, Asher, Stewart, & ISAAC Phase III Study Group, 2010; Frissell et al., 2004; Mellor, Rapoport, & Maliniak, 2008; Tigges, 2003), i.e. in effect a minority of parents actively refuse permission for their child to participate. In comparison, in studies conducted in the last 10 years, participation rates under active consent conditions ranged between 27% - 76% (Courser, Shamblen, Lavrakas, Collins, & Ditterline, 2009; Ellwood et al., 2010; Frissell et al., 2004; Mellor et al., 2008; Unger et al., 2004). Participation rates dropped from over 90% to below 40% when consent procedures were switched from passive to active consent in Australian samples in the International Study of Asthma and Allergies in Childhood (ISAAC) (Ellwood et al., 2010). In the Australian Covert Bullying Prevalence Study (ACBPS) the mean consent rates were 36% when active consent was required and 96% when using an active-passive procedure (Cross, Shaw, et al., 2009).

Higher participation rates of 80% (Esbensen, Melde, Taylor, & Peterson, 2008) and 95% (Secor-Turner, Sieving, Widome, Plowman, & Vanden Berk, 2010) have been achieved under active only consent conditions, but with considerable expenditure of resources by the research team and schools to obtain data on consent from parents. The financial cost was estimated at US$7 per participant in one study (Esbensen et al., 2008) and in another, non-responding parents were contacted individually by telephone and incentives were provided, including movie tickets per responding family (Secor-Turner et al., 2010).

Comparisons of students with active only consent to those with other forms of parental consent found the active consent group constituted a biased sample under-representing students who were male (Courser et al., 2009; Dent et al., 1993; Unger et al., 2004), older (Courser et al., 2009), less academically competent (Henry, Smith, & Hopkins, 2002; Unger et al., 2004), from minority groups (Dent et al., 1993; Tigges, 2003; Unger et al., 2004), who were absent more often from school (Henry et al., 2002; Secor-Turner et al., 2010), less involved with their parents and less likely to live with both parents (Dent et al., 1993). Students of non-responding parents were also more involved in and more at risk of involvement in problem behaviours such as aggression, tobacco use and alcohol and other

drug use (Courser et al., 2009; Dent et al., 1993; Tigges, 2003; Unger et al., 2004; White, Hill, & Effendi, 2004), and more likely to be overweight or at risk of being overweight (Mellor et al., 2008).

While limited information is available on the characteristics of non-responding parents, available research shows that 87% of non-responders to a passive consent process had received and understood the materials, and had consciously decided to allow participation (Ellickson & Hawes, 1989). Thus, non-response was more likely to indicate consent than refusal (Ellickson & Hawes, 1989). While non-responding parents were more likely to be employed, their attitudes to research were more similar to consenting parents, than those who refused consent (Baker, Yardley, & McCaul, 2001).

Passive or opt-out consent procedures are recommended for low-risk research to avoid non-response bias (Lacy et al., 2012; Stubbs & Achat, 2009). Research is defined in the National Statement as "low-risk" if the "only foreseeable risk is one of discomfort" (National Health and Medical Research Council et al., 2007). Limited studies have assessed the impact on young people of completing health surveys. In a study by Langhinrichsen-Rohling and colleagues (Langhinrichsen-Rohling, Arata, O'Brien, Bowers, & Klibert, 2006), 4.4% of students often felt upset while completing a survey which included questions on sensitive topics such as suicidal behaviour and physical and sexual abuse. Previous experience of sensitive events in their lives predicted students' level of emotional response to completing the survey. However, no adverse consequences were observed or reported during data collection in the three years of the study. Furthermore, the passive consent group were no more likely to report feeling upset than the active group, indicating no increased risk for students without active parental consent. In a two-year smoking prevention trial no harm to students from participation was observed, and no parental complaints of negative outcomes were received as a result of their child's participation (Leakey, Lunde, Koga, & Glanz, 2004).

Motivation and aim of this study

Two studies conducted by the Child Health Promotion Research Centre (CHPRC) at Edith Cowan University investigating bullying behaviours in schools in Western Australia recruited students in government schools using active parental consent procedures as mandated by the Department of Education (Department of Education. Western Australia, 2009). Parental consent rates for government school students were 15% and 18% respectively, with 5% and 7% of parents respectively refusing permission (Cross, Brown,

Epstein, & Read, 2009; Hall, Cordin, Bruce, & Paki, 2011). These low response rates were obtained despite using numerous strategies to increase parental participation in the consent process. This included up to three rounds of follow-up, the provision of small incentives and reply paid envelopes for the return of consent forms and multiple methods of contacting parents, including information letters and consent forms being sent directly to home addresses by the schools (Cross, Brown, et al., 2009; Hall et al., 2011). In light of these low rates, the aim of this paper was to explore the potential for bias in samples recruited under active only consent conditions, on a range of demographic, bullying and social-emotional variables. We also assessed the potential impact of these biases on the associations between bullying and other outcomes. Students with active and passive parental consent were compared, utilising a sample recruited in non-government schools using an active-passive consent procedure.

While the focus in this paper is on parental consent, the need to ensure a young person's participation is voluntary and as informed as possible, and to obtain his/her consent is also fully recognised.

4.2 Methods

Procedure

The Cyber Friendly Schools Project was a group-randomised intervention trial conducted in secondary schools in Western Australia in 2010-2012 (Child Health Promotion Research Centre, 2010). All Perth metropolitan non-government secondary schools were approached for participation and 36 (68%) agreed. Most of the participating schools were co-educational (n=25), with seven girls only and four boys only schools, whilst 16 were Catholic and 20 other non-government schools. All Year 8 students in recruited schools were eligible for participation and the recruited cohort was followed until the end of Year 10. The baseline 2010 data were analysed here.

An active-passive consent process was followed to obtain consent from the Year 8 students' parents and caregivers for the students to complete a survey. Thus, each school was provided with stamped, pre-packaged envelopes for school staff to attach address labels and mail to the students' home addresses. The envelopes contained an information letter describing the study and requesting active consent for the child's participation, a consent form and reply paid envelope. Parents who had not responded after two weeks

were mailed a follow-up letter requesting passive consent for their child to participate in the study, a consent form which allowed the parent to decline participation, and reply paid envelope to return the form if they did not want their child to participate. Through this process steps were taken to firstly obtain active consent and then active refusal, and if no response was received from the parent in any of the phases, parental consent was deemed to be passively present. Parents were asked to discuss the research with their child prior to signing the consent forms. Student consent was obtained prior to survey administration, students were informed of the purpose of the survey and that their participation was voluntary.

Students completed online surveys in Term 2, 2010 in controlled conditions during their normal classroom lessons. The surveys were administered by trained research staff according to a strict procedural protocol. Students were assured of the confidentiality of their responses. At the completion of the survey, all students were encouraged to speak with an adult they trust and received information regarding help they could access confidentially, should the survey raise any concerns for them.

Ethics approval for the study was obtained from the Human Research Ethics Committee at Edith Cowan University and the Catholic Education Office of WA. Independent schools affiliated with the Association of Independent Schools of WA independently provided consent.

Measures

Demographic variables

Individual demographic variables included gender, age (12, 13, 14 years), self-reported academic performance, living with one parent or two parents/adults (used as a proxy measure for family socio-economic status). School variables included school sector (Catholic, Independent) and school type (co-educational, girls only, boys only).

Social-emotional outcomes

Symptoms of depression and anxiety were assessed using the 14 item Depression Anxiety Stress Scale (DASS). Items had four response options ranging from 'Does not apply to me' to 'Most of the time'. The validity of DASS scores has been shown previously (Szabó, 2010). Mean scores were calculated, with higher scores reflecting greater symptoms of depression (α = .92) and anxiety (α = .82).

The five subscales of the Strengths and Difficulties Questionnaire (SDQ) assessed emotional symptoms, conduct problems, peer problems, hyperactivity and pro-social behaviour (Goodman, 1997). Each subscale comprised five items with response options 'Not true', 'Somewhat true', 'Certainly true'. Means were obtained for each subscale, with higher scores representing greater levels of emotional symptoms (α=.73), conduct problems (α=.55), peer problems (α=.54), hyperactivity (α=.68) and pro-social behaviours (α=.72), and a total difficulties score calculated using four of the subscales (α=.80).

The peer support scale (11 items, e.g. "How often would other students invite you to do things with them") was adapted from the Perceptions of Peer Social Support Scale (Ladd, Kochenderfer, & Coleman, 1996). The response options were 'Lots of times', 'Sometimes', 'Never'. The higher the mean score the greater the respondent's perception of support from his/her peers (α=.86).

Connectedness to school was measured using a five item scale (e.g. "I feel close to people at my school") adapted from the scale of Resnick and McNeely (Resnick et al., 1997), with response options ranging from 'Never' to 'Always'. Mean scores corresponded to higher connectedness (α=.78).

Bullying outcomes

Involvement in different forms of bullying victimisation and perpetration were measured using the respective versions of the Forms of Bullying Scale (FBS) (Shaw, Dooley, Cross, Zubrick, & Waters, In press) and cyberbullying victimisation and perpetration using two 11 item scales. The scales were preceded by definitions of different forms of traditional bullying and cyberbullying, including examples and illustrations. The bullying and cyberbullying definitions were adapted from those developed by Olweus (Solberg & Olweus, 2003) and Smith (Smith et al., 2008) respectively. The response options correspond to the frequency of bullying within the previous term (10 weeks): 'This did not happen to me/I did not do this'; 'Once or twice'; 'Every few weeks'; 'About once a week'; and 'Several times a week or more'. Higher mean scores represented greater involvement (victimisation: α=.87; perpetration: α=.85; cybervictimisation α=.86; cyberperpetration α=.91).

Consent status

Students were categorised as having active parental consent if their parents returned a written consent form agreeing to their child's participation, or passive parental consent if they did not return a form indicating either consent or non-consent.

Statistical analyses

Logistic regression was applied in Stata 12.0 to identify predictors of consent status to determine differences on the demographic, social-emotional and bullying variables between the consent groups. The impact of consent status on associations between the bullying and social-emotional variables was assessed by testing for interaction effects between consent status and the social-emotional variables on the bullying outcomes using tobit regression, as the bullying variables were highly skewed with percentages of 28% or more at the minimum value. Random intercepts were included in all models to account for school level clustering.

4.3 Results

At baseline, 7% of parents of the Year 8 cohort returned forms indicating they did not consent to their child's participation. In total, 3,496 students completed surveys (91% of the students with consent), 35% with active parental consent and 65% with passive consent. The sample comprised 52% female students, most were 13 years of age (85%), lived in a household with two parents/adult care-givers (86%) and attended co-educational schools (72%). About half attended non-Catholic non-government schools (51%). A minority perceived themselves to be academically less competent than their peers (9%). Table 4-1 describes the percentages of students with active consent within each demographic group, e.g. relatively fewer male (34%) than female (37%) students had active consent.

Table 4-1. Active parental consent rates by demographic group

Demographic variable	Active consent rate
Gender	
Male	34%
Female	37%
Age	
12 years	42%
13 years	35%
14 years	25%
Family structure	
Single adult family	31%
Two adult family	36%
Academically compared to peers	
Worse	30%
Same or better	36%
School type	
Co-educational	34%
Girls only	40%
Boys only	39%
School sector	
Catholic	31%
Other non-government	40%

Note: Sample size varies between 3,433 and 3,495 due to missing values

Consent status was not predicted by gender or school type, but the parents of older students (p = .014), students living with one parent/adult care-giver (*OR* = 0.80, 95% CI 0.65 - 0.98), attending Catholic schools (*OR* = 0.70, 95% CI 0.57 - 0.85) and who reported they did not perform as well as their peers at school (*OR* = 0.76, 95% CI 0.58 - 0.98), were less likely to provide active consent (

Table 4-2). Age and sector remained as independent predictors in multivariable analyses, and academic status was on the border of statistical significance.

Table 4-2. Demographic predictors of active parental consent

	Unadjusted odds ratio	(95% CI)	Adjusted odds ratio[a]	(95% CI)
Girl	1.16	(0.99-1.36)		
Age				
12 years	1.00	*p*=.014	1.00	*p*=.023
13 years	0.78	(0.63-0.96)	0.81	(0.66-1.00)
14 years	0.51	(0.30-0.88)	0.50	(0.29-0.87)
Single adult family	0.80	(0.65-0.98)		
Academically worse than peers	0.76	(0.58-0.98)	0.77	(0.59-1.00)
School type				
Co-educational	1.00	*p*=.127		
Girls only	1.29	(0.97-1.73)		
Boys only	1.26	(0.89-1.79)		
Catholic school	0.70	(0.57-0.85)	0.71	(0.58-0.87)

[a] Multivariable model including age, academic status and school sector; 'Single adult family' dropped as not significant in multivariable model.

No association was found between measures of student mental health (depression, anxiety or emotional symptoms) and consent status (Table 4-3). However, students who reported higher levels of conduct problems (*OR* = 0.76, 95% CI 0.61 - 0.95) and those who reported having bullied others (*OR* = 0.68, 95% CI 0.50 - 0.94), had significantly lower odds of active parental consent and thus, proportionally fewer of these students would be included in this group. In comparison, the odds of active consent increased with higher scores on the pro-social scale (*OR* = 1.37, 95% CI 1.13 – 1.65). Only pro-social skills remained as clearly significant (*OR* = 1.38, 95% CI 1.14 – 1.67) after controlling for the demographic variables identified as independent predictors in the multivariable model reported in

Table 4-2.

Table 4-3. Socio-emotional and bullying variables as predictors of active parental consent

	Unadjusted odds ratio	(95% CI)	Adjusted odds ratio[a]	(95% CI)
Social-emotional variables				
Depression	1.02	(0.90-1.16)		
Anxiety	1.01	(0.84-1.22)		
SDQ Emotional symptoms	0.92	(0.79-1.08)		
SDQ Conduct problems	0.76	(0.61-0.95)	0.79	(0.63-1.00)
SDQ Hyperactivity	0.95	(0.81-1.12)		
SDQ Peer problems	0.93	(0.74-1.15)		
SDQ Pro-social	1.37	(1.13-1.65)	1.38	(1.14-1.67)
SDQ Total difficulties	0.81	(0.62-1.06)		
School connectedness	1.11	(0.99-1.26)		
Peer support	1.02	(0.85-1.24)		
Bullying variables				
Victimisation	1.00	(0.88-1.14)		
Perpetration	0.68	(0.50-0.94)	0.72	(0.52-1.00)
Cyber victimisation	0.90	(0.67-1.21)		
Cyber perpetration	0.72	(0.40-1.28)		

[a] Adjusted for age, academic status and school sector

Consent status was found to be a significant moderator of the associations between three predictors and victimisation, with significant interactions identified in the tobit regressions between consent status and school connectedness (z = 2.46, p = .014), emotional symptoms (z = 2.67, p = .008), and the SDQ total difficulties score (z = 2.21, p = .027) respectively. These interactions indicate the values for the measures of association between the social-emotional variable and victimisation differed significantly between the two consent groups. The associations between each of school connectedness, emotional symptoms and the SDQ total difficulties scores and victimisation were significant within each consent group but stronger, in terms of larger regression coefficient values, in the group of students with active consent than those with passive consent. Therefore, these associations in the general student population are over-estimated based on the active

consent only group. No moderating effects were found for the other bullying outcomes, namely perpetration, cybervictimisation or cyberperpetration.

4.4 Discussion

Study findings

In accordance with the findings of others (Courser et al., 2009; Dent et al., 1993; Henry et al., 2002; Tigges, 2003; Unger et al., 2004; White et al., 2004) this study's results demonstrate that students with active parental consent differ systematically from the larger student population. Students involved in problem behaviours and with lower pro-social scores were less represented in this group. Also, older and less academically competent students, and those living with one parent/adult care-giver were less represented.

Students in single parent families may be less likely to have forms returned due to the increased pressures on these households, while parents of older students may be less engaged with their child's school. More pro-social students presumably belong to families where these attitudes are valued, hence their parents may be more likely to return consent forms. It is unclear why parents of students in Catholic schools were less likely to return consent forms. School SES has been found to be a predictor of active consent rates (Esbensen et al., 2008) and the higher socio-economic status of several of the non-Catholic schools may explain their higher response rates. The lower participation of students engaged in problem behaviours, suggests higher risk students who are most able to inform and benefit from the research, may be excluded.

Stronger associations were present in the active consent group between bullying victimisation and school connectedness, emotional symptoms and the SDQ total difficulties score, than in the group with passive consent. Thus, results based on the active consent only group would overestimate the associations in the wider student population. This has implications for studies aimed at estimating associations and effects, as the estimates may be biased (Esbensen et al., 2008). Although the correlations based on the active consent group were overestimated in this study, it is unclear in which direction the bias may operate for different outcomes and different parental response rates. Critically, when assessing the effects of health programs, if students engaged in the higher risk behaviours targeted by the program, i.e. those with the most potential to shift their behaviours as a

result of the program, are under-represented in the sample, it is likely that the impact of effective programs will be underestimated. Further information on the impact or otherwise of active versus passive parental consent on study findings in health-related studies will be provided by a Cochrane review which is currently underway (Priest et al., 2012).

Costs and benefits

Mandating active parental consent has costs and consequences which need to be weighed against the benefits of increased participation rates and resultant greater validity of research findings, as well as a more equitable distribution of the research burden, gained by allowing passive parental consent. The potential harm experienced by young people from completing surveys is a further consideration.

Various strategies are proposed for improving response rates under active consent conditions (Secor-Turner et al., 2010; Tigges, 2003; Wolfenden et al., 2009), but these were costed at between US$7 and US$32 per completed survey (Esbensen et al., 2008; Tigges, 2003). Whilst rigorous passive consent procedures also incur costs, high response rates under active only conditions require expensive strategies such as telephoning non-responding parents, repeated visits to each school by researchers and incentives (Esbensen et al., 2008; Secor-Turner et al., 2010; Unger et al., 2004). A further cost may be the increased sample size required to account for higher intraclass correlations resulting from greater homogeneity between actively consented students (Shaw & Cross, 2012; White et al., 2004). Researchers, conducting research requiring large samples, will have difficulty funding these additional costs with limited budgets. An opportunity cost is also present where increased expenditure for improved response rates means less available funding for research activities such as intervention development and implementation. Furthermore, these costly measures taken to improve active response rates, still do not guarantee unbiased samples, even with participation rates as high as 79% (Esbensen et al., 2008) and 76% (Unger et al., 2004). At issue is the extent to which participation is associated with the outcomes of interest in the study (Courser et al., 2009), the stronger this association and the smaller the percentage of responding parents, the greater the likelihood of a biased sample.

Some strategies used to increase response rates may not be possible. . For example, in Australia, researchers are unlikely to be given parents' personal contact details to directly follow-up non-responding parents. Moreover, the incentives to participation provided by

others (Esbensen et al., 2008; Secor-Turner et al., 2010), may be seen as inducements or coercive, and not approved by research ethics committees.

The National Statement refers to the need for the research burden to be equitably distributed (National Health and Medical Research Council et al., 2007). Under active only consent conditions, an added burden may be imposed on certain schools if research is found to be impracticable in others. There may also be an added burden on certain students for data collection. Further, increased time and effort is often asked of teachers as their role is critical in obtaining consent forms from parents. Indeed, some schools have chosen to use passive consent procedures because of the extra time and effort required by school staff to secure active consent (White et al., 2004).

Research and anecdotal evidence indicates a low probability of harm to students as a result of completion of health surveys (Langhinrichsen-Rohling et al., 2006; Leakey et al., 2004). This is the experience of the research team at the Child Health Promotion Research Centre, with no adverse student outcomes from survey completion observed or reported in the conduct of ten large-scale projects in schools within the last 12 years. Whilst it is important to implement adequate measures to assist the small number of students who may be impacted negatively, current evidence suggests that completing health surveys (especially those without highly sensitive items) is unlikely to lead to distress or other harm to students, and that this type of research can be viewed as low-risk (National Health and Medical Research Council et al., 2007).

Use of active-passive procedures

Despite the cautions against mandating active only consent presented here, the aim is not to promote the use of passive only procedures, i.e. where parents are approached once and asked to return the consent form only if permission is refused. Instead an active-passive approach is recommended with two to three rounds of active consent followed by a round of passive, given the limited benefits gained beyond three follow-ups (Ji, Pokorny, & Jason, 2004). Given evidence that non-response to a passive consent process more likely indicates consent than refusal (Ellickson & Hawes, 1989), an opt-out process which provides adequate opportunities for parents to indicate their non-consent allows for informed individual decision making and maintains parents' autonomy. Importantly, participants recruited using any of these consent processes are entitled to withdraw from the study at any time.

It is imperative that all procedures which include a passive component are appropriately designed and implemented to meet the ethical research requirements of voluntary participation based on sufficient information (National Health and Medical Research Council et al., 2007). Interpreting non-response as consent assumes parents have received and understood the information, have had multiple opportunities to refuse their child's participation and the methods of returning forms indicating non-consent are as infallible as possible. Researchers should therefore make every effort to ensure parents not only receive, but engage with and understand the information (Ellickson & Hawes, 1989), and are provided with the opportunity to discuss the research with either the researchers or an independent body such as a representative from a university ethics committee (National Health and Medical Research Council et al., 2007).

Limitations

This study is subject to limitations. Only non-government metropolitan schools in Western Australia were included, however the results are likely to be broadly generalisable to other contexts. The data are self-reported, and bias will be introduced into the results if the validity of the self-reported data differs between the actively and passively consented groups. Whilst statistically significant, the observed effects are small. However, they are in accordance with biases reported in the literature and could be expected to be larger in samples with lower active consent rates than observed here. Further research on a range of health outcomes is required to gain a better understanding of the biases resulting from active only consent procedures, the consent status of non-responding parents and the impact on young people of completing health surveys.

4.5 Conclusion

Education authorities and ethics committees who mandate active parental consent prior to the participation of minors in research, principally to avoid the risks of harm to the minor, need to be cognisant of the consequences and opportunity costs of this requirement, relative to the benefits of allowing procedures including passive consent, particularly for low risk research. Researchers and funding bodies need to factor in the additional resources needed, both financial and time, to secure good response rates if active parental consent is required. Given the potential for response rates lower than 40%, even with additional resource expenditure, the inferences drawn from such research may lead to

miss-targeted policies and interventions. Moreover, the conduct of research under active consent conditions may simply be impracticable. Active-passive consent procedures result in higher participation rates, lower costs, and reduced burden on teachers and schools, and are recommended when parental consent is sought for minors to complete low risk health surveys.

Chapter 5 : The clustering of bullying and cyberbullying behaviour within Australian schools

Citation

Shaw, T., & Cross, D. (2012). The clustering of bullying and cyberbullying behaviour within Australian schools. *Australian Journal of Education, 56*(2), 142-162.

Date submitted: **10 February 2012**

Date accepted: **15 March 2012**

Contribution of authors

The candidate was responsible for the literature review, conceptual framework, data analyses and discussion involved in this research paper. Professor Cross assisted in reviewing and commenting on the manuscript.

Relevance to the thesis

This chapter presents discussion and analyses central to Research Question 4 of this thesis. Methods for the calculation of sample sizes in evaluations of school-based programs were presented, and the results and discussion informed the broader aim of this thesis by adding to the body of knowledge around the rigorous conduct of school-based program evaluations.

Abstract

Bullying between students at school can seriously affect students' health and academic outcomes. To date little is known regarding the extent to which bullying behaviours are clustered within certain schools rather than similarly prevalent across all schools. Additionally, studies of bullying behaviours in schools that do not account for clustering of such behaviours by students within the same school are likely to be under-powered and yield imprecise estimates. This paper presents intraclass correlation (ICC) values for bullying victimisation and perpetration measures based on a large representative sample of 106 Australian schools. Results show that bullying is not confined to specific schools and school differences contribute little to explaining students' bullying behaviours. Despite this, seemingly negligible ICC values can impact substantially on the sample sizes required to attain sufficiently powered studies, when large numbers of students are sampled per school. Sample size calculations are illustrated.

Keywords

intraclass correlation, bullying, sample size, cluster sampling, group-randomised trials

Acknowledgments

The Australian Covert Bullying Prevalence Study was funded by the Australian Federal Department of Education, Employment and Workplace Relations. We would like to thank the students, parents and staff at participating schools and Melanie Epstein and other staff at the Child Health Promotion Research Centre (CHPRC) at Edith Cowan University for their contributions to the Australian Covert Bullying Prevalence Study. We would also like to thank the editor and reviewers, as well as Professor Stephen Zubrick, for helpful comments which have resulted in improvements to the paper.

5.1 Introduction

Bullying is defined as aggressive behaviour repeated over a period of time, characterised by a real or perceived imbalance of power perpetrated with the intent to harm the target (Olweus, 1996). Bullying between students at school can seriously affect the social, physical and psychological well-being as well as the academic achievement of both the perpetrators and those who are victimised (Arseneault, Bowes, & Shakoor, 2010; Kaltiala-Heino, Rimpelä, Marttunen, Rimpelä, & Rantanen, 1999; Kaltiala-Heino, Rimpela, Rantanen, & Rimpela, 2000; Nansel et al., 2001; Wolke, Woods, Bloomfield, & Karstadt, 2001). The Australian Covert Bullying Prevalence Study found that just over one quarter (27%) of Australian school students aged 8 to 14 years reported being frequently bullied and 9% reported frequently bullying others (every few weeks or more often) (Cross et al., 2009). Approximately 7% of students in Years 4 to 9 reported being cyberbullied every few weeks or more often in their last term at school (Cross, et al., 2009). While cyberbullying occurs with less frequency than traditional bullying, its prevalence is still appreciable and possibly increasing in Australia, as elsewhere in the world (Smith & Slonje, 2009). The high prevalence of school bullying and its significant detrimental effects have prompted, especially in recent years, much research to better understand this behaviour and to intervene to reduce harm associated with bullying (Farrington & Ttofi, 2009).

Research on behavioural phenomena amongst school students, such as bullying, must take account of the clustering of students within schools. This is due not only to the study designs used but also to the contextual influences on the variables of interest. Firstly, cluster sampling designs, where schools are selected in the first stage of sampling and individuals in the second, are often used in studies of young people, as schools facilitate access to the target population and survey administration (Carlin & Hocking, 1999; Heeringa, West, & Berglund, 2010). Secondly, students' experiences of bullying at school or within other contexts are dependent on the behaviours and norms within the particular group, for example, a number of students in a school may be victimised by the same perpetrator. Additionally, young people's behaviour (and particularly problem behaviours) may be influenced by their peers (Dishion & Owen, 2002; Kiesner, Dishion, & Poulin, 2001). Furthermore, in intervention research wide use is made of group-randomised trials, in which whole groups such as schools are randomised to conditions and interdependence of outcomes exist, particularly if the interventions have a whole-of-school focus.

The consequences of these clustering factors are that students within a particular school will be more alike with regard to bullying behaviours than students from different schools. This homogeneity within schools is measured by the intraclass correlation (ICC). In a two-level design with students nested within schools, the ICC can be interpreted as the extent to which students from the same school are more similar than students from different schools. The ICC is calculated as the ratio of the variation between schools relative to the total variation (at school and individual level) in the variable of interest such as bullying between students. ICC values vary between zero and one. Greater variation or differences between schools implies greater similarities within schools and hence a higher ICC value (Twisk, 2006). An ICC of zero would imply no variation between schools, i.e. the variation in bullying outcomes of students aggregated within schools is equal to the variation among students across all schools. At the other extreme, an ICC of one (an unlikely value) would mean that all of the variation between students is due to school differences (i.e. there are no differences within schools).

Intraclass correlation values vary according to the outcome measure for which the value is calculated and the study population (Carlin & Hocking, 1999; Murray, 1998). Additional factors such as the time of the year of the survey and the gender, ethnicity and year level of the students can also influence the size of the correlation (Murray et al., 1994; Resnicow et al., 2010; Siddiqui, Hedeker, Flay, & Hu, 1996). Intraclass correlations have been found to be below .1 in value for a range of health outcomes for school-based data, including for measures of tobacco (Murray, et al., 1994; Siddiqui, et al., 1996), alcohol and other drug use (Carlin & Hocking, 1999; Murray, Clark, & Wagenaar, 2000; Scheier, Griffin, Doyle, & Botvin, 2002), nutritional intake (Murray, Phillips, Birnbaum, & Lytle, 2001) and physical activity (Murray et al., 2006). Hutchison (2004) compared ICC values across a range of variables for primary and secondary students in schools in the Third International Mathematics and Science Study (TIMSS) survey in the UK and Wales (Hutchison, 2004). Results showed the strongest clustering effects for the various educational outcomes (mean ICC=.18) and ethnic variables (mean ICC=.14) compared with demographic variables (mean ICC=.02), leisure activities (mean ICC=.04) and home characteristics relevant to educational attainment (mean ICC=.04). Australian data from the 2009 Programme for International Student Assessment (PISA) study similarly produced ICC values of .2 for each of Reading, Mathematics and Science scores (OECD, 2010).

Limited data are available on ICC values for bullying outcomes and thus, the extent to which a culture of bullying may be stronger in certain schools compared with others. In 2009 Bradshaw et al (Bradshaw, Sawyer, & O'Brennan, 2009) reported an ICC for victimisation amongst elementary school students in a district in Maryland, USA of 0.019 and for middle school students for victimisation as 0.006 and bullying others as 0.009. The evaluation of the KiVa Antibullying Program conducted in Grades 4-6 in 78 schools in Finland found school-level ICC values of .02 for both victimisation and perpetration (Kärnä et al., 2011). Two other cluster-randomised intervention trials, presented the average and ranges, rather than specific values for individual outcomes, of school-based ICC values. Fonagy and colleagues (Fonagy et al., 2009) reported an average ICC of .04 for the bullying-related outcomes, which included measures of aggression and victimisation, in their study of third to fifth grade students in nine elementary schools in a city of the Mid-West, USA. Australian data from the Gatehouse Project (Bond et al., 2004), conducted amongst secondary students in 26 schools in Victoria, found ICC values between .01 and .06 for a range of emotional well-being and drug use outcomes, including bullying victimisation.

A second motivation exists for estimation of ICC values for bullying outcomes. When planning school-based studies, account needs to be taken of the impact of the non-independence of the subjects in calculating required sample sizes or underpowered studies will result (Murray & Short, 1997). The homogeneity within clusters results in cluster samples having a smaller effective sample size in terms of the precision with which parameters are estimated and hence the power to detect a statistically significant difference, than a simple random sample of the same number of subjects (Heeringa, et al., 2010). The design effect (deff) is a measure of the performance of a complex sampling design (such as cluster sampling) compared to what would be achieved with simple random sampling. A means of determining the sample size required for a cluster sample with sufficient power is to incorporate the design effect in the calculations. The design effect for cluster sampling depends on the cluster sizes and the strength of the correlation within the clusters, hence the need for estimates of ICC values.

The type of analyses conducted influences the operative ICC value i.e. the value that will be operating when the results of the study are determined (Carlin & Hocking, 1999; Murray, 1998) and hence the value to be accounted for in sample size calculations. In longitudinal or pretest-posttest studies, repeated measures analyses or adjustment for the baseline value of the outcome measure can result in substantive reductions in operative ICC values

compared to cross-sectional designs (Murray & Blitstein, 2003). Furthermore, the addition of explanatory variables or covariates in statistical models can reduce the impact of the ICC and increase the power of a given analysis, if the variables reduce the variation between schools (Murray, 1998; Snijders & Bosker, 1999). School level variables correlated with the outcome of interest are the most likely to reduce this variation.

To our knowledge to date there are no Australian data published that describe ICC values for bullying outcomes and there are no published data from Australia or elsewhere describing ICC values for cyberbullying outcomes. This paper presents ICC values for self-reported bullying victimisation and perpetration measures, including cyberbullying, based on a representative sample of Australian school students. The aim is to explore the extent to which differences in students' victimisation and perpetration of bullying behaviours are related to the school they attend i.e. the extent to which bullying is clustered within specific schools. Differences between demographic groups are illustrated. A secondary aim is to determine the impact of clustering effects on sample size requirements to assist researchers to plan cross-sectional surveys or group-randomised intervention trials for bullying outcomes. Guidelines for the calculation of sample sizes are discussed.

5.2 Method

Schools and Participants

The Australian Covert Bullying Prevalence Study (ACBPS) included a cross-sectional survey of Years 4 to 9 students (typically 9-15 years of age) conducted in Term 4 in 2007 in 106 schools (55 primary and 51 secondary, 46% response rate) (Cross, et al., 2009). A stratified (by State/Territory and location) cluster sampling design was utilised, with schools randomly sampled in the first and classes in the second stage of sampling. Because of differences in school structures between states in Australia, in four of the eight States/Territories (the Australian Capital Territory, New South Wales, Tasmania and Victoria) Year 7 students were drawn from secondary schools and from primary schools in the remaining three States (Western Australia, Queensland and South Australia) and the Northern Territory. The sampling population included all schools in Australia other than non-mainstream schools, those in remote areas and schools with less than 30 students in 2007 in each of the sampled year levels. Students with a disability which prevented them from completing the hard copy questionnaire were not included in the sample.

Consent was sought from parents of all students in selected classes with information and letters mailed directly to parents. Reply paid envelopes for the return of consent forms were provided. Active parental consent was required in Government schools in certain States/Territories (36% consent rate), and an active/passive consent procedure (where participants actively opt out) was used in all other instances (96% consent rate).

Students completed hard copy questionnaires in their classrooms administered by school staff according to a strict procedural and verbal protocol. The questionnaire was read aloud to Year 4-6 students. Alternative learning activities were provided for students without parental consent and those who declined or were unable to participate. Of the 8782 students with parental consent, useable surveys were returned from 7418 students (84%). The sample comprised 52% female students (n=3874), 37% (n=2779) in Government schools, 64% (n=4760) in metropolitan areas and between 14% and 19% in each of Years 4 to 9. Students ranged in age from 8 to 16 years (mean=12, SD=1.7 years).

Measures

Bullying victimisation and perpetration were measured using single items and scales. Consistent with previous research (Solberg & Olweus, 2003), students were provided with a definition of bullying as repeated behaviour which happens "to someone who finds it hard to stop it from happening", together with examples of different forms of bullying. All victimisation and perpetration questions referred to the previous term at school (past 10 weeks) and had specific response options i.e. "was not bullied/did not bully", "once or twice", "every few weeks", "about once a week" and "most days". In each instance, the same items were used for victimisation and perpetration, with wording adjusted appropriately.

Any type of bullying

Items adapted from the Olweus Bully\Victim Questionnaire (Olweus, 1996) and the Rigby and Slee Peer Relations Questionnaire (Rigby, 1998) were used to measure any type of victimisation ("This term, how often were you bullied again and again by another student or group of students") and perpetration. Test-retest reliability of these items was moderate (n=140, K_w=.54 and K_w=.45 respectively). Students were categorised as having been bullied/bullied others if they indicated the victimisation/perpetration occurred every few weeks or more frequently in the previous school term (Solberg & Olweus, 2003). Two 12 item scales were included to measure victimisation and perpetration of different forms of bullying, namely verbal, exclusion, social (eg. spreading rumours), physical and threatening

bullying behaviours and the extent of victimisation/perpetration. Cronbach's alphas for these scales were .91 and .88 respectively. A mean score (0-4) was calculated for each scale.

Cyberbullying

Cyberbullying perpetration and victimisation behaviours were measured using two 8 item scales addressing bullying behaviours perpetrated via mobile phone, email and/or the Internet (e.g. "sent nasty messages on the internet", "mean or nasty comments or pictures posted to websites", "ignored or left out of things over the Internet"). Cronbach's alphas of .86 and .88 were found for the cyber victimisation and perpetration scales respectively. Mean scores (0-4) were calculated for the cyberbullying victimisation and perpetration scales. The scale scores were also dichotomised (zero and > zero) to obtain binary measures of any exposure to and involvement in cyberbullying behaviours. Hence, these variables measure any involvement in cyberbullying behaviours (even a single instance such as being sent a nasty text message or hurtful comment on a social networking site) and should not be interpreted as defining students who are/are not cyber bullied nor did/did not cyber bully others.

Both binary and continuous measures of bullying behaviours are illustrated in this paper as the sample sizes required for each can differ markedly.

Demographic variables considered in this paper include student gender as well as those which represent ways of grouping schools commonly included in research studies and hence, where separate ICC values may be of interest. These are school sector (Government vs. non-Government), location (metropolitan vs. non-metropolitan), school level (primary vs. secondary) and school size. Schools were dichotomised into two equally sized groups according to the numbers of schools in the sample; smaller primary schools had up to 410 primary students and smaller secondary schools up to 666 secondary students. These groupings were chosen to ensure sufficient numbers of schools for the calculation of ICC values in each.

Calculation of ICC's and SE's

The ICC values in this paper were calculated using the 'analysis of variance' approach. This estimator of the intraclass correlation in a two-level design is the ratio of the variation due to differences between schools (σ_g^2) to the total variation in the outcome measure, where the total variation is the sum of the variation between individuals in the same school (σ_e^2)

and the school level variation: ICC = σ_g^2 / (σ_g^2 + σ_e^2) (Snijders & Bosker, 1999). Binary outcomes are commonly analysed by means of logistic regression, in which case the level 1 error term is assumed to follow a logistic distribution with a constant variance of $\sigma_e^2 = \pi^2/3 = 3.29$ (Rabe-Hesketh & Skrondal, 2008; Snijders & Bosker, 1999; Twisk, 2006).

Maximum likelihood estimates of the ICC's and their standard errors were calculated in Stata10 using the xtreg (with the mle option) and xtlogit procedures, fitting linear and logistic regression models respectively and including random intercepts to account for the school-level clustering (StataCorp, 2007). The ICC standard error for continuous outcomes is calculated using the delta method in xtreg (StataCorp, 2011). For binary outcomes, in xtlogit estimates of the standard errors are derived from the second derivative of the likelihood function (Rodriguez & Elo, 2003).

Only schools with five or more students were included (Rabe-Hesketh & Skrondal, 2008) and in most instances the ICC values are estimated based on at least 40 schools, to ensure adequate estimation of the variance components, ICC's and standard errors (Donner & Klar, 2004; Murray, Varnell, & Blitstein, 2004).

A simple means of ascertaining the required sample size for a cluster sample is to calculate the required sample size based on simple random sampling and then inflate this figure by the design effect, obtaining an adjusted sample size with the power required for the study. The design effect (also known as the variance inflation factor or VIF) is the increase in between-school variance due to the homogeneity of students within the same school. For a simple cluster sample it is: deff = 1+(m-1)*ICC, where m is the average number of students sampled per school (Kish, 1965). This approach to calculating a sample size will be illustrated in this paper.

5.3 Results

ICC values

The ICC values and standard errors for the binary and continuous bullying measures for the entire sample and broken down by demographic variables are presented in

Table 5-1. For the total sample, the ICC values range from .015 to .031 for the victimisation and between .037 and .071 for the perpetration measures. The standard errors varied between .003 and .033. The two largest ICC values also had the largest standard errors, indicating the most uncertainty with regard to these estimates. Similarities in ICC values were found for the binary and continuous scale measures for any type of victimisation (.025 and .023 respectively), the binary any type of perpetration and cyber perpetration measures (.071 and .067 respectively) and the continuous scale measuring any type of perpetration and cyber perpetration measures (.039 and .037 respectively) (

Table 5-1).

The estimates are based on between 43 and 106 schools, with cluster sizes ranging from 35 to 85 students per school (Table 5-2). The actual cluster sizes varied appreciably between schools, e.g. for the total sample, in the smallest school 11 and in the largest between 181 and 186 students responded (depending on the outcome measure). Note that, whilst the binary measures for any type of bullying perpetration and victimisation (columns 1-4 of

Table 5-1) represent behaviours that occur every few weeks or more often, those for cyberbullying perpetration and victimisation (columns 5-8 of

Table 5-1) represent any exposure to or involvement in cyberbullying behaviours.

The ICC values give some interesting insights into the clustering of bullying behaviours within schools. Higher ICC values indicate greater disparities between schools with regard to bullying behaviour and thus higher concentrations of bullying behaviours within particular schools. Lower ICC values indicate commonalities between schools. As the total ICC values are close to zero (range from .015 to .071), variation between schools is low and the occurrence of bullying behaviours seem, therefore, not particular to only certain schools. In almost all instances, the ICC values for the perpetration measures were higher than those for the victimisation measures. Differences between schools, therefore accounted for a greater percentage of the variation in perpetration than their contribution to the victimisation measures. The exception was for secondary schools, where the ICC values for the matching perpetration and victimisation measures were similar.

Group differences in values

The ICC values were higher for girls on each of the bullying measures, thus girls' perpetration of bullying and bullying victimisation behaviours differed to a greater extent between schools than was the case for boys.

When comparing ICC values for primary and secondary schools, the values were higher in primary schools for perpetration of bullying, but lower for bullying victimisation. This means that differences between primary schools were more pronounced and students within the same primary school were more similar in their perpetration behaviours than those within secondary schools, where students across schools were more similar. However, less diversity existed between primary schools on bullying victimisation than between secondary schools.

For most of the measures, the value of the ICC for the total sample lay within the range of the values for school level, school size and school sector, indicating variation between schools within groupings was greater or similar to the variation across the entire sample of schools. The exceptions were the ICC values for the two cyber perpetration measures (binary and continuous), where the ICC's for the total sample were higher than those for each of the primary and secondary school samples. For example, the value for the binary cyber perpetration measure of .067 is higher than the primary school value of .040 and secondary school value of .032, indicating greater diversity between the entire range of schools sampled, than between schools within the primary and secondary groupings. This

may be due to the relatively low proportion of students, particularly in primary schools, who report perpetrating cyberbullying behaviours leading to the seemingly more pronounced clustering of these behaviours within schools.

Apart from the two binary perpetration outcomes, the ICC values for the different sized schools did not differ by more than .01. In smaller primary schools, school level variation accounted for an estimated 9.2% of the total variation in perpetration of any type of bullying behaviours. Whilst the numbers of primary and secondary schools within each school size grouping were not sufficient to adequately estimate separate ICC values, subsequent analyses showed that the large ICC for perpetration of any type of bullying in smaller schools may be due to a high level of variability on this binary measure, particularly among the 28 smaller primary schools. School differences made up 8.3% of the total variation in involvement in cyberbullying in larger schools. This result could not be attributed to primary or secondary schools in particular (both levels had ICC values well below .083) and seems to be a consequence of differences between larger primary and secondary schools in perpetration of cyberbullying behaviours.

The ICC values were similar for the government and non-government sectors. Apart from the binary measure for perpetration of any type of bullying, the non-metropolitan schools had higher or similar values to the metropolitan schools on all the measures, signifying more pronounced clustering of bullying and cyberbullying behaviours within certain non-metropolitan schools.

Table 5-1. ICC values – Total sample and by demographic group.

ICC (*SE*)	Any type of bullying victimisation / perpetration				Cyberbullying victimisation / perpetration			
	Vict Yes/No	Perp Yes/No	Vict Scale	Perp Scale	Exp Yes/No	Invol Yes/No	Exp Scale	Involv Scale
Total	**.025**	**.071**	**.023**	**.039**	**.031**	**.067**	**.015**	**.037**
	(.007)	(.017)	(.005)	(.007)	(.007)	(.013)	(.004)	(.007)
Gender								
Female	**.035**	**.128**	**.048**	**.056**	**.038**	**.081**	**.022**	**.053**
	(.011)	(.033)	(.011)	(.012)	(.011)	(.018)	(.007)	(.012)
Male	**.022**	**.044**	**.008**	**.035**	**.011**	**.064**	**.012**	**.027**
	(.010)	(.019)	(.006)	(.009)	(.010)	(.017)	(.007)	(.008)
School level								
Primary	**.019**	**.083**	**.018**	**.041**	**.029**	**.040**	**.006**	**.030**
	(.007)	(.024)	(.006)	(.010)	(.010)	(.013)	(.003)	(.008)
Second-	**.032**	**.031**	**.028**	**.025**	**.032**	**.032**	**.019**	**.017**
ary	(.013)	(.017)	(.010)	(.009)	(.012)	(.012)	(.007)	(.007)
School size								
Smaller	**.021**	**.092**	**.023**	**.033**	**.031**	**.048**	**.012**	**.032**
	(.008)	(.027)	(.007)	(.009)	(.011)	(.014)	(.005)	(.009)
Larger	**.025**	**.033**	**.020**	**.040**	**.028**	**.083**	**.019**	**.042**
	(.010)	(.018)	(.008)	(.011)	(.010)	(.021)	(.007)	(.013)
School sector								
Govern-	**.024**	**.063**	**.020**	**.040**	**.029**	**.062**	**.011**	**.034**
ment	(.010)	(.024)	(.008)	(.011)	(.011)	(.019)	(.005)	(.010)
Non-gov	**.022**	**.077**	**.023**	**.036**	**.031**	**.066**	**.019**	**.036**
	(.008)	(.023)	(.007)	(.009)	(.010)	(.016)	(.007)	(.010)
Area								
Metro-	**.018**	**.076**	**.018**	**.032**	**.023**	**.069**	**.013**	**.037**
politan	(.007)	(.023)	(.006)	(.008)	(.008)	(.017)	(.005)	(.010)
Non-	**.029**	**.057**	**.025**	**.049**	**.041**	**.065**	**.019**	**.037**
metro	(.012)	(.023)	(.009)	(.014)	(.014)	(.020)	(.008)	(.011)

Note. Vict: Victimisation; Perp: Perpetration; Exp: Exposure to; Involv: Involvement in.
Binary measures for any type of bullying perpetration and victimisation represent behaviours that occur every few weeks or more often, binary measures for exposure to (Exp.) and involvement (Involv.) in cyberbullying include single instances of such behaviours (once or twice a term or more often).

Table 5-2. Numbers of schools and students

Grouping	Number of schools	Total number of students	Mean cluster size	Minimum cluster size	Maximum cluster size
Total	106	7238-7312	69	11	181-186
Gender					
Female	101	3795-3836	38	5	108-109
Male	99	3416-3454	35	5	98-101
School level					
Primary	55	4569-4606	83	20-21	181-186
Secondary	51	2669-2711	53	11	125-129
School size					
Smaller	54	3843-3882	72	11	181-186
Larger	52	3384-3430	66	12	128-131
School sector					
Government	52	2708-2749	53	11	148-152
Non-government	54	4530-4565	84	12	181-186
Area					
Metropolitan	63	4640-4701	74	11	181-186
Non-metropolitan	43	2592-2615	61	15	148-152

Note. Mean cluster size rounded to the nearest whole unit. Ranges given as numbers of students varied by outcome measure

The impact of other factors on ICC values

The power of an analysis can be improved by lowering the value of the operative ICC value through judicious statistical modelling, such as including variables which explain school-level variation in analyses (Murray & Blitstein, 2003). The reductions in ICC values resulting from the addition of covariates to statistical models are illustrated in Table 5-3 for logistic/linear regression models including different demographic variables. In particular, the impact of the addition of gender and Australian State/Territory is shown. The inclusion of gender does not have a large effect on the ICC values (comparing Models 1 & 2, and Models 3 & 4) whereas the inclusion of Australian State/Territory does (comparing Models 1 & 3, and Models 2 & 4). This finding is because gender is measured at the student level and it is not able to explain or reduce much of the variation between schools, unlike variables measured at the school level. In fact, increases in ICC values may result from

adjustment for student level variables when there is an imbalance in the variable among schools, such that they appear more similar than they are (Murray & Blitstein, 2003).

Importantly, some of the variation between States/Territories is likely to be due to differences in parental consent processes, with government sectors in certain States/Territories requiring active parental consent (rather than allowing active/passive consent procedures) prior to student participation in surveys. These requirements resulted in markedly lower participation rates among active consent-only schools and hence likely greater homogeneity of responding students. A further explanation may be differences in the location of Year 7 students, mostly in primary schools in certain States/Territories and in secondary schools within others.

Table 5-3. ICC values adjusted for demographic variables

	Any type of bullying victimisation / perpetration				Cyberbullying victimisation / perpetration			
	Vict Yes/No	Perp Yes/No	Vict Scale	Perp Scale	Exp Yes/No	Involv Yes/No	Exp Scale	Involv Scale
Unadj.	**.025**	**.071**	**.023**	**.039**	**.031**	**.067**	**.015**	**.037**
Model 1	.019	.057	.018	.030	.028	.037	.011	.022
Model 2	.019	.055	.020	.029	.026	.036	.010	.022
Model 3	.010	.040	.013	.020	.013	.026	.003	.018
Model 4	.010	.037	.015	.020	.009	.026	.002	.018

Note. Unadj.: Unadjusted; Vict: Victimisation; Perp: Perpetration; Exp: Exposure to; Involv: Involvement in.
Model 1: Adjusted for area, sector, year level, school size
Model 2: Adjusted for gender, area, sector, year level, school size
Model 3: Adjusted for area, sector, year level, school size, Australian State/Territory
Model 4: Adjusted for gender, area, sector, year level, school size, Australian State/Territory

Cluster sizes and design effects

Whilst the ICC values seem negligible and they indicate small clustering effects, their impact in terms of the design effect and therefore on the power of a study is not able to be ignored, especially when large numbers of students are sampled per school (Table 5-4). A selection of ICC values for primary and secondary schools from

Table 5-1 have been utilised for illustrative purposes, together with increments of 25 students (roughly one class) per school.

As expected, the design effects increase as the ICC values and the cluster sizes increase. The greater the homogeneity of students within schools and the more students sampled per school, the less independent information to be gained from each individual and the sample. Even for a small ICC of .006 and an average of 200 student respondents per school, the required sample size to achieve the same power for a cluster sample is more than double (2.2 times) that of a simple random sample.

The importance of an ICC value with regard to power and sample size determination is related to the number of students who will be sampled per school. A small ICC=.006 is not an issue if 25 students per school are sampled, as only a small increase in sample size is needed to attain the required power for the study. However, it is evident from the first column of the table the degree to which a larger sample is required for that same ICC value, as the number of students per school increases. Sample sizes need to be inflated by a factor of at least 1.5 for the higher ICC values, regardless of whether the numbers of students per school are 25 or 250. Thus, if design effects are ignored when designing studies aimed at measuring and testing bullying outcomes, underpowered samples will result.

Table 5-4. Design effect sizes for different cluster sizes

	ICC				
m	0.006	0.019	0.032	0.041	0.083
25	1.1	1.5	1.8	2	3
50	1.3	1.9	2.6	3	5.1
100	1.6	2.9	4.2	5.1	9.2
150	1.9	3.8	5.8	7.1	13.4
200	2.2	4.8	7.4	9.2	17.5
250	2.5	5.7	9	11.2	21.7

Note. m: cluster size/number of students per school

Calculation of required sample size

Studies of bullying behaviours in schools are conducted for multiple reasons and the purpose of the study is a determining factor in deciding on the form of the bullying measure to be used. Commonly, the prevalence of such behaviours is estimated or

compared e.g. in studies of anti-bullying interventions. Alternatively, a researcher may wish to explore the relationship between bullying behaviours and other individual-level factors such as students' mental health or academic outcomes. If prevalence is the focus, single questions that can be categorised to identify students who have been bullied or have bullied others would be appropriate. When investigating associations between individual characteristics and bullying, a multi-item scale from which a continuous composite score can be calculated as a measure of involvement in bullying behaviours, would give greater sensitivity and variability than a binary outcome.

Apart from the measurement scale of the bullying outcome and the design effect (as determined by the ICC and cluster size), the required sample size for a cluster sample of schools is dependent on the size of the effect to be determined and, for categorical outcomes, the prevalence of the outcome. Table 5-5 summarises the calculation of the required numbers of students and schools for cluster samples for the four measures of any bullying, the corresponding ICC values for each measure and different prevalence rates and effect sizes. The calculations are conducted separately for primary and secondary schools, assuming an average of 100 responding students per school (after accounting for consent and non-response rates), power of 80% (conventionally the minimum acceptable value) and based on simple two-sided tests of proportions or means in two independent samples. Prevalence rates of 10% to 30% were chosen in line with the rates found in the ACBPS study (9% and 27% for any perpetration and victimisation respectively) and small (.25) and moderate (.5) effect sizes for continuous outcomes (Cohen, 1988). In most cases the required numbers of schools are rounded up. To achieve power greater than 80%, larger sample sizes than presented here would be required.

As an illustration of the use of the design effect estimate to determine the required sample size for a cluster sample, consider a study with the major outcome of comparing the prevalence of bullying victimisation in two groups (e.g. in an intervention trial) in primary schools (ICC= .019 from

Table 5-1). Assuming that the prevalence of bullying victimisation is 20% in the group with the lower rate and wishing to have 80% power to detect a difference of 5% between the groups (i.e. 20% in one and 25% in the other), the required number of students per group for a simple random sample is 1140 students. With 100 students per school, the anticipated design effect is 2.9, resulting in a total required sample of 3306 rather than 1140 students per group. Given the assumption of 100 respondents per school, this equates to about 33 schools per group and 66 schools in total. Note that, whilst the number of schools does not figure directly in the calculation of the design effect, it is implicitly determined by the numbers of students to be sampled per school and it is therefore advantageous to sample fewer students per school and more schools, rather than more students in fewer schools.

Within Table 5-5 the values of the input parameters to the calculations are adjusted as appropriate for the various measures, but also to illustrate their impact on the sample size calculation. Firstly, the lower prevalence of the two groups was varied between 10%, 20% and 30% to show how the required sample size increases as this rate increases. Thus, for a binary outcome, 730, 1140 and 1420 students are required in a simple random sample to detect a difference of 5% for prevalence rates of 10%, 20% and 30% respectively. Secondly, the impact of effect size is illustrated in terms of both differences in percentages and means. For 20% prevalence, to detect a smaller difference of 5% requires a larger simple random sample size of 1140 compared to that of 320 to detect a difference of 10% between two groups. Similarly, a sample of 255 is required to determine a small effect size of .25 for a continuous outcome measure as statistically significant compared with 64 if only a moderate effect size of .5 was considered important. Thirdly, the much lower sample size requirements for testing outcomes measured on continuous compared with categorical scales are illustrated.

It is important to note the differences in required cluster sample sizes for the victimisation and perpetration outcomes measured on the same scale, due to the differences in ICC values for the perpetration outcomes. For example, a sample of 2117 was required for the binary bullying victimisation measure compared with 6716 for the perpetration measure. As sample size calculations need to account for all the key outcome measures of a study, the largest ICC value is most pertinent.

As a practical example of the effects of ignoring school-level clustering on the conclusions drawn from a study, consider the case of an intervention trial of an anti-bullying program in

secondary schools. Based on the ACBPS data, one could assume that the prevalence of bullying victimisation is about 30% (conservatively rounded up from 27%). Assume that a researcher, ignoring clustering effects, determines the sample size required for the trial as 380 students in order to have 80% power to detect a 10% decrease in bullying behaviours as statistically significant i.e. if the program results in a reduction from 30% to 20% of students victimised, it should be seen as effective. However, with a sample of 380 students, the program would actually need to achieve a reduction of at least 22% i.e. from 30% to 8% of students' victimised before a statistical test would show the program as having a significant impact.

Table 5-5. Required sample sizes for cluster samples

					Students		Schools	
	School level	ICC	Deff m=100	Parameters / Effect size	n_{SRS}	$n_{cluster}$	Per group	Total
Any vict	Primary	0.019	2.9	Prev.=10%, diff. of 5%	730	2117	22	44
Yes/no	Secondary	0.032	4.2	Prev.=10%, diff. of 5%	730	3066	31	62
	Primary	0.019	2.9	Prev.=20%, diff. of 5%	1140	3306	33	66
	Secondary	0.032	4.2	Prev.=20%, diff. of 5%	1140	4788	48	96
	Primary	0.019	2.9	Prev.=30%, diff. of 5%	1420	4118	42	84
	Secondary	0.032	4.2	Prev.=30%, diff. of 5%	1420	5964	60	120
	Primary	0.019	2.9	Prev.=20%, diff. of 10%	320	928	10	20
	Secondary	0.032	4.2	Prev.=20%, diff. of 10%	320	1344	14	28
	Primary	0.019	2.9	Prev.=30%, diff. of 10%	380	1102	12	24
	Secondary	0.032	4.2	Prev.=30%, diff. of 10%	380	1596	16	32
Any perp	Primary	0.083	9.2	Prev.=10%, diff. of 5%	730	6716	68	136
Yes/no	Secondary	0.031	4.1	Prev.=10%, diff. of 5%	730	2993	30	60
Any vict	Primary	0.018	2.8	Effect size .25	255	714	8	16
Scale	Secondary	0.028	3.8	Effect size .25	255	969	10	20
	Primary	0.018	2.8	Effect size .5	64	180	2	4
	Secondary	0.028	3.8	Effect size .5	64	244	3	6
Any perp	Primary	0.041	5.1	Effect size .25	255	1301	13	26
Scale	Secondary	0.025	3.5	Effect size .25	255	893	9	18
	Primary	0.041	5.1	Effect size .5	64	327	4	8
	Secondary	0.025	3.5	Effect size .5	64	224	3	6

Note. m: cluster size; Deff: design effect; n_{SRS}: sample size required for simple random sample; $n_{cluster}$: sample size required for cluster sample; Prev: Prevalence; diff: difference; Vict: Victimisation; Perp: Perpetration.

5.4 Discussion

Good estimates of ICC values offer insights into bullying behaviours in schools and are vital for planning group-randomised trials or studies utilising cluster sampling (Donner & Klar, 2004; Murray, et al., 2006; Scheier, et al., 2002; Siddiqui, et al., 1996). To our knowledge, limited information has been published regarding ICC values for bullying behaviours and nothing to date for cyberbullying. This paper presents ICC values for bullying and cyberbullying outcomes based on a large representative Australian sample. Each calculation is based on more than forty schools to ensure stability of the estimates (Donner & Klar, 2004).

Few studies reporting ICC values for bullying-related outcomes for mainstream students were identified in the literature. Values reported by Kärnä et al. (2011) are not directly comparable with those found in this study as they also accounted for classroom-level clustering, which is of greater importance in Finland than in Australia due to the stability of classroom structures in the Finnish system. The outcome measures utilised by Bradshaw et al (Bradshaw, et al., 2009) were similar to the single item binary any bullying measures used in this study, with similar dichotomisations following the work of Solberg & Olweus (Solberg & Olweus, 2003). The value obtained here for victimisation for primary school students of .019 was the same as that found by Bradshaw (Bradshaw, et al., 2009) for elementary students (Grades 4 & 5, n=76 schools). However, their reported values for middle school students (Grades 6-8), based on only 19 schools, of .006 for victimisation and .009 for perpetration were substantially lower than those in this study for similar aged students (.032 and .031 respectively). In contrast, the values for secondary students of .03 found here are within the range of .01-.06 reported for a range of outcomes including bullying victimisation in the Gatehouse Project conducted amongst secondary students in Australia (Bond, et al., 2004). No published ICC values for cyberbullying-related outcomes were found.

The low ICC values obtained in this study (below .1) show that little variation exists between schools i.e. bullying behaviours are not more concentrated in certain schools, but are prevalent to a similar extent across all schools. This is possibly a surprising finding, given bullying often occurs within the school context and may be perpetrated by a

relatively small number of students. Indeed, the values are in line with those of other health outcomes such as nutritional intake and physical activity (Carlin & Hocking, 1999; Murray, et al., 2000; Murray, et al., 2001; Murray, et al., 1994; Murray, et al., 2006; Scheier, et al., 2002; Siddiqui, et al., 1996), which one would possibly expect to be less influenced by the school environment than bullying. In contrast, ICC values for academic outcomes do display much stronger clustering effects (Hutchison, 2004; OECD, 2010). This lack of evidence that some schools have a stronger 'bullying culture' than others, together with the fact that about a quarter of Australian students are bullied every few weeks or more often, highlights the need for bullying reduction and management programs in all schools.

The ICC values were higher for the perpetration than the corresponding victimisation outcome measures, indicating students across all schools were more homogeneous in respect to their reports of bullying victimisation than perpetration of bullying. This was particularly evident for primary rather than secondary schools. These trends occurred for both the any type of perpetration and cyber perpetration measures, implying that clustering of cyberbullying behaviours are similar to those of bullying behaviours in general. This greater variability between schools with regard to perpetration than victimisation may be reflective of lower rates of self-reported bullying perpetration, highlighting differences between schools. Additionally, these differences may be related to school level social norms or normative expectations related to the reporting of victimisation and perpetration of bullying behaviours. It is possible that students in some schools are less likely to report bullying perpetration or victimisation than is the case in other schools, perhaps due to school climate or unhelpful staff responses, adding to the variability between schools. Self-serving attribution bias may be more evident for perpetration than victimisation suggesting students report bullying targeting them more highly, than their own perpetration of bullying (Österman et al., 1994).

Differences in ICC values were noted for various demographic groups. For example, higher ICC values for girls on all the measures indicated contextual effects were stronger for girls, with a greater concentration of bullying behaviours in certain schools for girls whilst occurring more commonly across all schools for boys. Whereas few studies report ICC values for gender separately, or the other demographic groups considered in this study, Siddiqui also reported larger clustering effects for female than male students on current smoking status (Siddiqui, et al., 1996). One conclusion from this finding is that bullying between girls is more context dependent than is the case for boys. Thus, the need for

bullying prevention and management interventions is uniform across all schools for boys, but the need may be greater within certain schools than others for girls.

Similarly, a trend towards higher ICC values for non-metropolitan than metropolitan schools signifies a greater clustering of bullying behaviours in certain non-metropolitan schools. This may be attributable to the diversity of schools and environments in non-metropolitan areas in Australia, which include schools in rural areas as well as large regional centres. School size and sector did not greatly influence ICC values.

As Murray and colleagues (Murray, et al., 2001; Resnicow, et al., 2010) have noted, the reductions in ICC values that can be achieved through the addition particularly of school-level variables to regression models, are also demonstrated here. Apart from demographic variables as considered in this study, the inclusion in models of other school-level factors correlated with the outcome of interest, such as teacher-student ratios, anti-bullying policy implementation, and quality of school leadership, may also result in reductions in ICC values. Additionally, if a longitudinal study and repeated measures analyses are planned, lower ICC values are operative than those from a cross-sectional study (Murray & Blitstein, 2003). Therefore, the values presented here are likely an upper limit of those that would apply should such models be applied in planned studies.

Nevertheless, whilst ICC values are typically less than 0.1 and appear negligibly small, their influence in reducing precision and thus the power of a study are substantial if the number of students sampled per school is large. The impact of an ICC as low as 0.006 in terms of the design effect and resultant increase in sample size required for a cluster sample to achieve the same level of power as a simple random sample, is illustrated. For bullying related outcomes, design effects are not negligible when samples of about 25 or more students are sampled per school. Sample sizes need to be inflated by a factor of at least 1.5 and sometimes substantially more, or imprecise estimates and underpowered studies will result. A lack of power will lead to erroneous conclusions, for example, a failure to identify factors associated with bullying outcomes or demonstrate effective interventions.

Further, the practice of assuming clustering effects may be ignored on the basis of the non-significance of an hypothesis test that the ICC is zero, is not recommended as such tests have limited power (Donner & Klar, 2004) and are thus unlikely to detect ICC values that do differ significantly from zero.

A simplified approach to the calculation of the required sample size for a cluster sample is presented. Whilst the calculations are valid for testing individual-level effects, often the number of schools sampled is also of relevance. From a sampling perspective, it may be difficult to obtain a representative sample of a target population of students from a limited number of schools. Also, in studies testing contextual school-level variables (eg. school size or policy), it is recommended that a minimum of 40 schools be sampled as a sufficient sample size to assess school-level outcomes (Donner & Klar, 2004; Murray, et al., 2004). In intervention trials where schools are assigned to study conditions, this equates to 20 or more schools per study condition.

This study is subject to a number of limitations which restrict the applicability of the presented ICC values. Firstly, no account was taken of possible classroom clustering effects. In Australia classroom clustering would only apply to primary schools where, unlike in secondary schools where students move between classes throughout the day, in a single school year students largely stay with the same classroom of students and teacher throughout the school day. The extent to which classroom level effects will be present depends on the extent that bullying behaviours tend to be perpetrated between students in the same classroom rather than more broadly. Unfortunately information on class membership was not available for the sample analysed for this paper, but analyses of data from another project held by the CHPRC revealed classroom-based ICC values two to five times higher in magnitude than school-based values in a sample of twenty primary schools. Thus, whilst the ICC values presented in this paper are appropriate for use when designing studies in secondary schools, they are underestimates of the relevant values for primary schools where class-level clustering is of importance and may need to be accounted for. Secondly, the ICC values represent a combination of school and cohort effects. As this is a cross-sectional study of a specific cohort of students, these effects could not be separated (Smolkowski, Biglan, Dent, & Seeley, 2006). Thirdly, due to the relatively small numbers of students per year level and likely cohort effects, it was not possible to reliably estimate ICC values per year level. Consequently the applicability of the values presented for primary and secondary school students is limited by the extent to which clustering effects are similar in year levels within primary and secondary schools. Fourthly, the data used in this study were collected in the last term of the school year. The time of the year students were surveyed was found to influence the ICC values related to physical activity outcomes (Murray, et al., 2006). Clustering effects may also differ by school term for bullying

outcomes, especially in certain year levels such as the first year of secondary school (Pellegrini & Bartini, 2000; Rigby, 1998; Smith, Madsen, & Moody, 1999).

Some further considerations in the interpretation of the results presented here are pertinent. The data analysed were self-reported, clustering effects are likely to be higher for peer-report of bullying involvement. Kärnä et al (2011) reported an ICC=.13 for peer nomination of victimisation. Greater homogeneity of peer nominations may be partly due to a phenomenon known as reputation bias, where students' perceptions of some of their peers as "victims" or "bullies" persist despite behavioural changes that may occur (Hymel, Wagner, & Butler, 1990). If teacher-report is used, teacher-level clustering is an added strong source of variation to be accounted for.

Some authors have used linear procedures to calculate ICC values for binary outcomes. We compared the values obtained using xtlogit and xtreg and found the values using the linear procedure were lower. Taking a conservative approach we have presented the ICC values obtained from the logistic regression procedure, as this is in accordance with how the data are likely to be analysed and therefore arguably the more relevant ICC value.

Clearly smaller sample sizes are required with continuous than categorical outcomes, and studies can be powered to detect small effect sizes with relatively few schools. As mentioned, however, the numbers of schools sampled is also a critical consideration. In addition, the choice of outcome measure to be used should be based on theoretical considerations and the study's research questions.

Murray et al (2004) have described the need for researchers to use ICC estimates 'in their power analyses that closely reflect the endpoints, target population, and primary analyses planned for the trial'. The values reported in this paper may have assisted in this regard. A simplified method of determining the required sample size for a school-based cluster sample targeted at the measurement and testing of bullying outcomes is also described. The approach is applicable to any outcome measure and setting for which appropriate ICC values are available.

5.5 Conclusions

Results from this study suggest that bullying behaviours are relatively uniform across schools in Australia, with no marked differences in the bullying culture between schools.

Indeed, school context is more strongly associated with academic outcomes than bullying. This highlights the importance of providing anti-bullying interventions in all schools, for both boys and girls and regardless of school level, size, geographic location or sector.

A number of factors impact on the required sample size for a cluster study design to achieve a certain precision and power. Greater homogeneity of students within schools, as measured by higher ICC values, leads to larger design effects and thus, the amount by which the sample needs to be inflated to achieve the same precision and power as a simple random sample. Similarly, the larger the number of students that will be sampled per school, the larger the sample size for a cluster sample will need to be. Bullying outcomes measured on a continuous scale often require substantially smaller sample sizes than binary outcomes. Larger sample sizes are required if greater precision in estimation is desired or if it is important for the study to detect a smaller effect.

Although the ICC values for bullying outcomes are small, they are not able to be ignored and need to be accounted for when designing studies, particularly when large numbers of students are sampled per school. Sample sizes need to be inflated by a factor of at least 1.5 and sometimes substantially more, or estimates of bullying outcomes will lack precision and underpowered studies will result.

When designing studies to test school contextual variables or for intervention trials, the number of schools sampled is of vital importance to the validity of the findings. Samples of 40 schools are recommended, 20 per study condition in group-randomised trials, to adequately test for school-level variables or intervention effects. Studies based on small numbers of schools are likely to be underpowered and subject to a number of biases. Whilst in general it is advantageous to sample more schools with fewer students in each, rather than fewer schools with more students, in intervention trials the sample size requirements also need to be assessed in light of the resources available to the research team to support intervention implementation in study schools.

Chapter 6 : An empirical comparison of statistical models to test for school anti-bullying program effects

Relevance to the thesis

This chapter presents discussion and analyses central to Research Question 5 of this thesis. This expository chapter describes the use and advantages of traditional and innovative statistical methods that could be applied to test for program effects, using empirical data. While the statistical methods described herein are not new, their application to bullying outcomes is.

6.1 Introduction

Valid conclusions from a trial assessing the impact of an anti-bullying program are predicated on the use of appropriate statistical methods in estimating the program effect, and the extent to which the data meet the assumptions of the models applied. When bullying outcomes are used as measures of program impact, the choice as to the statistical model to apply and validity of the relevant assumptions need to be carefully considered, due to the highly skewed nature of these variables or invalid inferences may be drawn regarding the effectiveness or otherwise of programs. Thus, conventional methods of analysis may not be applicable to these data and alternate statistical models need to be considered. Additionally, different approaches and statistical models can lead to differing conclusions as, for example, they may address different substantive questions or have different levels of statistical power. The aim in this chapter is to describe the use of conventional and less used, innovative models when evaluating program impact on the basis of bullying outcomes and applied to empirical data – insights that can be gained from the different techniques and the assumptions associated with each will be discussed.

As a multitude of statistical techniques are available for data analysis, it is important to set the context within which the statistical models are to be applied and thus, the relevant techniques to consider. Study designs which include control conditions (or generate control data) where longitudinal data are collected on at least two occasions i.e., prior to and after the program implementation, provide the strongest evidence on program impact. For causal inference, i.e. to interpret group differences as program effects, it is necessary that the changes in the control group accurately represent what would have occurred in the program group had the program not been implemented (London & Wright, 2012; Wright, 2006). Random assignment of sufficient numbers of respondents (e.g. students) or the units to which they belong (e.g. schools), to the study groups, is a means of ensuring this comparability of the groups. This chapter, therefore, focuses on the analysis of repeated measures data from respondents in two comparable study groups, one group which does and one which does not receive the program to be evaluated.

Another important assumption in this discussion, is that the primary dependent variables by which the program is to be evaluated, measure the frequency of involvement in bullying victimisation and/or perpetration. Frequency of involvement in bullying behaviours is often measured using self-report on a single global item or a scale with multiple items describing different forms of bullying (Swearer, Siebecker, Johnsen-Frerichs, & Wang, 2010). Typically,

four to five response options are used for both types of questions, ranging from none to highly frequent involvement e.g. almost every day. A variety of dependent variables can be formed from the responses to these questions, but the impact of anti-bullying programs is typically assessed based on variables assumed to be on a binary or a continuous measurement scale (Merrell, Gueldner, Ross, & Isava, 2008). These formats will be the focus of this chapter, in particular, binary variables created as dichotomisations of a single multi-category global item and mean scores calculated from multi-item scales. The results from the different models applied to each bullying outcome variable will be compared.

Three commonly applied statistical approaches to the analysis of longitudinal data will be considered, namely cross-sectional analyses, ANCOVA-type analyses and growth modelling. In the first, cross-sectional analyses at each time point will be conducted. These are presented as illustrations of the simplest form of analyses that are applied by researchers to test for group differences at each time point, but also as they are equivalent to the manner in which effect sizes are often calculated in meta-analyses of program evaluations, the basis on which conclusions regarding the effectiveness of school anti-bullying programs are made (Farrington & Ttofi, 2009; Merrell et al., 2008; Smith, Schneider, Smith, & Ananiadou, 2004). The second approach to modelling considered is the use of analysis of covariance or ANCOVA models, where the pre-program values of the dependent bullying variable are included as a predictor in the model and hence controlled for statistically. In group-randomised trials, schools rather than students are randomised to conditions and it is possible despite the randomisation, especially when the study includes a smaller number of schools, that dissimilarities will be present in the pre-program measures of the study groups. This is also the case for age-cohort designs, where different cohorts of students act as controls for each other. Thus, statistically controlling for pre-program differences on the outcome variables may need to be considered in the analyses. These analyses have the added advantage over models which do not include the baseline value of the dependent variable, of increased power to detect group differences and increased precision of estimates of effects. Using the third approach, a variety of growth models for longitudinal data will be applied. Growth models compare trajectories or "growth" in bullying behaviours between study groups, hence trends over time are the focus rather than differences between the groups at particular points in time, as is the case for cross-sectional and ANCOVA analyses.

6.2 Empirical study

Data from the Supportive Schools Project (SSP) were utilised for illustrative purposes in this chapter, as the SSP was a program evaluation trial in which longitudinal data on bullying outcomes were collected. This group-randomised trial was conducted in Catholic secondary schools in metropolitan Perth, Western Australia (Waters, Epstein, Cross, & Shaw, 2008). The whole-of-school program aimed to reduce bullying and enhance students' social and emotional wellbeing. Data collected at three time points during 2006-2007 from the cohort of participating students in 20 schools were analysed here.

Sampled schools (20 of the 29 eligible) were randomised to study groups – the 10 intervention schools implemented the SSP program while the 10 control schools followed their normal strategies and programs. All Grade 8 students (aged 13 years on average) in each school were eligible to participate (parental consent rate = 92.9%). Data collected when the students were in Term 1 Year 8 (n = 3,068, 96.2% response rate), Term 3 Year 8 (n = 2,966, 96.7% follow-up rate) and Term 3 Year 9 (n = 2,739, 89.3% follow-up rate), are analysed here. These three time points will be designated as Time 1 to Time 3 respectively. Trained researchers administered the hard copy surveys to the students with consent in a usual class period following a strict procedural and verbal protocol. School co-ordinators were asked to ensure students absent on the day completed surveys on their return to school. Half of the students were boys (n = 1,541, 50.2%).

Both single item global questions and multi-item scales measuring frequency of bullying victimisation and perpetration were included in the SSP surveys. The questions analysed in this chapter are given in Appendix 6-A – the global questions correspond to Questions 2 and 4 and the scales to Questions 1 and 3. The global questions were analysed with the five categories dichotomised in line with recommendations for identifying victimised and perpetrating students (Solberg & Olweus, 2003), to create binary variables with values 0 (once or twice or less often in the term) and 1 (every few weeks or more often). This cut-off is also used as it is in accordance with the definition of bullying which requires the behaviour to be repeated over a period of time and not one or two isolated instances. Aggregated scores for the scales were calculated as means of the items (reverse coded and rescaled from 1-5 to 0-4). These mean scores represent a sum of the frequency and the number of different ways in which a student was bullied or bullied others, with higher scores representing greater exposure to or perpetration of bullying behaviours and zero indicating no involvement.

6.3 Distributions of bullying outcomes

Statistical models are based on assumptions regarding the distributions of the analysed variables, in particular that of the dependent variable(s). A critical factor to consider when analysing bullying outcomes is the skew in the distributions of the variables, whether these are single questions or aggregated scores from multi-item scales. Many students are not involved in bullying, either as a targeted student or a perpetrator and this leads to a high percentage of students reporting no involvement. Furthermore, relatively few involved students are bullied/bully others at the most frequent levels, leading to a skewed distribution for the values above the minimum. The percentages of students within each of the categories of the single item global questions for each of the three time points in the SSP are presented in

Table 6-1. A majority of the students reported no involvement in bullying behaviours, 63% or more had not been bullied in the term and 66% or more had not bullied others. Further, the percentages drop in value from between 18% – 26% in the category “1-2 a term” to 5% or less for “Most days”.

The distributions of the mean scale scores are also highly skewed, as illustrated for all the mean scores in

Table 6-1, and as an example for victimisation at Time 1 in **Error! Reference source not found.**. Not only are the distributions skewed, but there is a preponderance of values at the minimum indicating “zero” involvement. Between 30% and 38% of the students reported they were not exposed to any of the seven forms of bullying, i.e. had a zero mean score on the victimisation scale. The high percentage of zeroes is greater for perpetration – 46% or more of the sample have a zero score. Variables with distributions such as these may be considered as semicontinuous rather than measured on a continuous scale. (Although technically the most prevalent value could be the minimum or the maximum, depending on the way in which the response options are presented, one can assume without loss of generality that it is the minimum and the value is zero. Also, although mean scores were calculated here, other linear composites of the items such as weighted factor scores will be similarly distributed as the means.)

Table 6-1. Descriptive statistics for the bullying outcomes in the SSP

Global Questions

	None	1-2 in the term	Every few weeks	About once a week	Most days	Total
	% (*n*)	% (*n*)	% (*n*)	% (*n*)	% (*n*)	*n*
Victimisation						
Time 1	69.9 (*2,114*)	23.0 (*699*)	3.1 (*94*)	2.3 (*71*)	1.9 (*59*)	*3,037*
Time 2	62.5 (*1,869*)	26.1 (*782*)	4.5 (*134*)	3.2 (*95*)	3.7 (*111*)	*2,991*
Time 3	62.9 (*1,737*)	24.3 (*670*)	4.5 (*125*)	3.7 (*101*)	4.6 (*128*)	*2,761*
Perpetration						
Time 1	79.5 (*2,395*)	17.8 (*536*)	1.5 (*44*)	0.9 (*28*)	0.4 (*11*)	*3,014*
Time 2	67.7 (1,993)	26.8 (*787*)	2.4 (*70*)	1.6 (*47*)	1.5 (*45*)	*2,942*
Time 3	65.5 (1,779)	26.2 (*712*)	3.4 (*91*)	2.1 (*58*)	2.8 (*75*)	*2,715*

Mean scale scores

	M	*SD*	Range	% at zero
Victimisation				
Time 1	0.34	0.52	0 – 4	37.9
Time 2	0.45	0.63	0 – 4	31.0
Time 3	0.54	0.80	0 – 4	30.3
Perpetration				
Time 1	0.15	0.29	0 – 4	58.7
Time 2	0.23	0.46	0 – 4	50.7
Time 3	0.34	0.67	0 – 4	46.1

Note. Time 1 = Term 1 Year 8, Time 2 = Term 3 Year 8, Time 3 = Term 3 Year 9

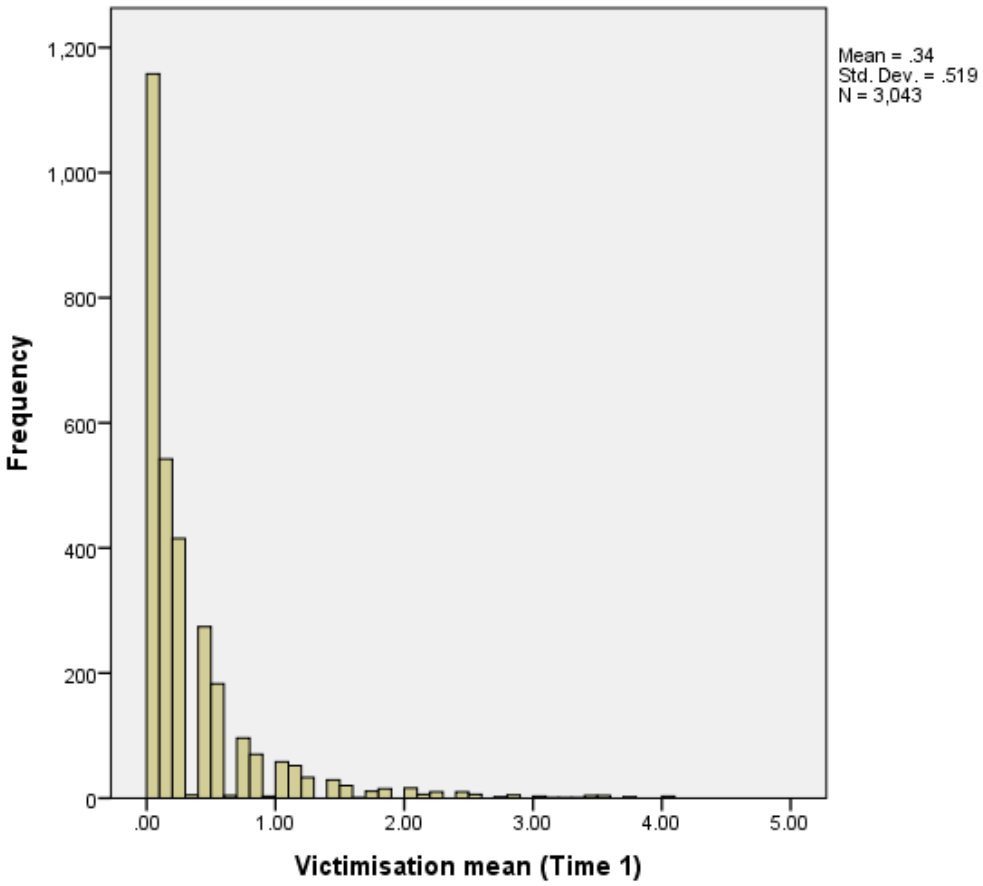

Figure 6-1. Victimisation mean scores for Time 1

6.4 Assumptions of statistical models

Statistical models are based on various assumptions, the violation of these threaten the validity of findings from the statistical analyses. Particularly pertinent, given the highly skewed nature of the mean scale scores, are the assumptions of normality and homogeneity of error variance which underlie traditional linear regression methods for continuous variables. When these assumptions are not met, estimates of regression coefficients and their standard errors may be biased, and invalid inferences may be drawn from the associated significance tests (Finney & DiStefano, 2006; McClendon, 1994; Vittinghoff, Glidden, Shiboski, & McCulloch, 2011; West, Finch, & Curran, 1995). In the context of evaluation studies, this means that ineffective programs may be assumed to have had an impact and vice versa.

As an example, Osgood and colleagues' analyses of self-reported frequency of involvement in delinquent behaviours revealed underestimates of associations and lower power when applying ordinary least squares methods, as compared with tobit regression models, to data with distributions as skewed as bullying outcomes (Osgood, Finken, & McMorris, 2002). To prevent erroneous conclusions of null effects, they recommended the use of alternate models than those based on the usual assumptions, when effect sizes are small. This highlights the need to consider models and estimation methods suited to the data distribution when evaluating anti-bullying program effects, as these are often found to be small to moderate in size (Farrington & Ttofi, 2009; Merrell et al., 2008; Smith et al., 2004). Even estimation methods robust to non-normality will not produce unbiased estimates of standard errors when the data are as skewed as these with an over-inflation of zeroes (Muthén & Asparouhov, 2011). The usual remedy for skewed data, namely applying non-linear transformations such as logarithmic transformations to the data, will only assist with regards the non-zero values and the preponderance of values at the minimum will remain.

A further assumption of statistical analyses is independent observations. Even though methods may appropriately account for the dependency in the repeated measures obtained from participants in longitudinal studies, this assumption is violated in school-based studies as homogeneity exists between the students in a school. Students from the same school will have more similar bullying experiences and involvement than students from different schools. The clustering of students within schools needs to be accounted for in the analyses or the standard errors of the program effects will be underestimated, and potentially erroneous conclusions made as a result of inflated Type I errors (Murray, 1998). Approaches to accounting for the clustering of students within schools (and classes in schools, e.g. in primary schools, where the students remain largely with the same group throughout the school day), include the use of multilevel modelling, random effects models, generalised estimating equations (GEE) methods and robust standard error estimation (Freedman, 2006; Snijders & Bosker, 2011; Twisk, 2003). As the illustration of different statistical models to the application of skewed data is the focus of this chapter, these techniques will not be discussed further and only one method of accounting for clustering will be used in the analyses, namely sandwich estimates which are robust to the reduced variation in responses between students in the same school. This is the default method within the software package (Mplus) utilised to conduct the analyses in this chapter.

6.5 Statistical models for bullying outcomes

A plethora of statistical models are available for use when analysing data from anti-bullying program evaluations and it would be impossible to consider them all. The specific models considered in this chapter for the two outcomes and using the three approaches described above to analyse data from group-randomised trials, are as follows:

For the binary outcomes

- Binary logistic regression (cross-sectional analysis)
- Binary logistic regression (ANCOVA analysis)
- Logistic growth model

For the scale mean scores

- Tobit regression (ANCOVA analysis)
- Linear growth model
- Two-part growth model

The specific statistical models were chosen either for comparative purposes as those traditionally used for longitudinal data or because they are specifically designed for variables with skewed distributions such as the scale mean scores.

All these analyses assume comparability of or a lack of systematic differences between the study groups to enhance the validity of the intervention-control group comparison as a measure of the program effect. For example, ANCOVA analyses test for differences between the groups post-program while statistically controlling for pre-program differences. The post-program difference could be interpreted as the estimated program effect when comparing two students with the same initial score. This interpretation is only valid, however, if one can assume the two students would "develop" in the same way over time, if neither student were exposed to the program or alternatively if both were exposed (London & Wright, 2012; Wright, 2006). This implies there are no systematic differences between the study groups which would result in different trends in bullying behaviours over time or differential impact of the program in the two groups. An example of systematic differences is where the development of bullying perpetration over time in a group with higher levels of bullying as a norm differs from that of a group with a lower

norm and furthermore, due to these different norms, the impact of the program is not the same in the two groups. The same restrictions on interpretations apply to growth model and cross-sectional analyses. The discussions in this chapter assume comparability of the study groups, one way of achieving this is random allocation of sufficient numbers of schools to each group.

The binary outcomes were analysed using logistic regression, using three approaches. First, cross-sectional models were fitted at each time point to illustrate the simplest models that may be applied and corresponding with the analyses that are often conducted in meta-analyses of program impact studies. Second, ANCOVA models were fitted to the Time 2 and the Time 3 variables, controlling for Time 1 responses. These two sets of analyses illustrate the differing results that may be obtained when pre-program differences are present and are not controlled for. Third, binary logistic growth models were applied as a first example of how growth modelling differs from cross-sectional or ANCOVA analyses in terms of the manner in which program effects are parameterised and estimated.

In recognition of the need to apply models which accounted appropriately for the highly skewed nature of the mean scale scores, tobit regression and two-part growth models were chosen. Descriptions of these models follow. The linear growth models are also included for comparative purposes as these are still widely used, even though under these models the robust methods employed would not adequately account for the semicontinuous nature of the bullying outcomes (Muthén & Asparouhov, 2011).

Tobit regression

Tobit regression models fall within the class of censored normal models (Long, 1997). A linear model is fitted to an assumed continuous latent variable which has been censored by a random mechanism. The non-zero data are seen as observations on the latent variable above the censoring point. Zeroes occur when the value of the latent variable is not observed i.e. the values fall below the censoring point. The mean of the latent variable and the probability of being censored are modelled jointly, using a linear model for the uncensored data and a probit model for the censored data. As a normal error distribution is assumed for the uncensored data, it is often appropriate to transform these values to more closely approximate a symmetric or normal distribution. The regression coefficients for the predictors in a tobit regression are measures of the association between the predictor and the latent variable. In the case of an ANCOVA analysis comparing two study groups, as applied here, it is the estimated difference in bullying victimisation or perpetration

between the two groups when comparing two students with the same status pre-program. Some examples of the outcome variables to which these models have been applied are self-reported involvement in problem behaviours (Osgood et al., 2002; Wang, Selman, Dishion, & Stormshak, 2010), alcohol-related problems (Delva, Grogan-Kaylor, Steinhoff, Shin, & Siefert, 2007), and depressive symptoms (Perren, Dooley, Shaw, & Cross, 2010).

Two-part models

Alternate models that can be applied to semicontinuous data are two-part models (Olsen & Schafer, 2001). These models are based on the concept that two processes are operating – one which determines whether a response is zero or not, and another which determines the level of the observed variable if the value is above zero. These two processes lead to two outcomes, for example for victimisation, a binary outcome defining whether the person experienced any bullying behaviours or not, and a continuous outcome which is the extent of the exposure, if any occurred. For cross-sectional data, the usual binary logistic and linear (on the non-zero values) regression can be fitted separately, but the interdependence of the binary and continuous outcomes need to be accounted for in longitudinal data.

Two-part models fit two random effects growth models with correlated random effects. The zeroes are modelled in a logistic model and non-zero responses in a conditional linear model, and the covariance between the random effects links the two parts of the overall model. This covariance reflects the correlation between the two outcomes over time, for example, being exposed or not exposed to bullying on one occasion influences the extent of the exposure that occurs at other occasions and vice versa.

One is able to include different independent variables in the two parts of the model, allowing for different predictors of the likelihood of the behaviours from those that may be associated with the extent or severity of the behaviours. Once again it is appropriate to transform the non-zero values of the dependent variable, as the linear part of the model assumes a normal error distribution. The models have been applied to parent and teacher ratings of children's aggressive behaviour and victimisation (Kim & Muthén, 2009; Runions & Shaw; Runions et al.), adolescent self-reports of substance use (Lee, Mun, White, & Simon, 2010; Muthén & Asparouhov, 2011; Olsen & Schafer, 2001), the evaluation of a school-based program addressing substance use (Brown, Catalano, Fleming, Haggerty, & Abbott, 2005), and the modelling of coronary artery calcium, a possible contributor to cardiovascular disease (McPherson & Barbosa-Leiker, 2012).

Different conceptualisations of zeroes

Tobit and two-part models differ in the manner in which the zeroes in the data are conceptualized. Tobit regression treats the zeroes as left-censored or truncated, and therefore as proxies for values that cannot be detected based on the instrument used. They are viewed as values on the latent scale that fall between zero and one, i.e. low levels of victimisation / perpetration. In comparison, two-part models assume the data are generated through two processes and that the zeroes are valid self-representing values, i.e. there are simply large numbers of students who have no involvement in bullying behaviours.

Software and settings

To compare the results from the various statistical models, some consistency was sought with regard to other factors that may impact on parameter estimates, such as the method of accounting for clustering in the data, the particular estimator used in the software package for the applied model, and the manner in which the package dealt with missing data. For this reason, all analyses were conducted using Mplus 6.11 (Muthen & Muthen, 1998-2009).

The clustering of students within schools was accounted for through the use of clustered sandwich estimates of standard errors i.e. the "TYPE = COMPLEX" option in Mplus. The MLR estimator, which applies maximum likelihood estimation with robust standard errors, was used for all analyses as this estimator is somewhat robust to non-normality, and is used for two-part models and the default with the "COMPLEX" option. Full information maximum likelihood (FIML) strategies in Mplus assume data are "missing at random" (MAR) and ensure a minimum number of cases are excluded from the analyses due to missing values, thus maximising the use of the available data (Muthen & Muthen, 1998-2009). Lastly, the simplest models e.g. not allowing for random slopes, which would be most comparable between the different methods, were fitted.

The estimated study group differences were adjusted by controlling for the following demographic variables in all the models: gender, school SES (schools grouped as below the median and above the median A-index used to represent the socio-economic status of schools affiliated with the Catholic Education Office), school size (the sampled schools were divided into three equally sized groups according to the total number of students enrolled, namely as smaller schools with 710 students or less, medium schools and large schools with more than 800 students), and school type (co-educational, girls' school, boys' school).

In the tobit and two-part models, natural logarithm transformations were applied to the non-zero values. This is recommended to reduce the skew in the values above the minimum (Osgood et al., 2002). In the linear regressions, no transformations were applied as the arbitrary choice of a constant to add to the data to enable the calculations of log values can introduce bias (Osgood et al., 2002). Further, the fact that MLR estimates are stated to be robust to non-normality (Muthen & Muthen, 1998-2009) leads many researchers to apply these methods to raw data as highly skewed and semicontinuous as these.

Growth models in Mplus are formulated as multivariate models of latent variables which represent the intercept and slope for the trajectory of the outcome over time – the intercept represents the status of the outcome variable (victimisation or perpetration) at a particular time and the slope represents the "growth" or development in the outcome over time (Muthen & Muthen, 1998-2009). In two-part models, a set of intercepts and slopes is estimated for each part of the model. For example, for victimisation, the intercept in the binary part represents the odds of being exposed to any bullying behaviours at that particular point in time and the slope, the change in those odds over time. Similarly, for the linear part of the model, the intercept is the mean level of victimisation at that time and the slope is the change in the level or extent of victimisation over time for exposed students. In the context of growth models, the slopes are the most relevant parameters as measures of program effects.

In preliminary analyses, the linearity of the slope terms over the three data points in the growth models was assessed and this assumption was found not to hold, i.e. "growth" did not increase consistently from Time 1 to Time 3 (i.e. from the beginning of Year 8 to the end of Year 9). Therefore, all growth models were fitted with piecewise slopes. This also allowed for the explicit modelling and testing of the slopes between the two sets of time periods, i.e. from Time 1 to Time 2 and from Time 2 to Time 3. Thus, the development of the bullying outcomes for the period prior to the program to immediately post-program (i.e. from Term 1 to Term 3 of Year 8), but also for the period between the first and second post-program data collections (i.e. from Term 3 Year 8 to Term 3 Year 9), could be described. More importantly, variations between the intervention and control groups in that development could be compared, allowing for separate tests of program effects for each of the two periods of the study.

6.6 Analyses – Results and Discussion

Results from the six statistical models which were applied are given in Table 6-2 to Table 6-8, and the inferences from each discussed and contrasted with interpretations from others. Two sets of analyses are presented, the first correspond to analyses of the binary outcome variables and the second the scale mean scores. The binary variables classify students as having been bullied or not and having bullied others or not (every few weeks or more often in the term, since bullying is repeated behaviour). The scale mean scores are the students' level of exposure to and perpetration of various bullying behaviours (a combination of frequency and variety of behaviours). The regression coefficients, standard errors and significance tests given in the tables are for the study group variable, and in all instances the values are for the intervention group as compared to the control. All models also include gender, school SES, school size and school type to control for possible confounding effects from these variables – these results are not shown.

Analyses on binary variables

The results from the cross-sectional binary logistic regressions applied to the binary variables derived from the global questions, are presented in Table 6-2 (**Analysis 1**). Based on these analyses, there are no statistically significant differences between the intervention and control groups in the odds of being bullied or of bullying others at any of the time points. From the results for Time 1, one could assume that the groups did not differ prior to implementation of the program, hence the randomisation "worked" in terms of creating equivalent groups at the start of the study. Furthermore, the non-significant comparisons between the groups at the post program time points indicate the absence of any program effects.

Table 6-2. Analysis 1: Cross-sectional binary logistic regressions, study group results

Dependent variable	Coefficient [*OR*]	*SE*	*t*	*p*
Victimisation				
Time 1 (*n* = 3,034)	.159 [1.172]	.213	.744	.457
Time 2 (*n* = 2,988)	-.069 [.933]	.146	-.475	.635
Time 3 (*n* = 2,751)	-.075 [.928]	.122	-.614	.539
Perpetration				
Time 1 (*n* = 3,011)	.141 [1.151]	.180	.782	.434
Time 2 (*n* = 2,938)	-.418 [.658]	.230	-1.821	.069
Time 3 (*n* = 2,706)	.211 [1.236]	.182	1.162	.245

Note. The coefficient is the log odds of being victimised / perpetrating bullying (every few weeks or more often) in the intervention group relative to the control group, controlling for gender, school SES, size and type. * $p < .05$, ** $p < .01$, *** $p < .001$.

In the second set of analyses (**Analysis 2**), binary logistic regression is applied to the same binary variables as in the first analyses, however in the ANCOVA approach the Time 1 variable is included as an independent variable in the model. The odds ratios calculated here are the ratios of the odds of the outcome at Time 2 (or Time 3), controlling for pre-program dissimilarities between the groups. Results from these analyses (**Table 6-3**) indicate no differences with regard to victimisation at either post-program time point and none for perpetration at Time 3, but significantly lower odds of perpetration in the intervention compared with the control group at Time 2 ($p = .041$). Thus, when comparing two students with the same perpetration status at Time 1, students in the intervention group were 0.613 times less likely to report bullying others at Time 2 than those in the control group.

Table 6-3. Analysis 2: ANCOVA binary logistic regressions, study group results

Dependent variable	Coefficient [*OR*]	*SE*	*t*	*p*
Victimisation				
Time 2 (*n* = 2,910)	-.121 [.886]	.163	-.744	.457
Time 3 (*n* = 2,684)	-.114 [.892]	.126	-.902	.367
Perpetration				
Time 2 (*n* = 2,844)	-.490 [.613]	.240	-2.046	.041*
Time 3 (*n* = 2,622)	.219 [1.245]	.192	1.138	.255

Note. The coefficient is the log odds of being victimised/ perpetrating bullying (every few weeks or more often) in the intervention group relative to the control group, controlling for victimisation / perpetration status at Time 1, gender, school SES, size and type. * $p < .05$, ** $p < .01$, *** $p < .001$.

Binary logistic growth models were applied to the same data in **Analysis 3**. In Mplus, intercept and slope variables are modelled as a function of independent variables, in this case the study group variable is the only independent variable of interest. For binary logistic growth models the group coefficients for the intercepts represent differences between the groups in the log odds of the outcome (victimisation or perpetration) at a particular time. The group coefficients for the slopes represent the variations in the "growth" or development in those log odds in the groups over time. Piecewise slopes were applied here to estimate the group differences for each time period, i.e. over the six months from Term 1 to Term 3 of Year 8 (Time 1 to Time 2) and the twelve months from Term 3 Year 8 to Term 3 Year 9 (Time 2 to Time 3).

The trends in the log odds of victimisation and perpetration were first estimated fitting null models excluding any independent variables. There were significant positive increases for victimisation between the time points, the estimate for the mean slope for Time 1 to Time 2 was $M = 0.648$ ($SE = 0.165$, $p < .001$) and for Time 2 to Time 3 was $M = 0.213$ ($SE = 0.078$, $p = .006$). Thus, the log odds of being victimised increased significantly from Term 1 to Term 3 of Grade 8, and also over the twelve months from Term 3 of Year 8 to Term 3 of Year 9. Similarly, the log odds of bullying others also increased significantly across these two periods (Time 1 to Time 2: $M = 0.881$, $SE = 0.177$, $p < .001$ and Time 2 to Time 3: $M = 0.590$,

SE = 0.182, *p* = .001). The discrepancies between the slopes for the two time periods, for both victimisation and perpetration, are also evident from these estimates. Hence, the piecewise fitting of the slopes is appropriate.

The models were then fitted including the study group variable and the demographic variables. The models were repeated with the intercept changed to represent victimisation / perpetration status at a different time point, in order to obtain comparisons between the groups for each data collection point. As in previous tables, Summary of analyses on binary variables

Analysis 1 illustrates the approach often used in meta-analyses i.e. use of binary variables and testing for differences between groups in prevalence at different time points. These analyses would lead to the conclusion that the program had no effect on either bullying victimisation or perpetration at either post-program data collection. In comparison, conclusions from Analysis 2 results would be, while no effects were found for victimisation, the program was associated with significantly lower odds of perpetration immediately after the program (Time 2)†. This positive outcome was not sustained a year later (Time 3).

† Comment: To ensure the differing conclusions for perpetration at Time 2 were not a consequence of the slightly different samples used in the two sets of analyses (missing data at Time 1 resulted in fewer cases in the ANCOVA analysis), the cross-sectional logistic regression were redone based on the same sample (n = 2,844) as used in the ANCOVA logistic regression. Small changes in the odds ratio from .658 to .641 and the p value from .069 to .063 negated this explanation for the differing results.

Table 6-4 presents the results from the models for the study group variable only. No differences were found for victimisation, all group coefficients for the intercept and slope variables were non-significant. For perpetration, while the groups did not differ significantly at any of the time points (based on the intercept terms), the trends over time were different between the groups. The negative slope from Time 1 to Time 2 (coefficient = -0.718, p = .003) indicates a lesser increase in the intervention group from the beginning to the end of Year 8 – a positive outcome associated with the program. (Note the shift from a positive coefficient of 0.230 at Time 1 to a negative one of -0.488 at Time 2, i.e. non-significant but higher odds in the intervention than the control group in Term 1 and then lower at the end of the year.) However, the trends for the period from the end of Year 8 to the end of Year 9 indicate a reversal of the initially positive outcome, as the increase in the log odds of perpetration was steeper in the intervention group (coefficient = 0.796, p = .004).

Summary of analyses on binary variables

Analysis 1 illustrates the approach often used in meta-analyses i.e. use of binary variables and testing for differences between groups in prevalence at different time points. These analyses would lead to the conclusion that the program had no effect on either bullying victimisation or perpetration at either post-program data collection. In comparison, conclusions from Analysis 2 results would be, while no effects were found for victimisation, the program was associated with significantly lower odds of perpetration immediately after the program (Time 2)†. This positive outcome was not sustained a year later (Time 3).

† Comment: To ensure the differing conclusions for perpetration at Time 2 were not a consequence of the slightly different samples used in the two sets of analyses (missing data at Time 1 resulted in fewer cases in the ANCOVA analysis), the cross-sectional logistic regression were redone based on the same sample (n = 2,844) as used in the ANCOVA logistic regression. Small changes in the odds ratio from .658 to .641 and the p value from .069 to .063 negated this explanation for the differing results.

Table 6-4. Analysis 3: Binary logistic growth models, study group results

Dependent variable	Coefficient [*OR*]	*SE*	*t*	*p*
Victimisation				
Intercept: Time 1	.198 [1.22]	.290	.682	.495
Intercept: Time 2	-.104 [0.90]	.203	-.513	.608
Intercept: Time 3	-.129 [0.88]	.162	-.793	.428
Slope: Time 1 to Time 2	-.302	.310	-0.975	.329
Slope: Time 2 to Time 3	-.025	.143	-.171	.864
Perpetration				
Intercept: Time 1	.230 [1.26]	.224	1.026	.305
Intercept: Time 2	-.488 [0.61]	.290	-1.679	.093
Intercept: Time 3	.308 [1.36]	.239	1.287	.198
Slope: Time 1 to Time 2	-.718	.241	-2.985	.003***
Slope: Time 2 to Time 3	.796	.274	2.903	.004***

Note. The coefficients for the intercepts are the log odds of being victimised/ perpetrating bullying (every few weeks or more often) in the intervention group relative to the control group at each time point. The coefficients for the slopes are the differences in trends in the log odds between the intervention and control groups. All group coefficients are estimated controlling for gender, school SES, size and type.* $p < .05$, ** $p < .01$, *** $p < .001$.

Note the trends in the odds ratios in **Table 6-2**. For victimisation, the odds ratio switches from a value greater than one at Time 1 (1.172), indicating slightly higher odds of being bullied in the intervention than the control group, to values below one at Time 2 (0.933) and Time 3 (0.928), indicating lower odds post-program. The trend differs for perpetration, with increased (1.151), decreased (0.658) and then increased (1.236) odds at each of the time points respectively. While the odds ratios at each time point are not substantive and the cross-sectional comparisons not significantly different, the changes in direction of the odds over time are indicative of possible differences between the groups.

This illustrates the importance of accounting for baseline discrepancies, despite the use of randomisation in the study design, and such initial differences being tested and found not

to be significant. It also demonstrates the increased power ANCOVA analyses have in comparison to cross-sectional tests (Wright, 2006).These differences in interpretation also raise concerns regarding the conclusions from meta-analyses where study groups are often compared cross-sectionally.

The conclusions from the growth models (Analysis 3) concur with those of the ANCOVA models (Analysis 2), i.e. that the program is associated with an initial decrease in the odds of perpetration, but the group differences are not sustained in Year 9. The growth models also illustrate the reasons for the discrepancies between the cross-sectional (Analysis 1) and ANCOVA (Analysis 2) analyses for the perpetration variable. The growth models reproduce the results from the cross-sectional analyses – the coefficients for the intercepts in the growth models correspond to the coefficients in the cross-sectional analyses. Both sets of analyses indicate non-significant differences between the study groups at each of the time points when tested cross-sectionally. However, the growth models additionally assess variations in changes in the log odds over time between the groups.

The reduced increase from the first to the second time point in the intervention compared to the control group, is a result of the switch from higher to lower odds in the intervention at these two times. This "switch" is reflected in the ANCOVA result for Time 2, namely the significant difference at Time 2 controlling for Time 1 differences. Furthermore, the reduction of the odds in the intervention group to levels below those in the control group, could be an argument against interpretation of the observed effect as purely regression to the mean or a natural correction over time as a result of high initial values. This is especially the case as the students and schools were not originally selected on the basis of their involvement in bullying behaviours, e.g. schools with students at higher risk. Additionally, the schools were randomly assigned to the study groups and there is no evidence of systematic differences between the groups.

Analyses of scale mean scores

The mean scores from the scales were analysed using tobit, linear and two-part growth models. While similar results to those for the binary variables are expected since both are measures of bullying victimisation and perpetration, some differences in findings are also expected. The discrepancies between students' reports of these outcomes based on global questions and scale items have been well documented (Furlong, Sharkey, Felix, Tanigawa, & Greif-Green, 2010; Solberg & Olweus, 2003).

Tobit regression, which treats the data as censored values of a latent continuous variable, was applied in **Analysis 4**. The non-zero values were log transformed. The estimated mean values on the log scale for the latent (uncensored) victimisation variables are *M* = 0.19 (*SE* = 0.014) and *M* = 0.22 (*SE* = 0.016) for Times 2 and 3 respectively, and for the perpetration variables *M* = -0.02 (*SE* = 0.019) and *M* = 0.02 (*SE* = 0.019) respectively. Note that negative values for the latent variable on the log scale correspond to values between zero and one on the original scale, and a value of zero is equivalent to exposure to / perpetration of one type of bullying behaviour on 1-2 occasions.

As for the binary logistic regression ANCOVA analyses, four models were fitted, one for each of victimisation and perpetration for Times 2 and 3 respectively, including the relevant Time 1 variable as an independent variable in the model. The results from these analyses (**Table 6-5**) show significantly lower levels in the intervention group in the estimated mean of the latent variable at Time 2 for both victimisation ($p < .001$) and perpetration ($p = .027$), and no differences at Time 3.

Table 6-5. Analysis 4: ANCOVA Tobit regressions, study group results

Dependent variable	Coefficient	*SE*	*t*	*p*
Victimisation				
Time 2 (*n* = 2,924)	-.034	.006	-5.49	<.001***
Time 3 (*n* = 2,705)	-.029	.018	-1.60	.109
Perpetration				
Time 2 (*n* = 2,871)	-.048	.022	-2.21	.027*
Time 3 (*n* = 2,651)	.002	.024	.084	.933

Note. The coefficients are the estimated mean difference (on the log scale) in the latent victimisation/ perpetration variables between the intervention and control groups at each time point, controlling for the Time 1 scores, gender, school SES, size and type. * $p < .05$, ** $p < .01$, *** $p < .001$.

As for the logistic growth models, intercept and slope variables are modelled in linear growth models, but growth in mean levels of victimisation / perpetration over time rather than log odds are described. Initial tests of the trends over time were conducted by fitting null models by excluding any independent variables. Victimisation levels were found to increase on average from Time 1 to Time 2 (*M* = 0.116, *SE* = 0.015, $p < .001$) and from Time

2 to Time 3 (M = 0.095, SE = 0.026, p < .001). Significant increases were also evident for perpetration for both these periods, namely from Time 1 to Time 2 (M = 0.086, SE = 0.014, p < .001) and from Time 2 to Time 3 (M = 0.108, SE = 0.020, p < .001). Thus, based on the scale mean scores, mean levels of both victimisation and perpetration increased from the beginning of Year 8 to the end (Term 3) of Year 9.

Once again, group coefficients for each of Time 1 to Time 3 are presented (**Table 6-6**), to provide tests of differences in means between the groups at each point. For victimisation, no significant differences were found when comparing the groups at any of the time points. Whilst there was an overall increase in victimisation from Time 1 to Time 2 (as shown in the null model above), the significant negative coefficient for the first slope indicates "growth" in victimisation was slower from Time 1 to Time 2 in the intervention group than the control (coefficient = -0.56, p = .033). No difference in slope or trend between the groups from Time 2 to Time 3 was found. The findings are similar for perpetration, however the test for the first slope term did not reach statistical significance (coefficient = -0.47, p = .059).

Two-part growth models, which combine logistic and linear growth models were fitted to the scale mean scores in **Analysis 6**. The first part models the odds of being involved in bullying behaviours and the second the extent or severity of that involvement, if it occurs. The scale mean scores are dichotomised to zero versus any involvement (binary part) and the non-zero scores were log transformed (linear part). Coefficients for the study group variable are given for two sets of intercepts and slopes, one for each part of the model.

Note that the dichotomisation here differs from that used to create the binary variables analysed earlier. Here, the zeroes are modelled in the binary part, i.e. the data are split between those who were not involved at all and those who had some involvement. Thus for victimisation, the comparison here is between those who were exposed to any type of bullying behaviour, whether this occurred once or more often, versus no exposure at all and is *not* a comparison of bullied versus not bullied students, i.e. those who experienced such behaviours repeatedly. The linear part of the model explores the frequency and variety of behaviours to which students were exposed, for those with some involvement (non-zero values). The same interpretations apply for the perpetration of bullying behaviours.

Table 6-6. Analysis 5: Linear growth models, study group results

Dependent variable	Coefficient	*SE*	*t*	*p*
Victimisation (*n* = 3,115)				
Intercept: Time 1	.033	.031	1.053	.292
Intercept: Time 2	-.023	.024	-.938	.348
Intercept: Time 3	-.003	.056	-.059	.953
Slope: Time 1 to Time 2	-.056	.026	-2.137	.033*
Slope: Time 2 to Time 3	.020	.047	.412	.681
Perpetration (*n* = 3,113)				
Intercept: Time 1	.012	.015	.846	.397
Intercept: Time 2	-.035	.029	-1.201	.230
Intercept: Time 3	.015	.047	.316	.752
Slope: Time 1 to Time 2	-.047	.025	-1.891	.059
Slope: Time 2 to Time 3	.050	.034	1.456	.145

Note. The coefficients for the intercepts are the estimated mean differences in the victimisation / perpetration scale mean scores between the intervention and control groups at each time point, controlling for gender, school SES, size and type. The coefficients for the slopes are the differences in the slopes for the means for the intervention group compared to the control group, controlling for gender. * $p < .05$, ** $p < .01$, *** $p < .001$.

Note also that the linear part of the two part models does not equate to the linear growth models in Analysis 5. The linear models here are fitted to the non-zero scale mean scores (log transformed) whereas the models in Analysis 5 are applied to the full range of scores (including zero) assuming these are continuous normally distributed variables.

As before, null models were fitted initially to ascertain the trends over time in the outcome variables. The results are given in **Table 6-7**. All but one of the slopes differed significantly from zero. All were positive, denoting increases across the entire sample in either the odds of involvement or the mean level of involvement for each of the time periods. The non-significant increase corresponded to the slope for victimisation between Time 2 and Time 3, hence one can conclude that the log odds of experiencing any victimisation did not change from Term 3 Year 8 to Term 3 Year 9.

Table 6-7. Two-part growth models – Time trends

Dependent variable	Coefficient	*SE*	*t*	*p*
Victimisation (*n* = 3,115)				
Binary part				
Slope: Time 1 to Time 2	0.475	0.080	5.965	<.001***
Slope: Time 2 to Time 3	0.067	0.064	1.059	.290
Linear part				
Slope: Time 1 to Time 2	0.077	0.011	6.764	<.001***
Slope: Time 2 to Time 3	0.055	0.015	3.567	<.001***
Perpetration (*n* = 3,113)				
Binary part				
Slope: Time 1 to Time 2	0.508	0.086	5.894	<.001***
Slope: Time 2 to Time 3	0.320	0.102	3.138	.002***
Linear part				
Slope: Time 1 to Time 2	0.078	0.012	6.706	<.001***
Slope: Time 2 to Time 3	0.084	0.016	5.340	<.001***

Note. No independent variables included in models. * $p < .05$, ** $p < .01$, *** $p < .001$.

The results from the two-part models testing for group differences are given in **Table 6-8**. The two parts of the model are joined by allowing their random intercepts to covary. The covariance was significant in the victimisation model ($t = 7.9$, $p < .001$), indicating the odds of being bullied at one time point were highly correlated with the extent of victimisation at another, and vice versa. The covariance was also significant for perpetration ($t = 6.9$, $p < .001$), with a similar interpretation.

Focussing firstly on victimisation and the binary part of the model, group differences were found for Time 1 but not Times 2 or 3. The intervention students had significantly higher odds of being exposed to bullying behaviours than the control students at Time 1 (coefficient = 0.311, *OR* = 1.36, $p = .044$), indicating some pre-program differences. The group effect was significant for the first slope ($p < .001$) but not the second ($p = .968$). The reduced slope for the intervention group in the first period is reflected in the reversal in the direction of the group differences from significantly higher odds amongst intervention

students at Time 1 (coefficient = 0.311, *OR* = 1.36) to lower odds at Time 2 (coefficient = -0.102, *OR* = 0.903). Given the random assignment of the schools to the study groups, the switch from significantly higher to lower (rather than equal) odds at Time 2 in the intervention group may be indicative of a program effect, rather than due to regression to the mean. The non-significance of the second slope term signifies fairly parallel slopes in the two groups from Time 2 to Time 3, resulting in a similar odds ratio at Time 3 (coefficient = -0.101, *OR* = 0.904) as for Time 2. Therefore, the reductions in the first period in the likelihood of victimisation following implementation of the program, were maintained in the second period of the study.

The second linear part of the victimisation model looks at the extent of the bullying experiences, if they occur. No significant group effects were found in this part, for either the comparisons of the groups at each of the time points or the developments in level of victimisation.

Turning attention to perpetration and the binary results, the groups did not differ significantly at any of the time points. As for victimisation, the growth in the odds of being involved increased less steeply in the intervention than the control group between the first two time points (slope coefficient = -0.397, $p < .001$), but not the second (slope coefficient = 0.025, $p = .889$). The intervention group start with increased odds of perpetrating bullying at Time 1 (coefficient = -0.248, *OR* = 1.28) which change to reduced odds at Time 2 (coefficient = -0.147, *OR* = 0.863) and almost equal odds at Time 3 (coefficient = -0.024, *OR* = 0.976).

The extent of the perpetration, for those who perpetrated bullying behaviours, is tested in the linear part. Once again, tested cross-sectionally, the differences between the groups are non-significant. Nevertheless, the signs of the coefficients do shift between the time points, from a positive value (coefficient = 0.014) indicating slightly higher levels of perpetration in the intervention than the control group at Time 1, to a negative value at Time 2 (coefficient = -0.030) and then a positive value close to the Time 1 value at Time 3 (coefficient = 0.016). These changes are reflected in the significant slopes. For those involved, the development in the degree of engagement in bullying behaviours increased at a lower rate between Times 1 and 2 (slope coefficient = -0.045, $p = .010$), but at a higher rate between Times 2 and 3 (slope coefficient = 0.046, $p = .020$) amongst the intervention group students. Thus the gains associated with the program which were achieved following the first intervention period, were reversed in the second.

Table 6-8. Analysis 6: Two-part growth models, study group results

Dependent variable	Coefficient	*SE*	*t*	*p*
Victimisation (*n* = 3,115)				
Binary part				
Intercept: Time 1	.311	.154	2.016	.044*
Intercept: Time 2	-.102	.165	-0.617	.537
Intercept: Time 3	-.101	.143	-0.710	.478
Slope: Time 1 to Time 2	-.412	.106	-3.896	<.001***
Slope: Time 2 to Time 3	-.005	.113	-0.040	.968
Linear part				
Intercept: Time 1	.021	.022	0.931	.352
Intercept: Time 2	-.009	.012	-0.749	.454
Intercept: Time 3	-.008	.025	-0.303	.762
Slope: Time 1 to Time 2	-.030	.017	-1.710	.087
Slope: Time 2 to Time 3	.001	.019	0.061	.951
Covariance: Binary & linear random intercepts	.364	.046	7.885	<.001***
Perpetration (*n* = 3,113)				
Binary part				
Intercept: Time 1	.248	.187	1.327	.184
Intercept: Time 2	-.147	.192	-0.763	.445
Intercept: Time 3	-.024	.187	-0.127	.899
Slope: Time 1 to Time 2	-.397	.114	-3.479	.001***
Slope: Time 2 to Time 3	.025	.182	0.139	.889
Linear part				
Intercept: Time 1	.014	.014	0.995	.320
Intercept: Time 2	-.030	.021	-1.452	.146
Intercept: Time 3	.016	.031	0.513	.608
Slope: Time 1 to Time 2	-.045	.017	-2.588	.010*
Slope: Time 2 to Time 3	.046	.020	2.327	.020*
Covariance: Binary & linear random intercepts	0.273	0.039	6.920	<.001***

Note. Binary part: The coefficients for the intercepts are the log odds of being victimised/ perpetrating bullying (1-2 a term or more often) in the intervention group relative to the control group at each time point. The coefficients for the slopes are the differences in trends in the log odds between the intervention and control groups.

Linear part: The coefficients for the intercepts are the estimated mean differences in the victimisation/ perpetration scale mean scores (log transformed) between the intervention and control groups at each time point. The coefficients for the slopes are the differences in the slopes for the means for the intervention group compared to the control group.

All group coefficients are estimated controlling for gender, school SES, size and type. * $p < .05$, ** $p < .01$, *** $p < .001$

Summary of analyses on scale mean scores

Based on the tobit analyses (Analysis 4) one would conclude, controlling for any initial differences, the program was associated with significantly lower levels of victimisation and perpetration at Time 2, while no differences were found at Time 3. The linear growth model results (Analysis 5) would lead to a similar conclusion that the program was associated with a lesser increase in victimisation from Time 1 to Time 2 i.e. from pre-program to the first post-program measurement and not thereafter, but a differing conclusion that no impact on perpetration was apparent. Differing conclusions are expected as, despite the use of robust estimation methods, the linear models do not adequately account for the semicontinuous nature of the data (Muthén & Asparouhov, 2011), while the tobit regressions are formulated for censored outcomes. The discrepancy for perpetration is also likely in part a consequence of the two statistical approaches taken and hence the differences in the information contained within the parameters from the models. The tobit coefficient represents group differences at Time 2, controlling for initial values, i.e. comparing two students with the same level of perpetration at Time 1. The group coefficient for the slope in the linear growth model, compares the slopes or growth in perpetration in the two groups between Time 1 and Time 2. While these slopes may differ in part as a consequence of differences in the intercepts for each group, they do not control for any such dissimilarities in levels of perpetration at Time 1.

The two-part models (Analysis 6) give the most detailed analysis of the data, “pulling apart” the results from the other analyses. The conclusions regarding program effects drawn from these models would be that the program was associated with lesser increases in the likelihood of victimisation and perpetration between the first two time points, i.e. from pre-program to the first post-program measurement, and these differences were sustained in the second year of the study. For students exposed to bullying behaviours, the program did not have an impact on the frequency and variety of behaviours experienced. For

students involved in bullying perpetration, the extent of that perpetration "grew" at a slower rate in the intervention than the control group within the first period of the study, but at a faster rate in the second period so that the group differences at the end of the study were similar to pre-program levels.

As for the tobit regression analyses, the two-part models indicate the program is associated with positive outcomes for both victimisation and perpetration from the pre-program to the first post-program period of the study but not for the second period from Time 2 to Time 3. In fact the two-part models indicate a reversal of these positive outcomes for extent of perpetration, so that the group differences are similar to pre-program levels. Comparisons of the two-part growth models and the linear models in Analysis 5 are problematic, given in the latter the full range of scores are analysed and in the first separate models are fitted to the zero and non-zero values. For example, the growth in perpetration between Time 1 and 2 described by the single non-significant slope (p = .059) in the linear growth model (Analysis 5) is described by the non-significant (p = .087) and significant (p = .010) slopes in the binary and linear parts respectively, of the two-part model.

6.7 Considerations and Conclusions

The results and conclusions from analyses of a trial testing the impact of a program are dependent in the first instance on the outcome variable chosen and in the second on the statistical approach and model used to analyse that outcome. Hence, decisions regarding these factors need to be well-considered.

Outcome variables

Use of a global question or multi-item scale, and hence whether a categorical or continuous variable is analysed, is dependent on which is seen as a more valid operationalisation of the construct targeted by the program. The relative merits of each have been debated in the bullying literature (Cornell & Bandyopadhyay, 2010; Felix, Sharkey, Green, Furlong, & Tanigawa, 2011; Solberg & Olweus, 2003), without consensus concerning which provide more valid measures of bullying behaviours. In *statistical terms*, the greater variation in scale mean scores and resultant increased sensitivity to change is an advantage of continuous measures. Analyses of binary outcomes using logistic models, not only have reduced power to detect an effect but dichotomisation of the original question means a

loss of information regarding the extent of the involvement in bullying victimisation and perpetration. Based on statistical criteria, in general, continuous rather than categorical variables are preferred. However, these are not the only considerations when choosing measures to implement in evaluation studies.

Which statistical approach and model to use?

The choice of statistical approach depends in the first instance on the substantive question being asked. Is the focus of the analysis on differences in the changes within each group over time, or are differences between the study groups at particular time points of most interest? The first question has to do with the developmental trends in bullying victimisation / perpetration over the course of the study and whether the study groups differ on these, i.e. whether the program has impacted on the rate of change. The second compares outcomes for the study groups at particular time points, usually times where it is pertinent to assess the impact of the program, i.e. tests whether the prevalence or level of bullying victimisation / perpetration is the same at specific points after program implementation. The first question is best answered with growth models and the second through the use of ANCOVA-type models.

While inference from all of these analyses that observed group differences are indicative of program effects, is based on the assumption the study groups do not differ systematically, the cross-sectional testing also disregards any randomly occurring pre-program differences. In group-randomised trials where schools rather than students are allocated to conditions, particularly in studies where the numbers of schools randomised are not large, small random differences may occur. Results presented here based on the SSP study, highlight the need to control for pre-program differences, even if tests of these are not significant in cross-sectional analyses. In these circumstances, ANCOVA analyses are preferred to cross-sectional analyses as the program effect is estimated as the difference between the two groups as if one were comparing students with the same pre-program score. In studies with small numbers of schools, say less than five per study group, where the groups differ significantly pre-program on the outcome variables, regression to the mean needs to be considered as the explanation for any observed effects over time.

In comparison with growth models, ANCOVA-type analyses provide straightforward, easily understood tests of program impact. For example, the tobit regression analyses provided clear conclusions regarding program impact. An advantage of the two-part models is that both the study group differences at each time point and the differing trends over time are

explicitly modelled. When assessing program effects in growth models, the slopes are interpreted in conjunction with the intercepts terms. If the pre-program intercepts are equal so that the two groups start from the same point, then a difference in slope corresponds to a direct estimate of a program effect. If the starting points differ, then the slope estimate may incorporate effects plus other changes due to the initial differences between the groups.

The decision between tobit and two-part models depends not only on the substantive question being addressed and the number of data points available, but is also based on the conceptualisation of the zeroes. Tobit regression is appropriate if the data are seen as observed values of an underlying continuous latent variable and the zeroes are levels of bullying victimisation / perpetration that fall below a level that is detectable by the instrument used. In contrast, for the application of two-part models the data are seen as having arisen from two separate processes, the first of which generates the zeroes which represent no involvement in bullying behaviours. These models are applicable when it is of substantive interest to determine the program impact on the occurrence of bullying as well as on the extent of the bullying when it occurs. Effects on one or other of the two processes may be masked when using models other than two-part, such as linear or tobit regressions. The separate modelling of, for example, the probability of being exposed to any bullying from the extent or severity of that bullying for exposed students, is an attractive feature of two-part models which leads to greater insights into the specific impact of the program.

Limitations

A limited number of statistical models were discussed in this chapter – those frequently applied and those designed for highly skewed data, but not as yet commonly used for the analysis of bullying outcomes. This chapter is not intended to be a comprehensive review of all possible methods. The presented analyses were chosen to illustrate and highlight some of the major differences between statistical approaches and models.

As the focus in this chapter was to illustrate the use of differing statistical approaches and certain models, one estimator and method for accounting for school-based clustering was consistently applied in the analyses. Despite this potential limitation and the use of a single data set, the principles and findings presented here are relevant to other estimation methods and contexts.

6.8 Final Conclusions

Data which are non-normal and where a significant number of individuals are at the lowest value indicating absence of the behaviour, require approaches different than those traditionally utilised. The statistical model and approach taken needs to be carefully considered, particularly in analyses estimating program effects. Binary logistic regression, linear and tobit regression and two-part models were applied as cross-sectional analyses, ANCOVA-type analyses and/or growth models to bullying outcomes to illustrate differences between these approaches and models.

Cross-sectional analyses are not recommended for the analysis of longitudinal data. Linear growth models which rely on assumptions of normality, even with the use of robust estimation methods, are also not recommended. Methods such as two-part models and tobit regression analyses, which specifically account for semicontinuous and skewed data, are preferred. Detailed information regarding the nature of the impact of a program can be gained from two-part growth models, however users need to be aware that the application and interpretation of the models is relatively complex. Tobit models are straightforward to apply and give single tests of program impact, but are less useful as the available data increases beyond about three time points.

To interpret observed group differences as effects associated with the implemented program, the intervention and control groups need to be comparable, such that one can assume the trends in bullying behaviours in the control group approximate the trends that would have occurred in the intervention group had the group not been exposed to the program. Additionally, a lack of systematic differences between the groups implies that the program would have the same impact on average in the control as it does in the intervention group. The determination of program effects based on non-comparable groups is severely limited as one can no longer assume that the trends in the two groups would be the same and alternate explanations for any observed differences cannot be discounted.

A number of factors will be taken into consideration when a researcher makes a choice as to the statistical approach and model to apply to the analyses of their data. Ideally the decision will be based on considerations of the substantive question to be answered and informed by the distribution of the data and validity of the assumptions of the chosen model.

Chapter 7 : General discussion

7.1 Introduction

The short- and long-term negative consequences of bullying behaviours make the implementation of effective programs in schools to prevent and respond well to bullying an imperative. Significant research activity has resulted in the development of numerous school-based programs and evaluations of their impact. These individual studies have provided conflicting evidence as to the efficacy of anti-bullying programs in schools or indications of small positive effects (Ferguson, San Miguel, Kilburn, & Sanchez, 2007; Merrell, Gueldner, Ross, & Isava, 2008; J. D. Smith, Schneider, Smith, & Ananiadou, 2004; P. K. Smith, Ananiadou, & Cowie, 2003; Ttofi & Farrington, 2011; Vreeman & Carroll, 2007). To assist education systems and schools choose and implement policies and programs that reduce bullying behaviours and improve students' mental and social health, "program evaluators and prevention scientists have a critical and pressing role to provide clear and accurate information" on programs (Ryan & Smith, 2009, p.256). Several of the reviews of anti-bullying program evaluations conducted since 2003 have identified the shortcomings of many of the evaluations. Factors proposed or found to account for the inconsistent findings on program effects included those related to the program and its implementation, as well as methodological limitations (Hahn Fox, Farrington, & Ttofi, 2012; Jimerson & Huai, 2010; Ryan & Smith, 2009; P. K. Smith, 2011). Five of these major methodological challenges have been investigated in this doctoral research. The implications of the findings are discussed below.

7.2 Significant findings

Detailed findings and discussion related to the first five research questions addressed in this thesis are given in each of Chapters 2 to 6. The most pertinent to the evaluation of anti-bullying programs are summarised below. This chapter integrates these findings and discusses their implications for evaluators of anti-bullying programs and thereby, addresses Research Question 6 of this doctoral work.

1. What are the psychometric properties of the Forms of Bullying Scale (FBS) measuring bullying victimisation and perpetration?

   Responses to the FBS were found to be valid and concurrently reliable measures of self-reported frequency of bullying victimisation and perpetration for adolescents. Additionally, the responses were demonstrated to be measurement invariant over time. For evaluators who concur with the approach taken in the FBS to measuring bullying, the characteristics and psychometric properties of the FBS make it a suitable instrument for use in anti-bullying program evaluations in middle and secondary schools.

2. To what extent do shifts in perceptions of bullying occur differentially in intervention and control groups in anti-bullying program trials?

   Whilst response shift bias was not present in the Cyber Friendly Schools Project data, it is still a potential source of bias in other studies and can feasibly occur for bullying outcomes. As demonstrated in this thesis, testing for the presence of response shift bias can be easily achieved using two methods, namely analyses of retrospective pre-test questions and tests of measurement invariance using confirmatory factor analyses.

3. What are the implications for bullying-related research of requiring active only parental consent versus active-passive parental consent for student participation?

   Active only parental consent procedures produced a biased sample, under-representative of students who bullied others and with less pro-social behaviours. Additionally, biased estimates of correlations between bullying victimisation and other outcomes were obtained. Hence, active only parental consent procedures have the potential to bias estimates of program effects.

4. What are the sizes of the clustering effects present in school-based studies of bullying outcomes and how should these be accounted for when designing such studies?

   The estimated intraclass correlations were typically less than 0.1 in value, however their influence in reducing precision and the power of a study should not be discounted when large numbers of students are sampled per school, as is the case for school-based program evaluations. Based on the estimated values for the bullying outcomes, the

sample sizes for school-based program evaluations need to be increased by a factor of at least 1.5 or substantially more to account for the reduction in power resulting from the clustering of students in schools. A minimum of 40 schools (20 per study group) is recommended in controlled studies to ensure sufficient power to detect an important program effect and ensure the comparability of intervention and control groups.

5. Which multivariable statistical methods are appropriate for analysing the frequency of bullying behaviours?

   Two-part growth models are both suited to the analysis of bullying outcomes and useful methods. They enable identification of the specific impacts of the program, namely distinguishing between the program's effect on trends in the odds of the occurrence of bullying behaviour and its effect on trends in the extent or frequency with which the behaviour occurs. A viable alternative for identifying program effects, is ANCOVA tobit analyses, which have the advantage of ease of interpretation.

6. What are the implications for research practice of these major methodological challenges for research evaluating school bullying prevention and reduction programs?

   This research question is addressed in detail in the next section, which forms the basis of the recommendations provided in Section 7.5.

7.3 Implications for anti-bullying program evaluations

The implications for research practice from the above findings, as they pertain to specific threats to the construct, internal and statistical conclusion validity of anti-bullying program evaluations, are presented here. A discussion of study designs is also provided.

Construct validity

Apart from issues related to the validity of the program and its implementation, the constructs for which it is most relevant to demonstrate validity when conducting evaluations of anti-bullying programs, are the outcomes or targets of the program, namely bullying victimisation and perpetration. Construct validity refers to the extent to which the chosen operational definition can be generalised to the construct about which inference is

to be drawn (Farrington, 2003). In this instance, the measurement validity or extent to which responses to an instrument are measures of behaviours correspondent with the definition of bullying.

Evaluators of anti-bullying programs are faced with a number of considerations when choosing an instrument for their study. The first consideration is the extent to which the instrument will elicit responses which meet the three criteria of repetition, intent to harm and a power imbalance, which differentiate bullying from other forms of aggression (Bovaird, 2009; Cornell & Bandyopadhyay, 2010; Jimerson & Huai, 2010). Incorporation of repetition and intent are usually relatively straightforward – frequency of involvement can be assessed through relevant response categories and, as is the case with the FBS, the items can be specifically worded to convey intent using phrases such as "to hurt" and "deliberately". The third criterion is more problematic. For example, the authors of the California Bullying Victimization Scale attempted to measure the power imbalance by including a list of ways in which the perpetrator may be in a more powerful position than their target (Felix, Sharkey, Green, Furlong, & Tanigawa, 2011). By necessity the list is limited and the terms used, not necessarily ideal as there are many possible causes of the power differential (Greif & Furlong, 2006; Rigby, 2002; P. K. Smith, 2012). Perhaps it is more easily specified in terms of its consequences, namely that the targeted person has difficulty in stopping the behaviour. This is one of the many ways in which this criterion is thought to be conveyed in the FBS (Shaw, Dooley, Cross, Zubrick, & Waters, In press). Program evaluators will need to assess the extent to which the instrument they choose, incorporates the three criteria which define bullying and will elicit responses which validly reflect involvement in bullying behaviours.

Relevant to this consideration of the three criteria, is whether a definition of bullying will be provided and if the term "bullied" will be used in the wording of the questions. There is ongoing debate in the research literature as to whether these latter practices enhance or reduce the validity of responses to an instrument (Greif & Furlong, 2006; Griffin & Gross, 2004; Kert, Codding, Tryon, & Shiyko, 2010; Ortega et al., 2001; P. K. Smith, Cowie, Olafsson, & Liefooghe, 2002; Solberg & Olweus, 2003; Ybarra, Boyd, Korchmaros, & Oppenheim, 2012). This debate has led to the development of a variety of instruments to measure involvement in bullying behaviours and, as highlighted by many of the reviewers of evaluations, this lack of consistency between measures has hampered the comparison of

results between studies. In the absence of consensus and further research, the onus is on the researcher to choose the approach they deem most valid.

Apart from the wording and format of questions, another consideration for evaluators when choosing an instrument is the data source or informants who will be used. Differences between self-report, peer- and teacher report may be conceptualised as measures of different constructs or at least different operationalisations of a bullying construct (Bovaird, 2009). Where self-report may be seen as a measure of self-perception, peer and teacher report may be more aligned with social reputation (Hymel, Wagner, & Butler, 1990). Self-report is arguably the more appropriate measure for victimisation, as the perspective of the person experiencing the bullying is most pertinent when investigating the consequences of that victimisation (Felix et al., 2011; Juvonen, Nishina, & Graham, 2001). Peer or teacher report may be more appropriate for perpetration as it is less reliant on a respondent's self-concept which may lead to under-reporting (Cornell & Bandyopadhyay, 2010). The use of multiple data sources to obtain measures of bullying outcomes was a common recommendation of the reviewers of program evaluations (Baldry & Farrington, 2007; Jimerson & Huai, 2010; Ryan & Smith, 2009; J. D. Smith et al., 2004; P. K. Smith et al., 2003) and the supplementation of self-report with data from other sources would provide more robust evidence of program impact. The randomised controlled trials of the KiVa program in Finnish schools (Kärnä et al., 2011) and the Steps to Respect program in California (Brown, Low, Smith, & Haggerty, 2011) are two of the few studies to utilise multiple informants.

The format or type of question that will be used is a further consideration for an evaluator. Relevant to the decision is the manner in which bullying victimisation and perpetration are conceptualised by the researcher. If these are seen as categorical outcomes, then single item global questions are appropriate as people can be categorised according to their responses (Solberg & Olweus, 2003), for example as "bullies", "victims", and by combining the questions as "bully-victims". Where bullying victimisation is understood as a level of exposure to and bullying perpetration as level of involvement somewhere along a continuum in such behaviours, multi-item scales as measures of continuous latent variables are more appropriate measures. From a practical point of view, when estimating program effects, scores on a continuous scale provide greater sensitivity to change than categorical scores. With an understanding and operationalisation of bullying victimisation and perpetration as continuous outcomes, for a program to be seen to have a significant effect,

a shift along a continuum in the levels of involvement in the behaviours would be required. In contrast, use of categorical measures would require a shift of participants from one group to another.

Construct validity not only applies to the adequacy of the measures of the theoretical constructs of bullying victimisation and perpetration, but also the time frames within the operational definitions (Portney & Watkins, 2000). Thus, a fourth consideration when choosing an appropriate instrument is the referent time period for which the respondents are reporting. This period is important as it impacts on the validity of the measures of bullying involvement, but also in terms of the validity of the conclusions regarding program effects. In the first instance of measurement validity, the time frame chosen should be one which maximises a respondent's ability to accurately recall events and minimises reporting subject to memory distortions (Bovaird, 2009). The eight to twelve weeks prior to completion of the questionnaire have been recommended as a suitable period (Ortega et al., 2001; Solberg & Olweus, 2003). As an example, the FBS refers to the last term at school (usually a period of 10 weeks in Australian schools) as this forms a natural period of time within which students can more accurately recall their experiences. Importantly, the timing of post-program data collection(s) need to be scheduled such that the period on which the participants report provides a sufficient time frame for the program to have an effect. Given the limited periods within which data may be collected in schools without undue disruption to educational activities, careful planning of the timing of implementation of program components and collection of outcome data in a trial is required.

In summary, evaluators of anti-bullying programs need to choose an instrument most closely aligned with the operational definition they feel will produce valid measures of the constructs of bullying victimisation and perpetration. Careful consideration will need to be made of the most appropriate data sources or informants; the extent to which the instrument incorporates the three characteristics of bullying; whether the behaviour will be named and a definition provided; and the referent time period that will be used.

Internal validity

Two threats to internal validity, specifically addressed in this doctoral research, will be discussed here. These are response shift bias and bias resulting from parental consent procedures, which relate to history effects and selection bias respectively.

History effects

History effects, particularly when they occur differentially in the intervention and control groups, can bias the results from an evaluation study. A form of differential history effect is response shift bias (Sprangers & Schwartz, 1999), which may occur as a result of exposure to the program amongst the intervention group students and result in changes to the reporting of involvement in bullying behaviours, which are not present in the control group.

As described in Chapter 3, response shift regarding the bullying construct may occur due to reconceptualisation (a change in understanding of what bullying is), or reprioritisation (a change in the perceived importance or severity of different forms of bullying), and/or recalibration (a shift in the perceived frequency that different forms of bullying occur). One type of response shift may result in another, for example, recognition of indirect forms as bullying behaviours (reconceptualisation) may lead to a change in perception of the frequency such incidences occur (recalibration). Two methods for testing for response shift bias, namely analyses of retrospective pre-test questions and tests of measurement invariance using confirmatory factor analyses of responses to the FBS, were illustrated using pre- and post-program data from the CFSP.

Researchers have proposed sensitisation or raised awareness may result in increases in reported involvement in bullying behaviours (Nixon & Werner, 2010; Orpinas et al., 2000; J. D. Smith et al., 2004; P. K. Smith et al., 2003). Based on the retrospective pre-test data in the CFSP, however, there was no evidence of "new" reporting of victimisation or perpetration, as student report of non-involvement in the pre-program period in bullying behaviours was remarkably stable over time (with consistent reporting of non-involvement for 93% to 96% of respondents). Thus, the program did not lead to a change in the way these students perceived their past experiences/behaviours and the vast majority remained of the view they had not been bullied or bullied others.

However, as the majority of students remain uninvolved in bullying behaviours over time, the potential for response shift may be greater in the relatively small groups of frequently involved individuals who are the primary focus of a program. For example, as a result of the intervention a response shift may occur in the subgroup of frequent perpetrators within the intervention group, without a similar change being present amongst the frequent perpetrators in the control group. Whilst not significant, there was some indication of possibly differing trends in the study groups for the subgroup of students who initially reported bullying others (with a four-fold difference in odds between the groups,

equivalent to an effect size of 0.8). This shift in reporting in a relatively small group may be sufficient to bias the estimates of the program effects on actual behaviours.

Although little evidence of response shift bias was found in these data, low levels of implementation of this multi-faceted program, a focus on cyberbullying (rather than bullying in general), and low numbers of students reporting involvement in bullying behaviours may, in part, be explanations for the results. The null results found in the CFSP should, therefore, not be seen as definitive evidence that such a bias will not be present in other evaluations, particularly when the participants may not have an established understanding of bullying (the three criteria which define bullying and the different forms it comprises) and/or the program aims to achieve a common understanding of bullying amongst students and staff in schools (e.g. Cross et al., 2011). Evaluators of anti-bullying programs should consider testing for response shift bias in their studies, especially as even small response shifts can lead to biased estimates of true change (Schwartz et al., 2006). This is easily achieved through the inclusion of retrospective pre-test questions in post-program questionnaires (Barclay-Goddard, Epstein, & Mayo, 2009) and by testing for the different types of response shift using confirmatory factor analyses to assess measurement invariance over time (Oort, 2005). Given the different insights that may be gained from the two methods of testing for response shift, namely average group change versus change within specific subgroups of students, it is recommended that evaluators consider using both.

The inclusion in the study of a comparable control group is important when testing for response shift bias. No response shift was observed in the intervention group in the CFSP data, but had there been, for example, a reconceptualisation or recalibration in the intervention group, without parallel tests in a control group, it would not be possible to determine whether the observed changes in reporting were due to exposure to the program or other effects. Similar shifts in the intervention and control groups would indicate effects common to both, such as maturation or testing effects. Response shifts in the intervention but not the control group, would more likely be due to program exposure.

Selection bias

The possibility of a biased sample resulting from selection bias is the second threat to internal validity considered in this section. Selection bias in school-based trials can result from a number of sources, including low school recruitment rates, low student participation rates and attrition of students across the course of the trial – a sample will be

biased whenever the participants differ systematically from those not participating. A major potential source of bias is the procedure whereby parental consent is obtained (Courser, Shamblen, Lavrakas, Collins, & Ditterline, 2009; Tigges, 2003; Unger et al., 2004). Currently, many education sectors and ethics review boards are mandating active only parental consent for research of all levels of risk. That is, student participation is contingent upon not only their consent, but receipt of a signed consent form from their parent or carer indicating their consent to the young person's participation. Under these conditions, procedures which allow for passive or opt-out parental consent are not sufficient, even though the available evidence indicates a low risk of harm to students from completion of health surveys (Langhinrichsen-Rohling, Arata, O'Brien, Bowers, & Klibert, 2006; Leakey, Lunde, Koga, & Glanz, 2004). Passive consent procedures are recommended for low-risk research to avoid non-response bias (Lacy et al., 2012; Stubbs & Achat, 2009).

The bias that can result from active only parental consent procedures was illustrated in the study reported in Chapter 4. Students with active consent, 35% of the respondents, differed systematically from those with passive consent on a number of outcomes, including being older and less academically competent. Most importantly, they were less likely to engage in problem behaviours, more specifically bullying others, and to report less pro-social behaviours. Furthermore, the correlations between bullying victimisation and certain social-emotional variables differed in magnitude in the groups of students with and without active parental consent, indicating that biased estimates of these correlations could result from samples comprising only students with active parental consent.

Thus, a consequence of mandating active only parental consent procedures is the likely exclusion of the perpetrators of bullying behaviours, an important target group for an anti-bullying program. This is particularly pertinent as parental consent is usually sought for the data collection component of an evaluation of a universal anti-bullying program and not for participation in the program itself, since the latter is at the discretion of the school principal or school authority. Under-representation of the students most likely to shift their behaviour in the sample from which data are collected, could feasibly lead to an underestimate of the impact of an effective program.

Under these active only conditions, evaluators of anti-bullying programs need to implement numerous strategies to maximise response rates, within the constraints of their research funding and the guidelines of their ethical review boards. Where limited resources are available, researchers may simply decide the research is impracticable and the

potential for biased samples and misleading results too great to justify the conduct of the evaluation.

Statistical conclusion validity

Methods for countering the two major threats to valid conclusions from statistical tests were investigated in this doctoral research, namely determination of a sufficient sample size and use of statistical methods appropriate to the distribution of the data.

In evaluations of school-based programs, both the sample size calculations as well as the analyses need to account for the clustering of students in schools, and for studies in primary schools where students largely remain with the same teacher and peers throughout the school day, classroom level clustering. Given the program is implemented at a school or at least a classroom level, the bullying experiences of students at the school will be interdependent. Ignoring this lack of independence in the data will lead to underpowered studies and over liberal statistical tests (Donner & Klar, 2004; Heeringa, West, & Berglund, 2010; Murray, 1998; Murray, Varnell, & Blitstein, 2004).

A vital step in planning a program evaluation is the determination of an adequate sample size to ensure the study has sufficient power to detect a program effect of a certain size. The required sample size for school-based studies is dependent on reliable estimates of the intraclass correlations (ICC's) or levels of homogeneity between the students in a cluster (Murray et al., 2004). The estimates of these correlations specific to bullying outcomes reported in Chapter 5, can be utilised for this purpose. Although typically less than 0.1 in value, their influence in reducing precision and the power of a study cannot be discounted when large numbers of students are sampled per school, as is the case for school-based program evaluations. The sample sizes for such studies need to be increased by a factor of at least 1.5 or substantially more (dependent on the numbers of students recruited per school and the specific ICC value) or the lack of power could lead to a failure to detect an important program effect. Furthermore, the ICC's will increase in value under active only parental consent conditions, due to the increased homogeneity of the recruited sample, implying a larger sample will be required to achieve sufficient power than would be the case under less stringent conditions. A straightforward method of determining the sample size accounting for school-based clustering is presented in Chapter 5.

Apart from the requirement of sampling sufficient numbers of students, the numbers of schools in a trial is of vital importance as this is the sample size most relevant to the

estimation of a program's effect. While the program may aim to change students' behaviours, outcomes measured at the student level, the program is implemented at a school level and exposure to the program is a contextual or school level variable. With a limited number of schools, the program effect cannot be estimated with precision. Traditionally, a sample of about 20 units is required to meet normality assumptions and, from a practical point of view, this is about the numbers of schools in which the program would need to be implemented to test for program effects. Where, as is commonly the case, the estimation of the program effect and its standard error relies on large sample theory, for example when robust sandwich estimation is used, a minimum of 40 schools (20 per study group) has been recommended (Donner & Klar, 2004; Murray et al., 2004).

The assignment of insufficient numbers of schools to the intervention and control groups reduces the comparability of the study groups, even if the schools are randomised to the groups (Hahn Fox et al., 2012). To benefit from an experimental design and randomisation, an evaluation study must include sufficient schools per group so that possible confounding factors are not differentially present in the intervention and control groups. Ideally, the only difference between the groups should be exposure to the program. Hence, larger numbers of schools not only provide power and meet the requirements of statistical methods which rely on large samples for unbiased estimates of program effects and their standard errors, but also protect against biases which threaten the internal validity of a study.

A further reason for including many schools in the intervention group, is to enable the conduct of dose-response analyses as such analyses assist in corroborating a program effect (Olweus, 2005). In evaluation trials there are often differences between schools in the level of implementation of the program (e.g. Baldry & Farrington, 2007; Vreeman & Carroll, 2007). Unless the study comprises a large number of schools that adequately represent the variability in effort invested by the schools or could, if appropriate, be divided into groups of about five per level of implementation, an exploration of different doses of the program on bullying behaviours is not possible. Of course, careful monitoring of the implementation of the program in the schools to determine program dose, is also required.

The second major threat to the conclusions from statistical analyses considered here, is violation of the assumptions of the applied statistical methods. This is a particular issue in the analysis of major outcomes used to assess anti-bullying program effects, namely the

frequency of involvement in bullying behaviours. Since many students are not involved in bullying, either as a victimised student or a perpetrator, these bullying outcomes are highly skewed with a preponderance of values at the minimum.

As discussed in the section addressing construct validity, an evaluator will choose the measure, i.e. a global single item or multi-item scale, which best fits their conceptualisation of the bullying constructs targeted by the program. This understanding together with statistical considerations regarding the relative power and sensitivity to change associated with categorical versus continuous variables, will determine the outcome measures used in an evaluation. When the outcomes are binary or categorical variables, logistic regression methods are applicable. As these methods are designed for categorical data, no assumptions related to the symmetry of the data distribution are made. A number of approaches could, however, be taken to the analysis of composite scores from multi-item scales analysed as continuous outcomes. Since the distributions of these variables do not meet the usual assumptions of normality and homoscedasticity of traditional statistical techniques, specialised methods are required to ensure valid inferences are made (McClendon, 1994; Muthén & Asparouhov, 2011; Osgood, Finken, & McMorris, 2002; Vittinghoff, Glidden, Shiboski, & McCulloch, 2011).

The application of two alternate approaches to the analysis of highly skewed bullying outcomes, namely tobit regression (Osgood et al., 2002) and two-part growth models (Olsen & Schafer, 2001), were illustrated in Chapter 6. Both of these methods are particularly suited to these semi-continuous data distributions and their use is hence less likely to result in the erroneous inference that may result from applying methods for which the assumptions are clearly violated.

The advantages of applying tobit regression within an ANCOVA framework, is the ease of interpretation of the results. The program effect is estimated as the post-program difference between the study groups as if one were comparing students with the same pre-program score. The focus is on testing for group differences at particular post-program time points. This approach is particularly insightful when the data collections have been scheduled at points which allow for short- and longer-term program impact.

The ANCOVA-type approach is of most use when the study comprises two to three data collections – where observations on many time points are available, growth models which estimate trends over time are a natural choice. Although the interpretations from two-part

growth models are less straightforward than those from the tobit analyses, they allow for greater insights into likely program impact. Two-part models enable the evaluator to distinguish between the program's effect on trends in the odds of the occurrence of bullying behaviour and its effect on trends in the extent or frequency with which the behaviour occurs. Thus, the evaluator can determine whether the program resulted in reductions in the percentages of students involved in bullying behaviours as well as the impact on the level of the bullying experienced / perpetrated by involved students. Such information enables program developers to reassess the content of the intervention. For example, a program associated with changes in the odds of bullying perpetration but not the frequency with which perpetrators bully others, may be changed to incorporate more strategies targeted at students who bully others, to reduce their levels of perpetration.

Study design

The focus in this doctoral research has been on experimental study designs, in particular for school-based evaluations, the use of group-randomised controlled designs. This focus is appropriate as these designs are best able to provide valid inference on program effects (Donner & Klar, 2004; Murray, 1998) and, as recommended by the reviewers of anti-bullying programs (Baldry & Farrington, 2007; Merrell et al., 2008; Ryan & Smith, 2009; J. D. Smith et al., 2004; Ttofi & Farrington, 2011), are preferred above designs which do not incorporate a control group or randomisation of schools to study groups.

The valid interpretation of results from program evaluations rely on these strong study designs to protect against biases which threaten the validity of the interpretation of the findings as evidence for or against a program's effects. In the absence of randomisation to intervention and control groups, together with longitudinal data collected on at least two occasions, that is prior to and after the program has been implemented, plausible alternate explanations for observed effects cannot be discounted. Furthermore, for causal inference, i.e. to interpret group differences as program effects, it is necessary that the changes in the control group accurately represent what would have occurred in the intervention group had the program not been implemented and in addition, that the program would have the same impact on average in the control as it did in the intervention group (London & Wright, 2012; Wright, 2006). A lack of systematic differences between the groups is required to enable these inferences to be made. As discussed above, random assignment of sufficient numbers of schools to the study groups is a means of ensuring this comparability of the groups.

Group-randomised trials with large numbers of schools may not be practically possible within a given research budget. Such studies not only require resource expenditure on the collection of data from participants, but also in training and supporting school staff to implement the program with integrity. A staggered or delayed provision of the program to control group schools may also not be feasible. In these circumstances an age-cohort or selected cohorts design (Olweus, 2005) could be considered a viable alternative to a group-randomised trial (Farrington & Ttofi, 2009). This design, while not including a separate control group, allows for a smaller number of schools since the control or comparison data are collected in the study schools prior to program implementation. A limitation of the design is that there is no control for history or testing effects. Hence, it would be difficult to test for the presence of, for example response shift bias, in studies in which this design is applied. The possibility of these threats to validity need to be weighed against the benefits of this study design, e.g. reduced sample size, when designing a program evaluation.

7.4 Strengths and limitations of this doctoral research

This research has limitations. The Cyber Friendly Schools Project (CFSP) and the Supportive Schools Project (SSP) were conducted in non-government metropolitan Perth schools and the samples were skewed toward higher socio-economic families, the generalisability of the research findings based on these data beyond this context, is unknown. The data for the CFSP were largely collected using online surveys whereas those of the SSP and Australian Covert Bullying Prevalence Study (ACBPS) were obtained using hard copy surveys. The extent in each instance, of the impact the administration mode may have had on the results, is also unclear.

The data from the various studies included in this research were obtained using only self-report measures. However, the findings regarding, for example the presence of response shift bias and the bias resulting from parental consent procedures, are more widely applicable. What is of importance to researchers is an awareness of the possibility of these biases and the processes whereby their potential to bias the estimates of program effects are minimised.

The intraclass correlation (ICC) values presented in Chapter 5 are based on self-report and school-level clustering. These values based on self-report will likely underestimate the ICC's for bullying outcomes obtained through peer report and in particular, teacher report of

student involvement in bullying. They will also be underestimates of the ICC's operating in primary schools where clustering effects at the classroom level rather than the school level will be greatest.

Cyberbullying has not been specifically addressed in this thesis. An understanding of cyberbullying as bullying behaviour perpetrated through the use of technology, suggests the focus of anti-bullying programs should be on bullying behaviours, e.g. exclusion, rumour spreading to damage reputation etc., rather than the means by which these occur. A further justification for the focus on bullying rather than cyberbullying in school-based programs, is the much higher prevalence of bullying through offline than online means. Hence this doctoral research has addressed methodological issues related to the determination of effects from programs which address bullying behaviours in general, including but not only restricted to cyberbullying behaviours.

While a range of methodological issues that may threaten the validity of program evaluations were considered, it was beyond the scope of this doctoral research to investigate all threats to validity or sources of bias in such studies. The most obvious omission and of critical importance when evaluating programs, are issues related to the monitoring of program implementation in the intervention schools to assess not only the levels of implementation of the different components of a program, but also the integrity or fidelity of that implementation.

Notwithstanding these limitations, the broad range of methodological issues explored here with large data sets and data from group-randomised controlled trials enhances confidence in the findings and is a major strength of the work. So too are the advanced methods, particularly the statistical methods applied to address the research questions.

7.5 Recommendations and future directions

The following recommendations are made in light of the findings from this doctoral research as discussed in Section 7.3, and aim to strengthen the validity of evaluations of anti-bullying programs in schools and enhance the inference that may be drawn from such studies of program effects. These are followed by suggestions for future research activities arising from this doctoral research.

Recommendation 1: Use more than one data source to measure bullying behaviours, and choose instruments which have demonstrated validity and reliability and are sensitive to change.

Evaluators need to carefully consider the operational definitions most appropriate for the constructs of bullying victimisation and perpetration. The wording of the instrument(s) chosen for the evaluation and the context within which they are administered, e.g. whether a definition of bullying is provided or not, need to enhance the validity of the responses by for example, reflecting the three characteristics of bullying behaviours and the different forms of bullying. In this regard it is important to assess the demonstrated psychometric properties of the instrument and choose one that is most suited to the population to be studied, e.g. most appropriate for the age of the students and the cultural context.

Where possible, multiple data sources or informants should be used to avoid mono-method bias and enhance the robustness of the findings from the evaluation. Where self-report measures will be used, multi-item scales will offer greater sensitivity to change in bullying behaviours than categorical single item questions.

Recommendation 2: Test for response shift bias in measures of bullying outcomes resulting from program exposure.

Where single item global questions will be used as measures of bullying involvement, evaluators could consider also including single item retrospective pre-test questions in the post-program questionnaires. This will allow for the testing of response shift, particularly in subgroups of involved students who may be more susceptible to reconceptualisation, reprioritisation or recalibration of the bullying construct.

Where bullying involvement will be measured using multi-item scales, confirmatory factor analysis methods of assessing measurement invariance over time can be utilised to test for various types of response shift. Should measurement invariance be present in data, structural equation models can be utilised to test for true differences or program effects.

Where possible both approaches should be considered as they provide different information and insights into the data.

Recommendation 3: Consider the potential for bias in study findings resulting from parental consent procedures and put strategies in place to maximise parental response rates.

When active only parental consent is required prior to students completing surveys, evaluators need to firstly determine whether the potential for bias is such as to invalidate the study findings and hence render the research impracticable. Alternatively, programs may need to be evaluated in settings where less stringent consent procedures are required. Under these consent conditions, where possible, alternate sources of data from students participating and not participating in surveys should be explored and compared to assess the likely extent of any bias resulting from the active only parental consent requirement.

Where active only parental consent is not mandated and passive consent sufficient, researchers need to ensure the procedures for obtaining parental consent are such that parents' autonomy is maintained; multiple efforts are made using varied means to provide parents with the necessary information to make an informed decision; the information is in a format that parents can understand and engage with (including translation if necessary); parents have the opportunity to discuss the research with the researchers or an independent party; parents have multiple opportunities to indicate their non-consent and that the methods of returning forms indicating non-consent are as fail-safe as possible.

Whichever form of parental consent is required, researchers need to apply strategies to maximise response rates and to minimise selection bias.

Recommendation 4: Choose a strong study design for the anti-bullying program evaluation.

Group-randomised controlled study designs are preferred when resources allow for these designs. Pre-program data need to be collected to enable baseline comparisons of the study groups. Matching of schools prior to randomisation could be considered in circumstances were matching is likely to enhance the comparability of the groups. Age-cohort designs may also be considered if experimental studies are not feasible, program implementation cannot be delayed and testing and history effects are unlikely to severely bias the results.

Recommendation 5: Sample sufficient numbers of schools in the anti-bullying program evaluation.

Prior to the conduct of evaluations, formal sample size calculations based on the most appropriate intraclass correlation values available, are required to ensure the study has sufficient power to detect an effect at least moderate in size. Typically, many students from each school participate in a program evaluation, hence the numbers of schools rather than the numbers of students included in the study is critical. Where possible, the anti-bullying program should be implemented in 20 or more schools. In controlled studies, sufficient numbers of schools need to be allocated to each of the study groups to ensure their comparability and enhance the interpretation of the observed differences between the groups as indicative of the effects of the program.

Recommendation 6: Apply statistical methods appropriate to the distributions of bullying outcomes.

Statistical methods appropriate to the highly skewed distributions of the bullying outcomes, such as tobit regression or two-part growth models, should be applied in the data analyses to protect against invalid statistical inferences. In addition, estimation methods which account for the clustered nature of the data need to be used.

The following are suggested topics for future research endeavours. These are not only specific to the conduct of anti-bullying program evaluations and are thus, more widely applicable than the above recommendations.

Research Focus 1: The measurement of bullying.

There is much still to be learned regarding the measurement of bullying behaviours. In the first instance, the incorporation of the three characteristics of bullying, in particular the power imbalance, in instrument wording requires further investigation. For example, qualitative research with young people to increase researchers' understandings of their interpretations of survey items used to measure bullying victimisation and perpetration

would be helpful. Exploring the means of measuring the impact of the bullying on the victimised student, in addition to the frequency that it occurs, is also important. Insights that may be gained from such data could provide information regarding, for example, the forms of bullying or context within which the bullying occurs that are most distressing and harmful to young people. As young peoples' use of technology increases and their transitions between online and offline social interactions become more seamless, the extent to which young people incorporate online events when responding to a general bullying scale needs to be investigated. Furthermore, the measurement of cyberbullying or online bullying behaviours is still under considerable debate.

Research Focus 2: Exploration of the presence of response shift in a variety of contexts.

The presence of response shift was assessed in a particular context in this doctoral research, i.e. in non-government secondary schools and within the evaluation of a cyberbullying intervention trial. Further research on this phenomenon in other age groups and populations, especially those where the understanding of bullying may not be stable, and other anti-bullying programs, is needed to ascertain the extent to which this bias may be operating in evaluations of such programs.

Research Focus 3: Issues related to parental consent procedures and student completion of health surveys.

Research is needed to determine the consent status of parents who do not respond to approaches to obtain their consent, particularly within the context of seeking consent under passive or opt-out conditions. Further research into the biases resulting from opt-in parental consent procedures is required, not only in terms of the biased samples that may result, but also the bias that may be introduced into correlations or parameters of interest, and hence threaten the validity of findings from bullying-related studies. Of vital importance to inform the debate around this ethical issue, is research into the impact of the completion of health surveys on young people and hence the risk of harm, particularly on young people and children of different ages and for health outcomes of differing sensitivity or risk for harm.

Research Focus 4: Exploration of the biases resulting from applying inappropriate statistical methods to highly skewed data such as bullying outcomes.

The magnitude of the bias present in, for example, estimates of program effects and their standard errors, obtained when methods such as linear regression models are applied to the analysis of semi-continuous data, could be explored in a simulation or monte carlo study. Such research would be informative as to the impact on the validity of statistical inference drawn from analyses applying different statistical methods or models.

7.6 Contribution to the literature

This thesis has added to the body of knowledge on the methodology of evaluations of anti-bullying programs in schools. By investigating limitations and shortcomings of such evaluations identified in the literature, the findings have contributed to understandings of the particular issues of concern, such as the measurement of bullying, in the research community. Several of the issues, i.e. the means of testing for response shift bias; the impact of parental consent procedures in producing biased samples and biased estimates of correlations; the values for intraclass correlations for bullying outcomes required to determine adequate sample sizes for evaluations; and the appropriate statistical methods for modelling bullying outcomes, have not previously been investigated within the context of bullying-related research. It is hoped the findings and recommendations will aid in improving evaluations in the future.

7.7 Conclusions

Evaluations of bullying programs in schools can provide valid evidence of their efficacy – with careful planning and availability of sufficient resources. Apart from the usual methodological considerations around the conduct of an evaluation, the shortcomings of many evaluation studies can be addressed through the use of strong designs; of appropriate instruments to measure bullying behaviours and measures by which response shift can be assessed; suitable methods to determine sufficient sample sizes; strategies to maximise response and consent rates when recruiting subjects; and appropriate methods for the analyses of the data.

References

Achenbach, T. M. (1991). *Manual for the child behaviour checklist/4-18 and 1991 profile*. Burlington: University of Vermont, Department of Psychiatry.

Ahmed, S., Mayo, N. E., Corbiere, M., Wood-Dauphinee, S., Hanley, J., & Cohen, R. (2005). Change in quality of life of people with stroke over time: True change or response shift? *Quality of life research, 14*(3), 611-627.

Archer, J. (2004). Sex differences in aggression in real-world settings: A meta-analytic review. *Review of General Psychology, 8*(4), 291-322.

Arseneault, L., Bowes, L., & Shakoor, S. (2010). Bullying victimization in youths and mental health problems: 'Much ado about nothing'. *Psychological Medicine, 40*(5), 717–729. doi: 10.1017/S0033291709991383

Baker, J. R., Yardley, J. K., & McCaul, K. (2001). Characteristics of responding-, nonresponding- and refusing-parents in an adolescent lifestyle choice study. *Evaluation Review, 25*(6), 605-618.

Baldry, A. C., & Farrington, D. P. (2007). Effectiveness of programs to prevent school bullying. *Victims and Offenders, 2*(2), 183-204.

Barclay-Goddard, R., Epstein, J. D., & Mayo, N. E. (2009). Response shift: A brief overview and proposed research priorities. *Quality of Life Research, 18*(3), 335-346.

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological methods, 16*(3), 265-284.

Bond, L., Patton, G., Glover, S., Carlin, J. B., Butler, H., Thomas, L., & Bowes, G. (2004). The Gatehouse Project: can a multilevel school intervention affect emotional wellbeing and health risk behaviours? *Journal of Epidemiology and Community Health, 58*(12), 997-1003.

Bond, L., Wolfe, S., Tollit, M., Butler, H., & Patton, G. (2007). A comparison of the Gatehouse Bullying Scale and the Peer Relations Questionnaire for students in secondary school. *Journal of school health, 77*(2), 75-79.

Bovaird, J. A. (2009). Scales and surveys. Some problems with measuring bullying behavior. In S. R. Jimerson, S. M. Swearer & D. L. Espelage (Eds.), *Handbook of bullying in schools: An international perspective* (pp. 277-292). New York: Routledge, Taylor & Francis.

Bradshaw, C. P., Sawyer, A. L., & O'Brennan, L. M. (2009). A social disorganization perspective on bullying-related attitudes and behaviours: the influence of school context. *Am J Community Psychology, 43*, 204-220.

Breivik, K., & Olweus, D. (Under review). An item response theory analysis of the Olweus Bullying Scale. *Aggressive Behavior*.

Bronfenbrenner, U. (1995). Developmental ecology through space and time: A future perspective. In P. Moen, G. H. Elder & K. Luscher (Eds.), *Examining lives in context: Perspectives on the ecology of human development* (pp. 619-647). Washington, DC: American Psychological Association.

Brown, E. C., Catalano, R. F., Fleming, C. B., Haggerty, K. P., & Abbott, R. D. (2005). Adolescent substance use outcomes in the Raising Healthy Children project: A two-part latent growth curve analysis. *Journal of consulting and clinical psychology, 73*(4), 699-710.

Brown, E. C., Low, S., Smith, B. H., & Haggerty, K. P. (2011). Outcomes from a school-randomized controlled trial of Steps to Respect: A bullying prevention program. *School Psychology Review, 40*(3), 423-433.

Brown, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 136-162). Newbury Park: Sage.

Campbell, M. A. (2005). Cyber bullying: An old problem in a new guise? *Australian journal of Guidance and Counselling, 15*(1), 68-76.

Campbell, M. K., Piaggio, G., Elbourne, D. R., & Altman, D. G. (2012). Consort 2010 statement: extension to cluster randomised trials. *BMJ: British Medical Journal, 345*, e5661. doi: 10.1136/bmj.e5661

Card, N. A., & Hodges, E. V. E. (2008). Peer victimization among schoolchildren: Correlations, causes, consequences, and considerations in assessment and intervention. *School Psychology Quarterly; School Psychology Quarterly, 23*(4), 451-461.

Carlin, J. B., & Hocking, J. (1999). Design of cross-sectional surveys using cluster sampling: an overview with Australian case studies. *Aust & NZ J of Public Health, 23*(5), 546-551.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Child Health Promotion Research Centre. (2010). An empirical intervention to reduce cyber bullying in adolescents. Annual report to Healthway. Perth, Western Australia: Edith Cowan University, Child Health Promotion Research Centre.

CHPRC. (2010). *An empirical intervention to reduce cyber bullying in adolescents. Annual report to Healthway.* Perth, Western Australia: Child Health Promotion Research Centre, Edith Cowan University.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2 ed.). Hillsdale, NJ: Erlbaum.

Cook, C. R., Williams, K. R., Guerra, N. G., & Kim, T. (2009). Variability in the prevalence of bullying and victimization. In S. R. Jimerson, S. M. Swearer & D. L. Espelage (Eds.), *Handbook of bullying in schools: An international perspective* (pp. 347-362). New York: Routledge, Taylor & Francis.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Chicago: Rand McNally College Publishing Company.

Cornell, D. G., & Bandyopadhyay, S. (2010). The assessment of bullying. In S. R. Jimerson, S. M. Swearer & D. L. Espelage (Eds.), *Handbook of bullying in schools: An international perspective* (pp. 265-276). New York: Routledge, Taylor & Francis.

Courser, M. W., Shamblen, S. R., Lavrakas, P. J., Collins, D., & Ditterline, P. (2009). The impact of active consent procedures on nonresponse and nonresponse error in youth survey data. Evidence from a new experiment. *Evaluation Review, 33*(4), 370-395.

Craig, W., Harel-Fisch, Y., Fogel-Grinvald, H., Dostaler, S., Hetland, J., Simons-Morton, B., . . . Due, P. (2009). A cross-national profile of bullying and victimization among adolescents in 40 countries. *International Journal of Public Health, 54*(2), 216-224. doi: 10.1007/s00038-009-5413-9

Crick, N. R., & Grotpeter, J. K. (1995). Relational aggression, gender, and social-psychological adjustment. *Child Development, 66*(3), 710-722.

Cronbach, L. J., & Furby, L. (1970). How we should measure" change" – or should we? *Psychological Bulletin, 74*(1), 68-80.

Cross, D., Brown, D., Epstein, M., & Read, M. (2009). *Strengthening school and families' capacity to reduce the academic, social, and emotional harms secondary students' experience from cyber bullying*. Perth, Western Australia: Child Health Promotion Research Centre, Edith Cowan University.

Cross, D., Epstein, M., Hearn, L., Slee, P., Shaw, T., & Monks, H. (2011). National Safe Schools Framework: Policy and practice to reduce bullying in Australian schools. *International Journal of Behavioral Development, 35*(5), 398-404.

Cross, D., Monks, H., Hall, M., Shaw, T., Pintabona, Y., Erceg, E., . . . Lester, L. (2011). Three-year results of the Friendly Schools whole-of-school intervention on children's bullying behaviour. *British Educational Research Journal, 37*(1), 105-129.

Cross, D., Shaw, T., Hearn, L., Epstein, M., Monks, H., Lester, L., & Thomas, L. (2009). Australian Covert Bullying Prevalence Study (ACBPS) Retrieved from http://www.deewr.gov.au/Schooling/NationalSafeSchools/Pages/research.aspx

deLara, E. W. (2012). Why adolescents don't disclose incidents of bullying and harassment. *Journal of School Violence, 11*(4), 288-305.

Delva, J., Grogan-Kaylor, A., Steinhoff, E., Shin, D. E., & Siefert, K. (2007). Using tobit regression analysis to further understand the association of youth alcohol problems with depression and parental factors among Korean adolescent females. *Journal of Preventive Medicine and Public Health, 40*(2), 145-149.

Dent, C. W., Galaif, J., Sussman, S., Stacy, A., Burtun, D., & Flay, B. R. (1993). Demographic, psychosocial and behavioral differences in samples of actively and passively consented adolescents. *Addictive Behaviors, 18*(1), 51-56.

Department of Education. Western Australia. (2009). *Research conducted on Department of Education sites by external parties*. Perth: The Government of Western Australia. Retrieved from http://www.det.wa.edu.au/policies/detcms/portal/.

Dishion, T. J., & Owen, L. D. (2002). A longitudinal analysis of friendships and substance use: Bidirectional influence from adolescence to adulthood. *Developmental Psychology, 38*, 480-491.

Donner, A., & Klar, N. (2004). Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health, 94*(3), 416-422.

Dooley, J. J., Pyżalski, J., & Cross, D. (2009). Cyberbullying versus face-to-face bullying. *Zeitschrift für Psychologie/Journal of Psychology, 217*(4), 182-188.

Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods, 14*(2), 370-388.

Ellickson, P. L., & Hawes, J. A. (1989). An assessment of active versus passive methods for obtaining parental consent. *Evaluation Review, 13*(1), 45-55.

Ellwood, P., Asher, M. I., Stewart, A. W., & ISAAC Phase III Study Group. (2010). The impact of the method of consent on response rates in the ISAAC time trends study. *The International Journal of Tuberculosis and Lung Disease, 14*(8), 1059-1065.

Esbensen, F. A., Melde, C., Taylor, T. J., & Peterson, D. (2008). Active parental consent in school-based research. How much is enough and how do we get it? *Evaluation Review, 32*(4), 335-362.

Espelage, D. L., & Holt, M. K. (2001). Bullying and victimization during early adolescence: Peer influences and psychosocial correlates. *Journal of Emotional Abuse, 2*(2/3), 123-142.

Espelage, D. L., & Swearer, S. M. (2003). Research on school bullying and victimization: What have we learned and where do we go from here? *School Psychology Review, 32*(3), 365-383.

Farrington, D. P. (2003). Methodological quality standards for evaluation research. *The Annals of the American Academy of Political and Social Science, 587*(1), 49-68.

Farrington, D. P., & Ttofi, M. M. (2007). School-based programs to reduce bullying and victimization. *KiVa*, 6.

Farrington, D. P., & Ttofi, M. M. (2009a). Reducing school bullying: Evidence-based implications for policy. *Crime and Justice, 38*(1), 281-345.

Farrington, D. P., & Ttofi, M. M. (2009b). School-based programs to reduce bullying and victimization. *Campbell Systematic Reviews, 6*(10.4073).

Felix, E. D., Sharkey, J. D., Green, J. G., Furlong, M. J., & Tanigawa, D. (2011). Getting precise and pragmatic about the assessment of bullying: The development of the California Bullying Victimization Scale. *Aggressive Behavior, 37*, 234-247.

Ferguson, C. J., San Miguel, C., Kilburn, J. C., & Sanchez, P. (2007). The effectiveness of school-based anti-bullying programs A meta-analytic review. *Criminal Justice Review, 32*(4), 401-414.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269-314). Greenwich, Connecticut: Information Age Publishing.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., . . . Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science, 6*(3), 151-175.

Fonagy, P., Twemlow, S. W., Vernberg, E. M., Nelson, J. M., Dill, E. J., Little, T. D., & Sargent, J. A. (2009). A cluster randomized controlled trial of child focused psychiatric consultation and a school systems focused intervention to reduce aggression. *Journal of Child Psychology and Psychiatry, 50*(5), 607-616.

Freedman, D. A. (2006). On the so-called “Huber sandwich estimator” and “robust standard errors”. *The American Statistician, 60*(4), 299-302.

Frissell, K. C., McCarthy, D. M., D'Amico, E. J., Metrik, J., Ellingstad, T. P., & Brown, S. A. (2004). Impact of consent procedures on reported levels of adolescent alcohol use. *Psychology of addictive behaviors, 18*(4), 307-315.

Furlong, M. J., Sharkey, J. D., Felix, E., Tanigawa, D., & Greif-Green, J. (2010). Bullying assessment: A call for increased precision of self-reporting procedures. In S. R. Jimerson, S. Swearer & D. L. Espelage (Eds.), *Handbook of bullying in schools: An international perspective* (pp. 329-346). New York: Routledge, Taylor & Francis.

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *The Journal of Applied Behavioral Science, 12*(2), 133-157.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry, 38*(5), 581-586.

Greif, J. L., & Furlong, M. J. (2006). The assessment of school bullying. *Journal of School Violence, 5*(3), 33-50.

Griffin, R. S., & Gross, A. M. (2004). Childhood bullying: Current empirical findings and future directions for research. *Aggression and Violent Behavior, 9*(4), 379-400.

Hahn Fox, B., Farrington, D. P., & Ttofi, M. M. (2012). Successful bullying prevention programs: Influence of research design, implementation features, and program components. *International Journal of Conflict and Violence, 6*(2), 273-282.

Hall, M., Cordin, T., Bruce, K., & Paki, D. (2011). *Strengthening pastoral care to reduce secondary students' harm from tobacco. Final report to Healthway.* Perth, Western Australia: Child Health Promotion Research Centre, Edith Cowan University.

Hamburger, M. E., Basile, K. C., & Vivolo, A. M. (2011). *Measuring bullying victimization, perpetration, and bystander experiences: A compendium of assessment tools*: Centers for Disease Control and Prevention, National Center for Injury Prevention and Control, Division of Violence Prevention.

Hawker, D. S. J., & Boulton, M. J. (2000). Twenty years' research on peer victimization and psychosocial maladjustment: A meta-analytic review of cross-sectional studies. *Journal of Child Psychology and Psychiatry, 41*(4), 441-455.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied Survey Data Analysis*. Boca Raton, Florida: CRC Press.

Heller, S. S., Boothe, A., Keyes, A., Nagle, G., Sidell, M., & Rice, J. (2011). Implementation of a mental health consultation model and its impact on early childhood teachers' efficacy and competence. *Infant Mental Health Journal, 32*(2), 143-164.

Henry, K. L., Smith, E. A., & Hopkins, A. M. (2002). The effect of active parental consent on the ability to generalize the results of an alcohol, tobacco, and other drug prevention trial to rural adolescents. *Evaluation Review, 26*(6), 645-655.

Hill, L. G., & Betz, D. L. (2005). Revisiting the retrospective pretest. *American Journal of Evaluation, 26*(4), 501-517.

Howard, G. S. (1980). Response-shift bias. A problem in evaluating interventions with pre/post self-reports. *Evaluation Review, 4*(1), 93-106.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological methods, 3*(4), 424-453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Hunt, C., Peters, L., & Rapee, R. M. (2012). Development of a measure of the experience of being bullied in youth. *Psychological Assessment, 24*(1), 156-165.

Hutchison, D. (2004). *Clustering effects in TIMSS.* Paper presented at the International Research Conference (IRC-2004), University of Cyprus, Lefkosia.

Hymel, S., Wagner, E., & Butler, L. J. (1990). Reputational bias: View from the peer group. In S. R. Asher & J. D. Coie (Eds.), *Peer rejection in childhood* (pp. 156). Cambridge: Cambridge University Press.

Ji, P. Y., Pokorny, S. B., & Jason, L. A. (2004). Factors influencing middle and high schools' active parental consent return rates. *Evaluation Review, 28*(6), 578-591.

Jimerson, S. R., & Huai, N. (2010). International Perspectives on Bullying Prevention and Intervention. In S. Jimerson, S. Swearer & D. Espelage (Eds.), *Handbook of Bullying in Schools: An International Perspective* (pp. 571-592). New York: Routledge.

Joreskog, K. G., & Sorbom, D. (1993). *LISREL 8: Structural equation modelling with the SIMPLIS command language*. Chicago: Scientific Software International.

Junghans, C., Feder, G., Hemingway, H., Timmis, A., & Jones, M. (2005). Recruiting patients to medical research: Double blind randomised trial of "opt-in" versus "opt-out" strategies. *Bmj, 331*(7522), 940. doi: 10.1136/bmj.38583.625613.AE

Juvonen, J., Graham, S., & Schuster, M. A. (2003). Bullying among young adolescents: The strong, the weak, and the troubled. *Pediatrics, 112*(6), 1231-1237.

Juvonen, J., Nishina, A., & Graham, S. (2001). Self-views versus peer perceptions of victim status among early adolescents. In J. Juvonen & S. Graham (Eds.), *Peer harassment in school: A plight of the vulnerable and the victimized* (pp. 105-124). New York: Guilford.

Kaltiala-Heino, R., Rimpelä, M., Marttunen, M., Rimpelä, A., & Rantanen, P. (1999). Bullying, depression, and suicidal ideation in Finnish adolescents: school survey. *BMj, 319*, 348-351.

Kaltiala-Heino, R., Rimpela, M., Rantanen, P., & Rimpela, A. (2000). Bullying at school - an indicator of adolescents at risk for mental disorders. *Journal of adolescence, 23*(6), 661-674.

Kaltiala-Heino, R., Rimpelä, M., Rantanen, P., & Rimpelä, A. (2000). Bullying at school - an indicator of adolescents at risk for mental disorders. *Journal of adolescence, 23*(6), 661-674.

Kärnä, A., Voeten, M., Little, T. D., Poskiparta, E., Kaljonen, A., & Salmivalli, C. (2011a). A large-scale evaluation of the KiVa Antibullying Program: Grades 4–6. *Child Development, 82*(1), 311-330.

Kärnä, A., Voeten, M., Little, T. D., Poskiparta, E., Kaljonen, A., & Salmivalli, C. (2011b). A Large Scale Evaluation of the KiVa Antibullying Program: Grades 4–6. *Child development, 82*(1), 311-330.

Kendrick, K., Jutengren, G., & Stattin, H. (2012). The protective role of supportive friends against bullying perpetration and victimization. *Journal of adolescence, 35*(4), 1069-1080.

Kert, A. S., Codding, R. S., Tryon, G. S., & Shiyko, M. (2010). Impact of the word "bully" on the reported rate of bullying behavior. *Psychology in the Schools, 47*(2), 193-204.

Kiesner, J., Dishion, T. J., & Poulin, F. (2001). A reinforcement model of conduct problems in children and adolescents: Advances in theory and intervention. In J. Hill & B. Maughan (Eds.), *Conduct disorders in childhood and adolescence* (pp. 264-291). New York: Cambridge University Press.

Kim, Y. K., & Muthén, B. O. (2009). Two-part factor mixture modeling: Application to an aggressive behavior measurement instrument. *Structural Equation Modeling, 16*(4), 602-624.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology, 5*(3), 213-236.

Kumpfer, K. L., Xie, J., & O'Driscoll, R. (2012). Effectiveness of a culturally adapted Strengthening Families Program 12–16 Years for high-risk Irish families. *Child Youth Care Forum, 41*, 173-195.

Kutash, K., Banks, S., Duchnowski, A., & Lynn, N. (2007). Implications of nested designs in school-based mental health services research. *Evaluation and Program Planning, 30*(2), 161-171.

Lacy, K., Kremer, P., Silva-Sanigorski, A., Allender, S., Leslie, E., Jones, L., . . . Swinburn, B. (2012). The appropriateness of opt-out consent for monitoring childhood obesity in Australia. *Pediatric Obesity, 7*, e62-e67.

Ladd, G. W., Kochenderfer, B. J., & Coleman, C. C. (1996). Friendship quality as a predictor of young children's early school adjustment. *Child development, 67*(3), 1103-1118.

Lam, T. C. M., & Bengo, P. (2003). A comparison of three retrospective self-reporting methods of measuring change in instructional practice. *American Journal of Evaluation, 24*(1), 65-80.

Langhinrichsen-Rohling, J., Arata, C., O'Brien, N., Bowers, D., & Klibert, J. (2006). Sensitive research with adolescents: Just how upsetting are self-report surveys anyway? *Violence and Victims, 21*(4), 425-444.

Laslett, A.-M., Ferris, J., Dietze, P., & Room, R. (2012). Social demography of alcohol-related harm to children in Australia. *Addiction, ePub*(ePub), ePub-ePub.

Leakey, T., Lunde, K. B., Koga, K., & Glanz, K. (2004). Written parental consent and the use of incentives in a youth smoking prevention trial: A case study from project SPLASH. *American Journal of Evaluation, 25*(4), 509-523.

Lee, C., Mun, E. Y., White, H. R., & Simon, P. (2010). Substance use trajectories of black and white young men from adolescence to emerging adulthood: A two-part growth curve analysis. *Journal of Ethnicity in Substance Abuse, 9*(4), 301-319.

Limber, S. P., & Small, M. A. (2003). State laws and policies to address bullying in schools. *School Psychology Review, 32*(3), 445-455.

London, K., & Wright, D. B. (2012). Analyzing Change between Two or More Groups. Analysis of Variance versus Analysis of Covariance. In B. Laursen, T. D. Little & N. A. Card (Eds.), *Handbook of Developmental Research Methods*. New York: The Guilford Press.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables* (Vol. 7). Thousand Oaks, California: Sage Publications, Inc.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*(5), 304-305.

Lovibond, P., & Lovibond, S. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. *Behaviour research and therapy, 33*(3), 335-343.

Lovibond, S., & Lovibond, P. (1995). *Manual for the Depression Anxiety Stress Scales.* Sydney: Psychology Foundation.

McClendon, M. K. J. (1994). *Multiple Regression and Causal Analysis*. Itasca, Illionois: FE Peacock Publishers.

McPherson, S., & Barbosa-Leiker, C. (2012). An example of a two-part latent growth curve model for semicontinuous outcomes in the health sciences. *Journal of Applied Statistics, 39*(10), 2113-2128.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568-592.

Mellor, J. M., Rapoport, R. B., & Maliniak, D. (2008). The impact of child obesity on active parental consent in school-based survey research on healthy eating and physical activity. *Evaluation Review, 32*(3), 298-312.

Menesini, E. (2012). Cyberbullying: The right value of the phenomenon. Comments on the paper "Cyberbullying: An overrated phenomenon?". *European Journal of Developmental Psychology, 9*(5), 544-552.

Merrell, K. W., Gueldner, B. A., Ross, S. W., & Isava, D. M. (2008). How effective are school bullying intervention programs? A meta-analysis of intervention research. *School Psychology Quarterly, 23*(1), 26-42.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*(3), 479-515.

Monks, C. P., & Smith, P. K. (2006). Definitions of bullying: Age differences in understanding of the term, and the role of experience. *British Journal of Developmental Psychology, 24*(4), 801-821.

MPlus. (2012). http://www.statmodel.com/chidiff.shtml Retrieved January 2012

Murray, D. M. (1998). *Design and Analysis of Group-Randomized Trials*. New York: Oxford University Press.

Murray, D. M., & Blitstein, J. L. (2003). Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review, 27*(1), 79-103.

Murray, D. M., Clark, M. H., & Wagenaar, A. C. (2000). Intraclass correlations from a community-based alcohol prevention study: the effect of repeat observations on the same communities. *Journal of studies on alcohol, 61*(6), 881-890.

Murray, D. M., Phillips, G. A., Birnbaum, A. S., & Lytle, L. A. (2001). Intraclass correlation for measures from a middle school nutrition intervention study: estimates, correlates, and applications. *Health Education & Behavior, 28*(6), 666-679.

Murray, D. M., Rooney, B. L., Hannan, P. J., Peterson, A. V., Ary, D. V., Biglan, A., . . . Futterman, R. (1994). Intraclass correlation among common measures of adolescent smoking: estimates, correlates, and applications in smoking prevention studies. *American journal of epidemiology, 140*(11), 1038-1050.

Murray, D. M., & Short, B. J. (1997). Intraclass correlation among measures related to tobacco use by adolescents: estimates, correlates, and applications in intervention studies. *Addictive Behaviors, 22*(1), 1-12.

Murray, D. M., Stevens, J., Hannan, P. J., Catellier, D. J., Schmitz, K. H., Dowda, M., . . . Yang, S. (2006). School-level intraclass correlation for physical activity in sixth grade girls. *Medicine and Science in Sports and Exercise, 38*(5), 926-936.

Murray, D. M., Varnell, S. P., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: a review of recent methodological developments. *American Journal of Public Health, 94*(3), 423-432.

Muthén, B. (2006, 13 April 2006 - 11.03am). Mplus Discussion: Structural Equation Modelling. Retrieved from http://www.statmodel.com/discussion/messages/11/80.html#POST8182

Muthén, B., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of Advanced Multilevel Analysis* (pp. 15-40). New York: Taylor & Francis Group.

Muthen, L. K., & Muthen, B. O. (1998-2009). *Mplus User's Guide. 5th Edition*: Muthén & Muthén.

Mynard, H., & Joseph, S. (2000). Development of the multidimensional peer-victimization scale. *Aggressive Behavior, 26*(2), 169-178.

Nansel, T. R., Overpeck, M., Pilla, R. S., Ruan, W., Simons-Morton, B., & Scheidt, P. (2001). Bullying behaviors among US youth. *The Journal of the American Medical Association, 285*(16), 2094-2100.

National Health and Medical Research Council, Australian Research Council, & Australian Vice-Chancellors' Committee. (2007). *National Statement on Ethical Conduct in Human Research*. Canberra: Australian Government. Retrieved from http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72.pdf.

Nixon, C. L., & Werner, N. E. (2010). Reducing adolescents' involvement with relational aggression: Evaluating the effectiveness of the Creating A Safe School (CASS) intervention. *Psychology in the Schools, 47*(6), 606-620.

Nolte, S., Elsworth, G. R., Sinclair, A. J., & Osborne, R. H. (2012). The inclusion of 'then-test'questions in post-test questionnaires alters post-test responses: A randomized study of bias in health program evaluation. *Quality of life research, 21*(3), 487-494.

O'Brennan, L. M., Bradshaw, C. P., & Sawyer, A. L. (2009). Examining developmental differences in the social-emotional problems among frequent bullies, victims, and bully/victims. *Psychology in the Schools, 46*(2), 100-115.

OECD. (2010). *PISA 2009 Technical Report (Preliminary version)*. OECD Publishing. Paris.

Olsen, M. K., & Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association, 96*(454), 730-745.

Olweus, D. (1996). *The Revised Olweus Bully/Victim Questionnaire*. Bergen, Norway: Research Centre for Health Promotion, University of Bergen.

Olweus, D. (2005). A useful evaluation design, and effects of the Olweus Bullying Prevention Program. *Psychology, Crime & Law, 11*(4), 389-402.

Olweus, D. (2012). Cyberbullying: An overrated phenomenon? *European Journal of Developmental Psychology, 9*(5), 520-538.

Olweus, D., & Limber, S. P. (2010). Bullying in school: Evaluation and dissemination of the Olweus Bullying Prevention Program. *American Journal of Orthopsychiatry, 80*(1), 124-134.

Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research, 14*(3), 587-598.

Orpinas, P., Kelder, S., Frankowski, R., Murray, N., Zhang, Q., & McAlister, A. (2000). Outcome evaluation of a multi-component violence-prevention program for middle schools: The Students for Peace project. *Health Education Research, 15*(1), 45-58.

Ortega, R., Mora-Merchán, J. A., Singer, M., Smith, P. K., Pereira, B., & Menesini, E. (2001a). *Final Report of the Working Group on General Survey Questionnaires and Nomination Methods Concerning Bullying*. Seville: TMR Project: Nature and Prevention of Bullying.

Ortega, R., Mora-Merchán, J. A., Singer, M., Smith, P. K., Pereira, B., & Menesini, E. (2001b). Final report of the working group on general survey questionnaires and nomination methods concerning bullying. Seville: TMR Project - The causes and nature of bullying and social exclusion in schools, and ways of preventing them.

Osgood, D. W., Finken, L. L., & McMorris, B. J. (2002). Analyzing multiple-item measures of crime and deviance II: Tobit regression analysis of transformed scores. *Journal of Quantitative Criminology, 18*(4), 319-347.

Österman, K., Björkqvist, K., Lagerspetz, K. M. J., Kaukiainen, A., Huesmann, L. R., & Fraczek, A. (1994). Peer and self estimated aggression and victimization in 8 year old children from five ethnic groups. *Aggressive Behavior, 20*(6), 411-428.

Pearce, N., Cross, D., Monks, H., Waters, S., & Falconer, S. (2011). Current evidence of best practice in whole-school bullying intervention and its potential to inform cyberbullying interventions. *Australian journal of Guidance and Counselling, 21*(1), 1-21.

Pellegrini, A. D., & Bartini, M. (2000). A longitudinal study of bullying, victimization, and peer affiliation during the transition from primary school to middle school. *American Educational Research Journal, 37*(3), 699-725.

Perren, S., Dooley, J., Shaw, T., & Cross, D. (2010). Bullying in school and cyberspace: Associations with depressive symptoms in Swiss and Australian adolescents. *Child and Adolescent Psychiatry and Mental Health, 4*(28), 1-10.

Portney, L. G., & Watkins, M. P. (2000). *Foundations of Clinical Research. Applications to Practice* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Priest, N., Duncan, R., Yap, M. B. H., Redmond, G., Anderson, A., & Wade, C. (2012). Active versus passive parental consent for improving participant recruitment and outcomes in studies targeting children (Protocol). *The Cochrane Library*.

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. College Station, TX: StataCorp.

Renda, J., Vassallo, S., & Edwards, B. (2011). Bullying in early adolescence and its association with anti-social behaviour, criminality and violence 6 and 10 years later. *Criminal Behaviour and Mental Health, 21*(2), 117-127.

Resnick, M. D., Bearman, P. S., Blum, R. W., Bauman, K. E., Harris, K. M., Jones, J., . . . Shew, M. (1997). Protecting adolescents from harm. Findings from the National Longitudinal Study on Adolescent Health. *The Journal of the American Medical Association, 278*(10), 823-832.

Resnicow, K., Zhang, N., Vaughan, R. D., Reddy, S. P., James, S., & Murray, D. M. (2010). When intraclass correlation coefficients go awry: a case study from a school-based smoking prevention study in South Africa. *American Journal of Public Health, 100*(9), 1714-1718.

Reynolds, W. M. (2003). *Bully victimization: Reynolds Scale for Schools*. San Antonio, TX: Psychological Corporation.

Rigby, K. (1998a). *Manual for the Peer Relations Questionnaire*. Point Lonsdale, Victoria (Aust): The Professional Reading Guide for Educational Administrators.

Rigby, K. (1998b). *The Peer Relations Questionnaire*. Point Lonsdale, Victoria (Aust): The Professional Reading Guide for Educational Administrators.

Rigby, K. (2002). *New perspectives on bullying*. London: Jessica Kingsley Publishers.

Rigby, K. (2007). *Bullying in schools and what to do about it (Revised edition)*. Victoria: ACER Press.

Rodriguez, G., & Elo, I. (2003). Intra-class correlation in random-effects models for binary data. *The Stata Journal, 3*(1), 32-46.

Roland, E. (1993). Bullying: A developing tradition of research and management. In D. P. Tattum (Ed.), *Understanding and managing bullying* (pp. 15-30). Oxford: Heinemann Educational.

Room, R., & Rehm, J. (2011). Clear criteria based on absolute risk: Reforming the basis of guidelines on low-risk drinking. *Drug and Alcohol Review, ePub*(ePub), ePub-ePub.

Rothon, C., Head, J., Klineberg, E., & Stansfeld, S. (2011). Can social support protect bullied adolescents from adverse outcomes? A prospective study on the effects of bullying on the educational achievement and mental health of adolescents at secondary schools in East London. *Journal of Adolescence, 34*(3), 579-588.

Runions, K., & Shaw, T. Teacher-child relationship, child aggression and withdrawal in the development of peer victimization. *Journal of Applied Developmental Psychology, (Under review)*.

Runions, K., Vitaro, F., Shaw, T., Cross, D., Hall, M., & Boivin, M. Teacher-child relationship, parenting, and growth in likelihood and severity of physical aggression in the early school years. *Merrill-Palmer Quarterly, (In press)*.

Ryan, W., & Smith, J. D. (2009). Antibullying programs in schools: How effective are evaluation practices? *Prevention Science, 10*(3), 248-259.

Salmivalli, C., Kärnä, A., & Poskiparta, E. (2011). Counteracting bullying in Finland: The KiVa program and its effects on different forms of being bullied. *International Journal of Behavioral Development, 35*(5), 405-411.

Scheier, L. M., Griffin, K. W., Doyle, M. M., & Botvin, G. J. (2002). Estimates of intragroup dependence for drug use and skill measures in school-based drug abuse prevention trials: an empirical study of three independent samples. *Health Education & Behavior, 29*(1), 85-103.

Schwartz, C. E., Bode, R., Repucci, N., Becker, J., Sprangers, M. A. G., & Fayers, P. M. (2006). The clinical significance of adaptation to changing health: A meta-analysis of response shift. *Quality of Life Research, 15*(9), 1533-1550.

Schwartz, C. E., Sprangers, M., & Fayers, P. (2005). Response shift: You know it's there but how do you capture it? Challenges for the next phase of research. In P. Fayers & R. Hay (Eds.), *Assessing quality of life in clinical trials. Methods and Practice* (pp. 275-290). Oxford: Oxford University Press.

Schwartz, C. E., & Sprangers, M. A. G. (1999). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social science & medicine, 48*(11), 1531-1548.

Schwartz, C. E., & Sprangers, M. A. G. (2010). Guidelines for improving the stringency of response shift research using the thentest. *Quality of life research, 19*(4), 455-464.

Secor-Turner, M., Sieving, R., Widome, R., Plowman, S., & Vanden Berk, E. (2010). Active parent consent for health surveys with urban middle school students: Processes and outcomes*. *Journal of School Health, 80*(2), 73-79.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company.

Shaw, T., & Cross, D. (2012). The clustering of bullying and cyberbullying behaviour within Australian schools. *Australian Journal of Education, 56*(2), 142-162.

Shaw, T., Dooley, J. D., Cross, D., Zubrick, S. R., & Waters, S. (In press). The Forms of Bullying Scale (FBS): Validity and reliability estimates for a measure of bullying victimization and perpetration in early adolescence. *Psychological Assessment*.

Sibthorp, J., Paisley, K., Gookin, J., & Ward, P. (2007). Addressing response-shift bias: Retrospective pretests in recreation research and evaluation. *Journal of leisure research, 39*(2), 295-315.

Siddiqui, O., Hedeker, D., Flay, B. R., & Hu, F. B. (1996). Intraclass correlation estimates in a school-based smoking prevention study. *American journal of epidemiology, 144*(4), 425-433.

Slonje, R., & Smith, P. K. (2008). Cyberbullying: Another main type of bullying? *Scandinavian Journal of Psychology, 49*(2), 147-154.

Slonje, R., Smith, P. K., & Frisén, A. (2013). The nature of cyberbullying, and strategies for prevention. *Computers in Human Behavior, 29*, 26-32.

Smith, J. D., Schneider, B. H., Smith, P. K., & Ananiadou, K. (2004). The effectiveness of whole-school antibullying programs: A synthesis of evaluation research. *School Psychology Review, 33*(4), 547-560.

Smith, P. K. (2011). Why interventions to reduce bullying and violence in schools may (or may not) succeed: Comments on this Special Section. *International Journal of Behavioral Development, 35*(5), 419-423.

Smith, P. K. (2012). Cyberbullying: Challenges and opportunities for a research program - A response to Olweus (2012). *European Journal of Developmental Psychology, 9*(5), 553-558.

Smith, P. K., Ananiadou, K., & Cowie, H. (2003). Interventions to reduce school bullying. *Canadian Journal of Psychiatry, 48*(9), 591-599.

Smith, P. K., Cowie, H., Olafsson, R. F., & Liefooghe, A. P. D. (2002). Definitions of bullying: A comparison of terms used, and age and gender differences, in a fourteen-country international comparison. *Child Development, 73*(4), 1119-1133.

Smith, P. K., del Barrio, C., & Tokunaga, R. (In press). Definitions of bullying and cyberbullying: How useful are the terms? In S. Bauman, D. Cross & J. Walker (Eds.), *Principles of Cyberbullying Research: Definition, Methods and Measures*. New York: Routledge, Taylor & Francis.

Smith, P. K., Madsen, K. C., & Moody, J. C. (1999). What causes the age decline in reports of being bullied at school? Towards a developmental analysis of risks of being bullied. *Educational Research, 41*(3), 267-285.

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry, 49*(4), 376-385.

Smith, P. K., Salmivalli, C., & Cowie, H. (2012). Effectiveness of school-based programs to reduce bullying: a commentary. *Journal of Experimental Criminology, 8*(4), 433-441.

Smith, P. K., & Slonje, R. (2009). Cyberbullying: The nature and extent of a new kind of bullying, in and out of school. In S. Jimerson, S. Swearer & D. Espelage (Eds.), *Handbook of bullying in schools: An international perspective*. Hoboken: Routledge.

Smolkowski, K., Biglan, A., Dent, C., & Seeley, J. (2006). The multilevel structure of four adolescent problems. *Prevention Science, 7*(3), 239-256.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: SAGE.

Snijders, T. A. B. A. B., & Bosker, R. J. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*: Sage Publications Limited.

Solberg, M. E., & Olweus, D. (2003). Prevalence estimation of school bullying with the Olweus Bully/Victim Questionnaire. *Aggressive Behavior, 29*(3), 239-268.

Sourander, A., Klomek, A. B., Kumpulainen, K., Puustjarvi, A., Elonheimo, H., Ristkari, T., . . . Ronning, J. A. (2011). Bullying at age eight and criminality in adulthood: Findings from the Finnish Nationwide 1981 Birth Cohort Study. *Soc Psychiatry Psychiatr Epidemiol, 46*, 1211-1219.

Spears, B., Slee, P., Owens, L., Johnson, B., & Campbell, A. (2008). Behind the Scenes: Insights into the Human Dimension of Covert Bullying. Adelaide: University of South Australia.

Spiel, C., & Strohmeier, D. (2011). National strategy for violence prevention in the Austrian public school system: Development and implementation. *International Journal of Behavioral Development, 35*(5), 412-418.

Sprangers, M. A. G., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: A theoretical model. *Social Science & Medicine, 48*(11), 1507-1515.

StataCorp. (2007). Stata Statistical Software: Release 10. College Station, TX: StataCorp LP.

StataCorp (2011a). [Email communication].

StataCorp. (2011b). *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP.

Stubbs, J. M., & Achat, H. M. (2009). Individual rights over public good? The future of anthropometric monitoring of school children in the fight against obesity. *Medical Journal of Australia, 190*(3), 140.

Swearer, S. M., Siebecker, A. B., Johnsen-Frerichs, L. A., & Wang, C. (2010). Assessment of bullying/victimization: The problem of comparability across studies and across methodologies. In S. R. Jimerson, S. M. Swearer & D. L. Espelage (Eds.), *Handbook of bullying in schools: An international perspective* (pp. 305-328). New York: Routledge, Taylor & Francis.

Szabó, M. (2010). The short version of the Depression Anxiety Stress Scales (DASS-21): Factor structure in a young adolescent sample. *Journal of adolescence, 33*(1), 1-8.

Taylor, P. J., Russ-Eft, D. F., & Taylor, H. (2009). Gilding the outcome by tarnishing the past. Inflationary biases in retrospective pretests. *American Journal of Evaluation, 30*(1), 31-43.

Tigges, B. B. (2003). Parental consent and adolescent risk behavior research. *Journal of Nursing Scholarship, 35*(3), 283-289.

Ttofi, M. M., & Farrington, D. P. (2011). Effectiveness of school-based programs to reduce bullying: A systematic and meta-analytic review. *Journal of Experimental Criminology, 7*(1), 27-56.

Ttofi, M. M., Farrington, D. P., & Lösel, F. (2012). School bullying as a predictor of violence later in life: A systematic review and meta-analysis of prospective longitudinal studies. *Aggression and Violent Behavior, 17*, 405-418.

Twisk, J. W. R. (2003). *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide*: Cambridge University Press.

Twisk, J. W. R. (2006). *Applied multilevel analysis: A practical guide*. Cambridge: Cambridge University Press.

Underwood, M. K. (2002). Sticks and stones and social exclusion: Aggression among girls and boys. In P. K. Smith & C. H. Hart (Eds.), *Blackwell handbook of childhood social development* (pp. 533-548). Malden, MA: Blackwell.

Unger, J. B., Gallaher, P., Palmer, P. H., Baezconde-Garbanati, L., Trinidad, D. R., Cen, S., & Anderson Johnson, C. (2004). No News is Bad News. Characteristics of adolescents who provide neither parental consent nor refusal for participation in school-based survey research. *Evaluation Review, 28*(1), 52-63.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods, 3*(1), 4-70.

Varjas, K., Henrich, C. C., & Meyers, J. (2009). Urban middle school students' perceptions of bullying, cyberbullying, and school safety. *Journal of School Violence, 8*(2), 159-176.

Varjas, K., Meyers, J., Bellmoff, L., Lopp, E., Birckbichler, L., & Marshall, M. (2008). Missing voices: Fourth through eighth grade urban students' perceptions of bullying. *Journal of School Violence, 7*(4), 97-118.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2011). *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. New York: Springer.

Vreeman, R. C., & Carroll, A. E. (2007). A systematic review of school-based interventions to prevent bullying. *Archives of Pediatrics & Adolescent Medicine, 161*(1), 78-88.

Wang, M. T., Selman, R. L., Dishion, T. J., & Stormshak, E. A. (2010). A tobit regression analysis of the covariation between middle school students' perceived school climate and behavioral problems. *Journal of Research on Adolescence, 20*(2), 274-286.

Waters, S., Epstein, M., Cross, D., & Shaw, T. (2008). *A Randomised Control Trial to Reduce Bullying and Other Aggressive Behaviours in Secondary Schools. Final report*. Perth, Western Australia: Child Health Promotion Research Centre, Edith Cowan University.

Waters, S. K., Epstein, M., Cross, D., & Shaw, T. (2008). *A Randomised Control Trial to Reduce Bullying and Other Aggressive Behaviours in Secondary Schools. Final report*. Perth, Western Australia: Child Health Promotion Research Centre, Edith Cowan University.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural Equation Modeling: Concepts, Issues, and Applications* (pp. 56-75). Thousand Oaks, CA: Sage.

White, V. M., Hill, D. J., & Effendi, Y. (2004). How does active parental consent influence the findings of drug-use surveys in schools? *Evaluation Review, 28*(3), 246-260.

Windsor, R., Clark, N., Boyd, N. R., & Goodman, R. M. (2003). *Evaluation of Health Promotion, Health Education, and Disease Prevention Programs* (3rd ed.). New York: McGraw-Hill.

Wolfenden, L., Kypri, K., Freund, M., & Hodder, R. (2009). Obtaining active parental consent for school-based research: a guide for researchers. *Australian and New Zealand journal of public health, 33*(3), 270-275.

Wolke, D., Woods, S., Bloomfield, L., & Karstadt, L. (2001). Bullying involvement in primary school and common health problems. *Archives of Disease in Childhood, 85*(3), 197-201.

Wright, D. B. (2006). Comparing groups in a before–after design: When t test and ANCOVA produce different results. *British Journal of Educational Psychology, 76*(3), 663-675.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research and Evaluation, 12*(3), 1-26.

Ybarra, M. L., Boyd, D., Korchmaros, J. D., & Oppenheim, J. K. (2012). Defining and measuring cyberbullying within the larger context of bullying victimization. *Journal of Adolescent Health, 51*(1), 53-58.

Appendices

Appendix 2-A. Definition of Bullying and Cyberbullying

Bullying definition:

Please read the following information on **bullying** carefully.

Bullying is when one or more of the following things happen **AGAIN** and **AGAIN** to someone who finds it **hard to stop** it from happening again.

Bullying is when a person or a group of people **offline or online** (mobile phone or Internet):

- Make fun of / tease someone in a mean and hurtful way
- Tell lies or spread nasty rumours about someone to try to make others not like him/her
- Leave someone out on purpose or not allow him/her to join in
- Hit, kick or push someone around
- Deliberately damage, destroy or steal someone's things
- Threaten or make someone feel afraid of getting hurt

It is NOT bullying when:

- teasing is done in a friendly, playful way
- two people who are as strong as each other argue or fight.

[a] The forms of bullying behaviours were illustrated with pictures in color depicting online and offline scenarios

Cyberbullying definition:

Cyberbullying is bullying using a mobile phone and/or the Internet e.g. when a person:

- Is sent nasty or threatening emails or messages on the Internet or their mobile phone
- Has mean or nasty comments or pictures about them sent to websites e.g. MySpace; Facebook; MSN or to other students' mobile phones
- Is deliberately ignored or left out of things over the Internet
- Has someone else pretend to be them online to hurt them

Cyberbullying can happen through text messages/pictures/video-clips/emails etc. being **sent to you**, but also when these things are **sent to others, about you**.

[b] The cyberbullying definition was placed immediately after the bullying definition and followed by a colored picture depicting a cyberbullying scenario.

Appendix 2-B. Forms of Bullying Scale

Victimisation Version (FBS-V)

Qxx. Last term, how often were you bullied (INCLUDING cyber bullying) by one or more young people in the following ways?[a]

a	I was TEASED in nasty ways
b	SECRETS were told about me to others to hurt me
c	I was hurt by someone trying to BREAK UP A FRIENDSHIP
d	I was MADE TO FEEL AFRAID by what someone said he/she would do to me
e	I was deliberately HURT PHYSICALLY by someone and/or by a group GANGING UP on me
f	I was CALLED NAMES in nasty ways
g	Someone told me he/she WOULDN'T LIKE ME UNLESS I DID what he/she said
h	My THINGS were deliberately DAMAGED, DESTROYED or STOLEN
i	Others tried to hurt me by LEAVING ME OUT of a group or NOT TALKING TO ME
j	LIES were told and/or FALSE RUMOURS spread about me by someone, to make my friends or others NOT LIKE me

[a] For each item the respondent chooses one of the five response options as detailed in the Methods section titled "Construction of the FBS".

Perpetration Version (FBS-P)

Qxx. Last term, how often did you bully (or cyber bully) another young person(s) in the following ways (on your own or in a group)?[a]

a	I TEASED someone in nasty ways
b	I told SECRETS about someone to others to deliberately HURT him/her
c	I hurt someone by trying to BREAK UP A FRIENDSHIP they had
d	I deliberately FRIGHTENED or THREATENED someone
e	I deliberately PHYSICALLY HURT or GANGED UP on someone
f	I CALLED someone NAMES in nasty ways
g	I told someone I would NOT LIKE THEM UNLESS THEY DID what I said
h	I deliberately DAMAGED, DESTROYED and/or STOLE someone's things
i	I tried to hurt someone by LEAVING THEM OUT of a group or by NOT TALKING to them
j	I told LIES and/or spread FALSE RUMOURS about someone, to make their friends or others NOT LIKE them

[a] For each item the respondent chooses one of the five response options as detailed in the Methods section titled "Construction of the FBS".

Appendix 6-A. Bullying questions in Supportive Schools Project survey (Term 1, 2006)

The questions measuring bullying victimisation and perpetration given below were preceded in the survey by a definition of bullying based on that of Olweus (1996), which included examples and pictures illustrating the different forms of bullying, as well as two examples of behaviours that are not bullying. Students were asked to reflect on their experiences during Term 1, 2006.

1. So far this year (Term 1) how often did you have these things done to you by another student or students?

(please circle one number for each statement)

		Most days	About once a week	Every few weeks	Only once or twice	Never
a	I was made fun of and teased in a hurtful way	1	2	3	4	5
b	I was called mean and hurtful names	1	2	3	4	5
c	Students ignored me, didn't let me join in, or left me out on purpose	1	2	3	4	5
d	I was hit, kicked or pushed around	1	2	3	4	5
e	Students told lies about me and tried to make other students not like me	1	2	3	4	5
f	I had money or other things broken or taken away from me	1	2	3	4	5
g	I was made to feel afraid that I would get hurt	1	2	3	4	5

Bullying is when someone or people deliberately hurts a person <u>AGAIN AND AGAIN</u> and that person finds it hard to stop it from happening.

2. So far this year (2006), in TERM 1, how often did a student or group of students bully you?

(please circle one number)

a	I was bullied **MOST DAYS** in Term 1	1
b	I was bullied **ABOUT ONCE A WEEK** in Term 1	2
c	I was bullied **EVERY FEW WEEKS** in Term 1	3
d	I was bullied **ONLY ONCE OR TWICE** in Term 1	4
e	I was not bullied in Term 1	5

3. So far this year (Term 1), how often have you on your own or in a group, done these things to another student or students?

(please circle one number for each statement)

		Most days	About once a week	Every few weeks	Only once or twice	Never
a	I made fun of and teased another student or students in a hurtful way	1	2	3	4	5
b	I called another student or students mean and hurtful names	1	2	3	4	5
c	I ignored another student or students, didn't let them join in, or left them out of things on purpose	1	2	3	4	5
d	I hit, kicked or pushed another student or students around	1	2	3	4	5
e	I told lies or spread nasty stories about another student or students and tried to make other students not like them	1	2	3	4	5
f	I broke someone's things deliberately or took money or other things away from another student or students	1	2	3	4	5
g	I made another student or students feel afraid they would get hurt	1	2	3	4	5

4. So far this year (2006), in TERM 1, how often did you, on your own or in a group, bully another student or students?

(please circle one number)

a	I bullied someone **MOST DAYS** in Term 1	1
b	I bullied someone **ABOUT ONCE A WEEK** in Term 1	2
c	I bullied someone **EVERY FEW WEEKS** in Term 1	3
d	I bullied someone **ONLY ONCE OR TWICE** in Term 1	4
e	I did **NOT** bully anyone **AT ALL** in Term 1	5

Appendix A. Permission letter from journal Psychological Assessment

AMERICAN PSYCHOLOGICAL ASSOCIATION

INVOICE NO. **N/A**
Federal Tax I.D. 53-0205890
Date: July 11, 2013

IN MAKING PAYMENT REFER TO THE ABOVE INVOICE NUMBER

Therese Shaw
Edith Cowan University
2 Bradford St. Mt. Lawley
Perth, WA, 6050
Australia

APA Permissions Office
750 First Street, NE
Washington, DC 20002-4242
www.apa.org/about/copyright.html permissions@apa.org
202-336-5650 Fax: 202-336-5633

IF THE TERMS STATED BELOW ARE ACCEPTABLE, PLEASE SIGN AND RETURN ONE COPY TO APA. RETAIN ONE COPY FOR YOUR RECORDS. PLEASE NOTE THAT PERMISSION IS NOT OFFICIAL UNTIL APA RECEIVES THE COUNTERSIGNED FORM AND ANY APPLICABLE FEES.

Request is for the following APA-copyrighted material:

- Shaw, T., Dooley, J. J., Cross, D., Zubrick, S. R., & Waters, S. (2013, June 3). The Forms of Bullying Scale (FBS): Validity and Reliability Estimates for a Measure of Bullying Victimization and Perpetration in Adolescence. Psychological Assessment. Advance online publication. doi: 10.1037/a0032955

For the following use: **Non-Commercial Research or Educational Use in:** a) the print version of the final dissertation document;
b) the digital version of the final dissertation document posted online provided all conditions in item (3) below are met.

Permission is granted for the nonexclusive use of APA-copyrighted material specified above contingent upon fulfillment of the conditions indicated below:

1. The fee is waived.

2. The reproduced material must include the following credit line: Copyright © 2013 by the American Psychological Association. Reproduced [or Adapted] with permission. The official citation that should be used in referencing this material is [list the original APA bibliographic citation]. No further reproduction or distribution is permitted without written permission from the American Psychological Association.

3. **For online use:**
   (a) The credit line must appear on the first screen on which the APA content appears.

   (b) The online posting is permitted only on a non-commercial, secure and restricted web site. If condition (b) cannot be satisfied then the abstract may be reproduced along with a link to the journal table of contents at http://psycnet.apa.org/journals/pas/.

Note: APA does not grant you permission to reproduce the APA content on a Public Internet Site or in an online database of a commercial nature.

This agreement constitutes permission to reproduce only for the purposes specified on the attached request and does not extend to future editions or revisions, derivative works, translations, adaptations, promotional material, or any other formats or media. Permission

applies solely to publication and distribution in the English language throughout the world, unless otherwise stated. No changes, additions, or deletions to the material other than any authorized in this correspondence shall be made without prior written consent by APA.

This permission does not include permission to use any copyrighted matter obtained by APA or the author(s) from other sources that may be incorporated in the material. It is the responsibility of the applicant to obtain permission from such other sources.

ACCEPTED AND AGREED TO BY:	**PERMISSION GRANTED ON ABOVE TERMS:**
Applicant	for the American Psychological Association
15 July 2013	July 11, 2013
Date	Date

Appendix B. Permission letter from Australian Journal of Education

Australian J of Education
Sage Publications

27 May 2013

Dear Sir/Madam,

My name is Therese Shaw and I am completing a PhD thesis by publication at Edith Cowan University (ECU), Australia. Theses published at ECU are made digitally available on the World Wide Web for public access via the ECU Institutional Repository.
See URL: http://ro.ecu.edu.au/

My thesis includes the following research paper which is copyright material:

The clustering of bullying and cyberbullying behaviour within Australian schools

from the following journal for which you hold the exclusive license:

Australian Journal of Education

I wish to seek from you a limited, non-exclusive licence, for an indefinite period to include these materials for which you hold the exclusive license, in the hard bound copy and the digital copy of my thesis to be made available on the ECU Institutional Repository. Your works will of course be fully and correctly referenced.
Please sign below if you agree.

I, K. Ryan, agree to permit the non-exclusive licence for an indefinite period to include the above materials for which I am copyright owner, into the hard bound copy and the digital copy of your thesis to be included in the ECU Institutional Repository.

Position: Rights & Licensing Manager.
Date: 3 June 2013.

Yours sincerely,

Therese Shaw