Incredible Years

Follow-up Study

Published August 2014
Ministry of Social Development
PO Box 1556
Wellington 6140
New Zealand

Telephone: +64 4 916 3300
Facsimile: +64 4 918 0099

Email: info@msd.govt.nz
Web: www.msd.govt.nz

ISBN: 978-0-478-32361-0 (online)

Incredible Years
Follow-up Study

Long-term follow-up of the New Zealand Incredible Years Pilot Study

Fiona Sturrock, Dorian Gray, David Fergusson, John Horwood, Christina Smits

AUGUST 2014

Contents

List of Tables

List of Figures

Acknowledgements

The Incredible Years Pilot Study was a substantial cross-agency study conducted over 2 years. The Ministry of Education Special Education delivered the Incredible Years Parent courses evaluated in the study, the Ministry of Health provided the operational funding, and the evaluation team was located within the Ministry of Social Development.

We are grateful to the Incredible Years Evaluation Advisory Group for their guidance and advice on the Pilot Study. We also appreciate the support received from the Incredible Years professionals in the research sites, statistical experts at the Ministry of Social Development and the Christchurch Health and Development Study, administrative and collegial assistance from the Ministry of Social Development, the interviewers who conducted the fieldwork and, most importantly, the 166 families who participated in the Pilot Study.

The Incredible Years Follow-up Study built on the foundation provided by the Pilot Study. We are particularly indebted to the 136 families from the original study who responded to our request for further involvement in the New Zealand Incredible Years research.

Fiona Sturrock

Dorian Gray

David Fergusson

John Horwood

Christina Smits

Executive summary

- As part of the Drivers of Crime work programme the Ministries of Education, Health and Social Development established a pilot study of the Incredible Years Parent (IYP) programme to assess its effectiveness in reducing conduct problems in a New Zealand context.
- The New Zealand Incredible Years Pilot Study provided evidence to suggest that IYP, a programme developed overseas, can be successfully implemented in New Zealand and retain its general level of effectiveness for both Māori and non-Māori families.
- The effectiveness for Māori is particularly important given the higher rates of conduct problems reported for Māori children.
- The Pilot Study was a substantial 2-year, multiple-informant study that included mixed measurement methods, single case studies and a 6-month follow-up. A total of 166 parents took part.
- The main study of the Pilot was a repeated measures design in which all research participants were interviewed four times at home: at baseline before the IYP course began, mid-course, post-course and 6-month follow-up.
- Each interview covered a range of topics relating to child behaviour, parenting practices and relationships, and the family context. The Family Interview incorporated items from a number of previously validated measures to assess changes in outcome variables over the course of the study.
- The Follow-up Study investigated the long-term outcomes for 136 (82%) of the 166 children and parents who were in the original sample, 30 months on (range 28–32 months). The Follow-up Study Family Interview retained the suite of child behaviour, parenting practices and relationships and family context items used in the Pilot Study.
- The Follow-up Study sample of 136 participants was validated against any bias due to differential sample loss. No differences were found between the original and follow-up samples across a number of demographic variables, behaviour and parenting practices scores, and participation in the IYP programme.
- The key finding of the Follow-up Study is that the IYP programme outcomes were maintained over the 30-month follow-up with no diminution in the size of effects for almost all of the outcome measures. Specifically:
    - Compared to the baseline values, the findings at post-course showed clear and significant benefits in the areas of child behaviour, parenting, and family relationships. With a small number of exceptions, effect sizes were in the moderate to large range ($d>0.50$; $p<.001$).
    - Compared to the baseline values, the findings at the follow-up showed clear and significant benefits in the areas of child behaviour, parenting, and family relationships. As with the baseline/post-course comparisons, most effect sizes were in the moderate to large range ($d>.50$; $p<.001$).
    - Positive parenting and poor supervision were the only measures where significant differences were found between post-course and follow-up assessments.
- These conclusions suggest IYP was an effective programme that demonstrated benefits that persisted in the longer term. These findings are generally consistent with the findings of a small number of previous studies that have examined the long-term benefits of IYP.

- Other findings:
  - There were no differences between Māori and non-Māori families on measures of child behaviour, parenting, and family relationships at follow-up. These findings demonstrate that IYP is a programme that can be equally effective for both Māori and non-Māori families.
  - The large between-site variations in the efficacy of the IYP programme reported in the Pilot Study, with the effect sizes for the Mid-Central site being substantially greater than those for the Canterbury and Bay of Plenty sites, were evident at the 30-month follow-up.
  - Although children in both the clinical and sub-clinical ranges on the Eyberg Child Behaviour Inventory at baseline gained substantial benefits from the IYP programme, those in the clinical range tended to gain greater benefit than those in the sub-clinical range.
  - Parents reported high to moderate satisfaction with the IYP programme at the 30-month follow-up.
  - In general, these conclusions suggest that the pilot implementation of IYP has been highly successful and the programme has been demonstrated to show long-term benefits across a wide range of outcomes.

Background

This report presents the findings from a long-term follow-up evaluation of the Incredible Years Parent (IYP) programme in New Zealand. IYP is a parent management training programme developed for parents of children who have conduct problems.

The effective treatment and management of conduct problems are a high priority for New Zealand's health, education, justice and welfare sectors and form part of a broader cross-government priority work programme to address the Drivers of Crime. As part of this work programme, the Ministries of Education, Health and Social Development established a pilot study of the IYP programme to assess its effectiveness in reducing conduct problems in a New Zealand context. The project was influenced by the recommendations of the Government Advisory Group on Conduct Problems, the Ministry of Education's Positive Behaviour for Learning strategy and the desire to develop a new collaborative model to evaluate government-funded programmes.

Childhood conduct problems

Conduct problems are considered to be one of the most commonly occurring mental health issues among children and adolescents, with prevalence estimated to range from 5 to 10 percent of children both in New Zealand [1] and internationally [2]. Māori children in New Zealand have higher rates (15% to 20%) of conduct problems than non-Māori [3]. The far-reaching consequences of conduct problems for individual health, development and wellbeing are well documented [3]. Negative outcomes include antisocial behaviour, mental health difficulties, suicidal behaviours, substance abuse, teenage pregnancy, inter-partner violence and poor physical health [1, 4]. The wide range of negative outcomes associated with conduct problems has high social and fiscal costs [5, 6, 7].

Early intervention and the influence of parenting

Early intervention is recognised as a crucial element in reducing the onset of behaviour problems that often start in early childhood. A body of robust evidence shows early intervention can have a significant impact on child development and later life outcomes [5, 7, 8]. In addition, findings from a longitudinal study [9] demonstrated the enduring influence of parenting during the early years in a child's life. Children who experienced parenting that was warm, sensitive, cognitively stimulating and not intrusive or over-controlling early in life showed better cognitive functioning, academic achievement and social adjustment when in middle primary school. The opposite was true for children who did not experience this type of care.

Recognition that the early years lay the foundation for future development has led to investment in evidence-based prevention and treatment programmes for young children and their families [10]. These early intervention programmes seek to mitigate risks for vulnerable children by improving parental capabilities, addressing risk factors and enriching children's experiences.

Parent training programmes

Parent management training, based on social learning theory, is one of the most successful approaches to addressing conduct problems in early and middle childhood, particularly for children aged 3–7 years. Strong evidence from rigorous efficacy trials demonstrate that parent management programmes can improve parenting skills and reduce children's behavioural difficulties [11]. A range of manualised, well validated and widely used programmes are available, but few have as much empirical support as the IYP programme.

A review [12] of the numerous published randomised control group trials (RCT) conducted by Carolyn Webster-Stratton, the developer of IYP, and her colleagues documented the effectiveness of the IYP programme for young children with conduct problems [13]. These findings have been replicated in RCTs by independent investigators in England [14, 15], Wales [16, 17], Ireland [18] and Norway [19]. Studies using a matched control group methodology rather than a RCT [20, 21] have also demonstrated the effectiveness of IYP programmes in reducing childhood conduct problems and improving parenting practices.

The Incredible Years Parent, Teacher and Child Training series has been developed over the last 30 years at the University of Washington [22]. It has been implemented widely within the United States, the United Kingdom, Canada, Ireland, Norway, Denmark, Sweden, Australia and New Zealand. The Incredible Years series received a proven rating for IYP BASIC from the RAND Corporation [23], has been endorsed in a number of jurisdictions [24] and has been identified as one of 11 Blueprint interventions by the Centre for Violence Prevention at the University of Colorado, having satisfied stringent scientific criteria [17].

Cost-effectiveness

Childhood conduct problems are a major predictor of lifetime resource use that results in substantial costs in the education, health, justice and welfare sectors. Research shows the return from well implemented and well evaluated prevention, intervention, and treatment programmes for conduct problems is often very good, with programmes returning several times their costs as a result of reduced rates of crime, imprisonment and associated costs [1]. O'Neill's [18] cost-benefit analysis of the IYP programme suggests the long-term rate of return from parenting programmes is likely to be relatively high, while Scott [2] estimated the longer-term return from IYP training to be 10 times higher than its cost.

Although there is no guarantee cost-benefit analyses conducted overseas will apply in New Zealand, there is a universal consensus in the literature that a long-term investment strategy is likely to be highly cost-effective, providing the investment is made in well founded and well implemented evidence-based programmes [1].

The New Zealand Incredible Years Pilot Study

The New Zealand Incredible Years Pilot Study [25] provided evidence to suggest IYP, a programme developed overseas, can be successfully implemented in New Zealand and retain its general level of effectiveness for both Māori and non-Māori families. The effectiveness for Māori is particularly important. The higher rates of conduct problems reported for Māori children mean that, to reduce conduct problems in New Zealand, parenting interventions must be effective for and acceptable to Māori.

The Incredible Years Pilot Study was a substantial 2-year, multiple-informant study that included mixed measurement methods, single case studies and a 6-month follow-up. All parents enrolled in the 18-week IYP courses at three Special Education sites in 2011 were invited to take part in the evaluation; no parents were excluded from the opportunity to participate in the research.

The main study of the Pilot was a repeated measures design in which all research participants were interviewed four times: at baseline before the IYP course began, mid-course, post-course and 6-month follow-up. Trained interviewers employed for the Pilot Study conducted the four home-based interviews with the primary caregivers. Each interview covered a range of topics relating to child behaviour, parenting practices and relationships, and the family context. The Family Interview incorporated items from a number of previously validated measures to assess changes in outcome variables over the course of the study.

Detailed findings from the Pilot Study are available in the evaluation report [25]. Key findings were:

- There was clear evidence of child behaviour change, with effect sizes measured by Cohen's d in line with the international literature.
- There was clear evidence of parenting behaviour change, with effect sizes measured by Cohen's d in line with the international literature.
- Improvements were maintained 6 months following course completion.
- The benefits of the IYP training were broadly similar for Māori and non-Māori families. Nevertheless the evidence suggested the need for further work on maximising gains for Māori families, particularly in the maintenance of behaviour change.
- Although improvements were evident at all three sites, larger effect sizes were found in Mid-Central than in the Bay of Plenty or Canterbury.
- Children in both the clinical and sub-clinical ranges on the pre-course Eyberg Child Behaviour Inventory (ECBI) Scaled Intensity measure displayed evidence of improved behaviour, but those in the clinical range improved to a greater extent.
- Both Māori and non-Māori parents expressed high to moderate satisfaction with the programme.

The Pilot Study provided evidence of the short-term to medium-term efficacy of the IYP programme in New Zealand. Children and parents receiving IYP training showed significant improvements in the following areas: child behaviour median effect size of d=.65 (range d=.51 to d=.96), parenting practices median effect size of d=.54 (range d=.26 to d=.83) and relationships median effect size of d=.48 (range d=.21 to d=.60). The improvements evident at the completion of the IYP course were mostly sustained at the 6-month follow-up.

Significant improvements were maintained in the following areas: child behaviour median d=.71 (range d=.56 to d=1.0), parenting practices median d=.52 (range d=.25 to d=.79) and relationships median d=.43 (range d=.15 to d=.59).

The New Zealand Incredible Years Follow-up Study

The literature on the effectiveness of the IYP programme beyond 6 months post-course is limited. Significant post-intervention improvements in child and parent behaviour were reported at 12 months [26], 18 months [27], 2 years [21, 28] and 3 years [29] after the delivery of an IYP programme. The p-values for significant differences in child behaviour at these long-term follow-ups range from p<.001 to p=.05. One study [26] reports a large effect size of .97 for child behaviour change.

To assess the stability of post-intervention gains in the long term in New Zealand, the present study investigated outcomes for the Pilot Study children and their parents 30 months (range 28–32 months) after the start of their IYP programmes.

The following research questions guided the Follow-up Study:

- Are improvements in child behaviour and parenting practice maintained over the longer follow-up period?
- How much drop-off in efficacy is evident between the post-course and 30-month follow-up measures?
- Do sub-group differences identified in the Pilot Study persist?

Methodology

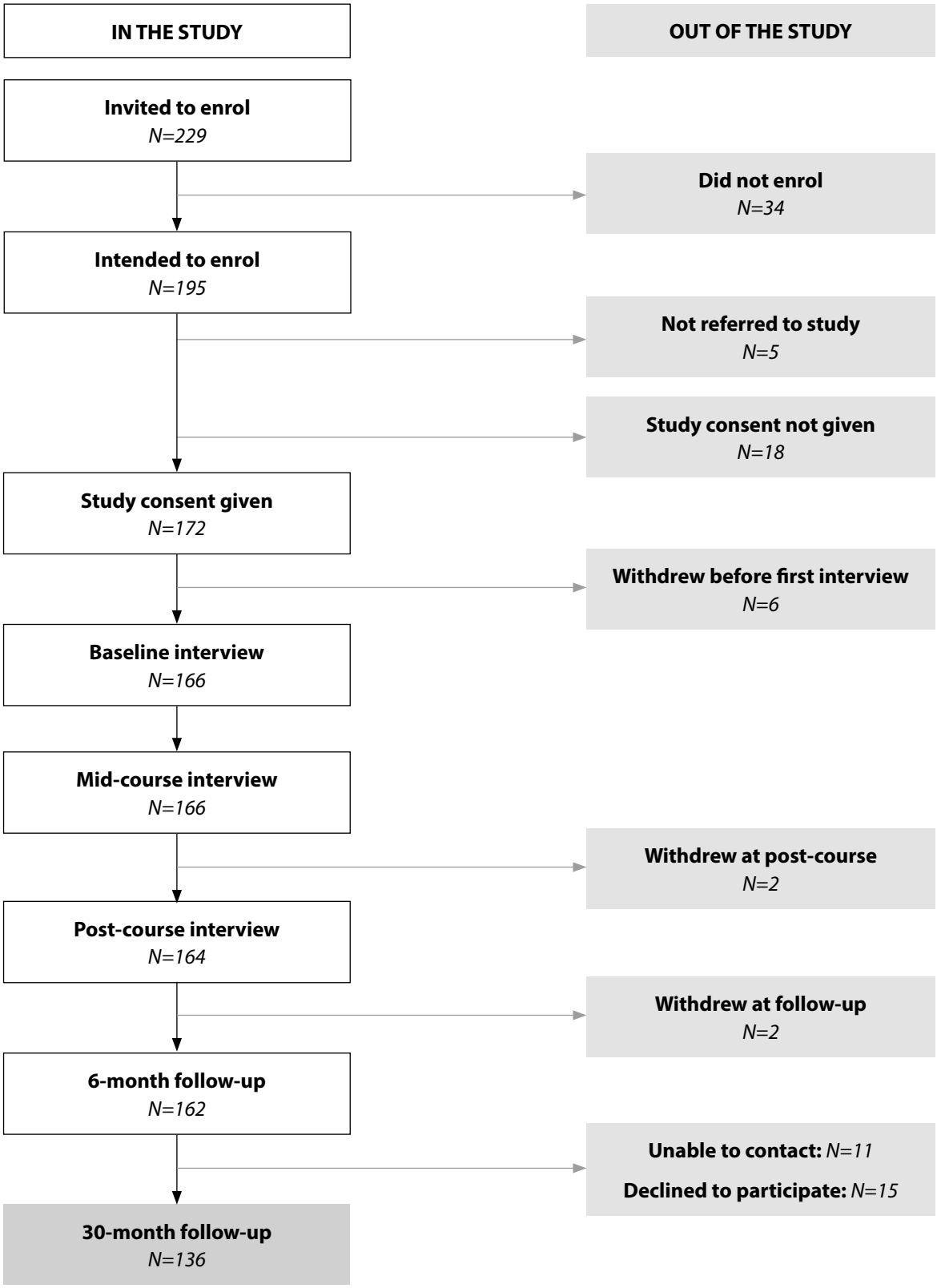
Parent recruitment and retention

Figure 1 shows the processes of sample selection, retention and loss from the point of initial recruitment for the Pilot Study to the 30-month follow-up. The original Pilot Study sample consisted of 166 parents of children aged 4–8 years referred to an IYP course because of their children's conduct problems. All parents enrolled in these courses, run in three Special Education sites in 2011, were invited to participate. Ninety-eight percent (162) of parents who completed the baseline interview remained in the research through to the 6-month follow-up.

Letters were sent to the 162 parents who had completed the 6-month interview explaining the reason for contacting them again. Interviewers then followed up the letters with a phone call to arrange the interview, either in the home or by telephone. The same trained interviewers employed for the Pilot Study conducted the follow-up interviews with primary caregivers.

Of the original 166 families who agreed to participate in the Pilot Study, 136 (82%) completed a further follow-up interview approximately 30 months (range 28–32 months) after the beginning of their IYP course. Participant loss at this long-term follow-up was due to withdrawal before the 6-month interview (N=4; 2%), no longer wanting to participate (N=15; 9%) and not reachable by phone or letter (N=11; 7%). This retention rate is consistent with rates reported for other long-term follow-up studies of the IYP programme. Retention rates in these studies range from 82 percent of mothers at 3 years [29], to 84 percent at 1 year [26], 88 percent at 18 months [27] and 91 percent at 2 years [28].

Figure 1: Incredible Years Pilot Study and Follow-up Study participant recruitment and retention



The Family Interview

The Follow-up Study Family Interview retained the suite of child behaviour, parenting practices and relationships, and family context items used in the Pilot Study [25]. The interview incorporated items from a number of previously validated measures to assess the maintenance of changes in outcome variables.

Parents answered questions about their child's behaviour during the past 4 weeks using 111 items based on four recognised scales: the Eyberg Child Behaviour Inventory [30, 31], the Strengths and Difficulties Questionnaire [32], the Incredible Years Social Competence Scale [33] and items from the 5-year evaluation of Early Start [34]. Confirmatory factor analysis of this item set for the Pilot Study [35] showed these items measured six correlated dimensions of child behaviour: Conduct Disorder (CD), Oppositional Defiant Disorder (ODD), Attention Deficit Hyperactivity Disorder (ADHD), Self-control, Anxiety/Withdrawal and Social Competence. Re-testing of instrument reliability at the 30-month follow-up detected no differences in the Cronbach's alpha coefficients obtained at baseline.

The parenting practices measures in the Family Interview consisted of two recognised instruments: the Alabama Parenting Questionnaire [36] and the Arnold-O'Leary Parenting Scale [37]. The revised Straus Parent/Child and Partner Conflict Tactics Scales [38, 39], the Dadds Parent Problem Checklist [40] and the Partner Attachment Scale [41] measured parents' relationships with their children and partners.

To gauge satisfaction with the programme, respondents were asked to rate their agreement with 13 statements about the IYP course overall. These items were based on the Incredible Years Parent Satisfaction Questionnaire. Several open-ended questions provided parents with the opportunity to expand on their experiences of IYP.

See Appendix 1 for a detailed description of the source instruments.

Statistical analyses

The analyses presented in this report used general linear modelling and the chi-square test of association to determine: i) the validity of the 30-month follow-up sample; ii) the outcomes of the IYP programme at post-course through to the follow-up; and iii) the sustainability of IYP outcomes between post-course and follow-up. Effect sizes were calculated to estimate the size of the outcomes. All data are based on primary caregivers' self-reports.

The IYP outcomes on child behavioural changes and parental behaviour changes were tested under the hypothesis that there were differences in the mean observed scores between the baseline and the post-course interviews. The sustainability of these outcomes was tested under two hypotheses: i) that there were differences in the mean observed scores between the baseline and the follow-up interviews; and ii) that there were *no* differences in the mean observed scores between the post-course and follow-up interviews.

The outcome tables in this report use superscripted letters of 'a', 'b', or 'c' after each mean score to show the results of the hypothesis tests described above. When comparing baseline, post-course and follow-up mean scores:

- Mean scores with the same superscript are not statistically different from each other.
- Mean scores with different superscripts (for example 'a' versus 'b' or 'c') are statistically different from each other with p-values of $p<.05$.
- Mean scores superscripted with 'a, b' indicate the score is different from the score with the superscript 'a' and the score with the superscript 'b'.

The size of the differences is expressed with Cohen's d effect-size estimations [42]. Cohen suggests that an effect size of d=.20 represents a small difference, d=.50 a medium-sized difference, and d=.80 is a large difference. Effect sizes for the post-course outcomes and the follow-up outcomes are provided in the tables along with 95% confidence intervals around the estimates.

The tables show the mean scores for each measure through the observation points of baseline, post-course, and follow-up. The mean scores of each measure are statistically tested for evidence of linear trends over time, which are shown as p-values. Because each table contains a number of measures and tests, there is an increased chance of finding a significant trend by chance. To account for this, the significance testing is adjusted using a Bonferroni correction to the α=.05 level by the number of measures in each table, thereby reducing the chance of false-positive conclusions.

The sustainability of IYP programme outcomes was compared between potential confounding categories of ethnicity, the research site, and the initial diagnosed conduct problem. These comparisons are presented in charts showing the estimated effect sizes for each measure. Statistical tests on the interactions demonstrate where the IYP programme outcomes did not differ within the categories of interest. In addition, multivariate analyses across all the measures in each chart were used to test that the overall IYP programme outcome was not different within the categories, thereby demonstrating that any significant difference or lack of difference within each category was not found by chance.

See Appendix 2 for a detailed description of the statistical analyses.

Results

Comparison of baseline and follow-up samples

At the 30-month follow-up interviews, 30 of the 166 original participants were either not contactable or declined to participate. The loss of 30 participants raises the question of whether or not the sample loss was non-random, which could pose a threat to study validity. Table 1 examines the sample loss by comparing the 136 participants in the follow-up sample with the 30 non-participants on a series of measures gathered at baseline. The measures tested to verify the sample validity include:

- demographics: child age, gender and ethnicity; caregiver age and education; and household characteristics
- IYP programme attendance
- measures of behavioural adjustment, parenting and family functioning.

These comparisons were tested for statistical significance using chi-square tests on categorical data and t-tests for continuous measures, and show that:

- In terms of demographic factors, there were no significant differences between those studied and those not studied at the follow-up interview.
- There were no significant differences between parents studied at the follow-up who had at least some exposure to IYP courses and those who had no exposure.
- For most comparisons on behavioural adjustment and parenting practices there were no significant differences between those parents studied and those who were not studied at the follow-up. However, significant differences ($p<.05$) were evident for the child's Conduct Disorder and Self-control, and the parent's positive parenting and lax discipline.

These findings suggest that overall there were few differences between the baseline sample and the sample assessed at follow-up. While some significant differences were noted, these could be due to chance as a result of multiple significance testing. To explore this possibility, a Bonferroni corrected p-value was used to assess the significance of the findings in Table 1. Using this value ($p<.002$), none of the differences in the table were significant. Collectively, these results suggest it was unlikely that sample losses posed a major threat to the validity of the findings in this report.

Table 1. Comparison of baseline characteristics of those studied and those not studied at follow-up

Measure	Not studied (N=30)	Studied (N=136)	P[10]
Child characteristics			
Mean (sd) age of child	6.3 (1.9)	5.8 (2.0)	.180
% Male	60.0	75.7	.079
% Māori	40.0	35.3	.627
Caregiver/household characteristics			
Mean (sd) age of caregiver	33.2 (8.8)	34.5 (7.4)	.379
Mean (sd) household size	3.8 (1.7)	4.2 (1.4)	.269
% No formal qualifications (primary caregiver)	46.7	30.2	.082
% Family employment[1]	50.0	57.4	.463
% Single parent family	56.7	42.7	.163
% Family in receipt of welfare	46.7	40.4	.531
Programme attendance			
Attended IYP training (or some exposure to)[2]	86.7	93.4	.215
Child behaviour (baseline)			
Mean (sd) Conduct Disorder (CD)	31.1 (8.6)	28.3 (6.4)	.043
Mean (sd) ODD[3]	37.2 (7.7)	35.4 (7.2)	.229
Mean (sd) ADHD[4]	36.3 (9.0)	33.8 (8.4)	.138
Mean (sd) Self-control	33.3 (8.4)	36.4 (7.5)	.049
Mean (sd) Anxiety/Withdrawal	29.1 (5.1)	27.1 (5.9)	.081
Mean (sd) Social Competence	52.7 (11.4)	55.0 (10.2)	.285
All mean scores (multivariate) interaction with 'observed/not observed': F(6,159)=1.31, p=.256.			
In-course behavioural assessments			
Mean (sd) ECBI Intensity Score (N=25, 126)[5]	65.0 (13.4)	65.1 (9.5)	.989
% in the clinical range	64.0	73.8	.317
Mean (sd) Social Competence Score (N=24, 126)[6]	18.0 (8.1)	16.9 (7.8)	.509
% in the clinical range	41.7	57.1	.163
Parenting practices (baseline)[7]			
Poor supervision	6.7 (1.2)	6.6 (1.3)	.639
Positive parenting	14.2 (1.2)	13.6 (1.6)	.040
Corporal punishment	3.9 (0.9)	3.7 (0.8)	.315
Parental involvement	12.1 (2.0)	11.8 (2.0)	.470
Inconsistent discipline	7.2 (2.4)	7.4 (2.2)	.583
All mean scores (multivariate) interaction with 'observed/not observed': F(5,160)=1.05, p=.389.			
Dealing with misbehaviour (baseline)[8]			
Lax discipline	8.2 (2.2)	8.2 (2.1)	.921
Over-reactive discipline	8.7 (2.6)	8.7 (2.1)	.944
Hostile discipline	4.6 (1.1)	4.4 (1.1)	.324
Total Scale[9]	1.7 (0.3)	1.6 (0.3)	.959
All mean scores (multivariate) interaction with 'observed/not observed': F(3,162)=0.47, p=.704.			

Notes:

1 Family employment means at least one of the parents is employed.

2 Some parents did not attend the intended IYP course but shifted to a different course or took the course at a different date. These parents are treated as having some exposure to the IYP programme.

3 Oppositional Defiant Disorder.

4 Attention Deficit Hyperactivity Disorder.

5 Not all of the participants' children had an in-course assessment recorded, hence for the ECBI assessment N(not observed, observed)=25, 126.

6 For the Social Competency Score (SCS) assessment, N(not observed, observed)=24, 126.

7 From the Alabama Parenting Questionnaire.

8 From the Arnold-O'Leary Parenting Scale.

9 The 'Total Scale' includes questions from the Arnold-O'Leary Parenting Scale that are not included in the Lax, Over-reactive, or Hostile discipline factors, and the 'Total scale' scores are standardised to an average score between 1 and 3.

10 For significance testing a Bonferroni adjustment to the p-value of $p<.05$ is made to account for the 28 tests. As a result, only p-values where $p<.002$ are considered significant.

Child behaviour at baseline, post-course and follow-up

Table 2 reports on a series of parentally reported child behaviour scores assessed at baseline, post-course and follow-up. These scores assess Conduct Disorder, Oppositional Defiant Disorder, Attention Deficit Hyperactivity Disorder, Self-control, Anxiety/Withdrawal and Social Competence. The post-course scores were assessed within 2 months of the IYP programme's completion and the follow-up scores were obtained 28–32 months following the start of the course.

The table contains means and standard deviations and also gives planned comparison tests of the differences between: i) baseline and post-course; ii) baseline and follow-up; and iii) post-course and follow-up. The results of these tests are shown by the superscripts ('a' and 'b') on the mean scores for each behaviour measure. Mean scores with the same superscript are not significantly different from each other. Mean scores with different superscripts are significantly different at p-values of $p<.05$. The table also shows estimates of effect size, using Cohen's d, for the baseline/post-course and baseline/follow-up comparisons. The results in Table 2 reveal a highly consistent set of findings as summarised here:

1) For all child behaviour outcomes, there was a highly significant ($p<.001$) linear trend of change.
2) The post-course and follow-up mean scores were significantly different ($p<.05$) from the baseline mean scores, suggesting improvements in child and parent behaviours following the IYP course.
3) The post-course and follow-up means were not significantly different from each other, suggesting the gains from the IYP course had not diminished by the follow-up.
4) All effect sizes for both baseline/post-course and baseline/follow-up comparisons were in the range classified as moderate to large ($d>.50$).

This pattern of results suggests the provision of the IYP programme was associated with substantial changes in a wide range of behavioural outcomes, and the benefits of the programme were still evident several years after the programme's completion, without any detectable diminution in the size of effect.

Table 2: Comparison of child behaviour scores at baseline, post-course and the 30-month follow-up

Behaviour measures[1]	Baseline mean (sd)		Post-course mean (sd)		Follow-up mean (sd)		*p*[4]	Post-course effect size *d (95% CI)*		Follow-up effect size *d (95% CI)*	
Conduct Disorder (CD)	28.3 [a]	(6.4)	23.9 [b]	(5.2)	23.8 [b]	(6.0)	<.001	.67	(.43 to .92)	.69	(.45 to .94)
ODD[2]	35.4 [a]	(7.2)	28.1 [b]	(6.3)	28.4 [b]	(7.0)	<.001	1.0	(.78 to 1.3)	.98	(.73 to 1.2)
ADHD[3]	33.8 [a]	(8.4)	28.8 [b]	(8.1)	28.3 [b]	(8.8)	<.001	.59	(.34 to .83)	.65	(.40 to .89)
Self-control	36.4 [a]	(7.5)	44.1 [b]	(7.8)	44.5 [b]	(9.1)	<.001	1.0	(.76 to 1.3)	1.1	(.81 to 1.3)
Anxiety/Withdrawal	27.1 [a]	(5.9)	24.0 [b]	(5.5)	23.8 [b]	(6.3)	<.001	.52	(.28 to .76)	.55	(.31 to .79)
Social Competence	55.0 [a]	(10.2)	62.4 [b]	(10.0)	62.6 [b]	(11.0)	<.001	.73	(.49 to .98)	.75	(.51 to 1.0)

Notes:

[1] Sample size N=136.

[2] Oppositional Defiant Disorder.

[3] Attention Deficit Hyperactivity Disorder.

[4] p-values of p<.008 may be considered significant after Bonferroni correction for multiple tests of significance (Bonferroni adjustment to α=.05 for six tests).

[5] The superscripts ('a' and 'b') on the mean scores denote the statistical differences between mean scores across baseline, post-course, and follow-up measures. Mean scores with different superscripts are significantly different (p<.05), while mean scores with the same superscript are not significantly different (p>.05).

Parenting and disciplinary practices at baseline, post-course and follow-up

Mean scores for a number of dimensions of parenting and child disciplinary practices assessed at baseline, post-course and follow-up are reported in Table 3. These measures spanned a number of domains including parental supervision, positive parenting practices, punishment practices, parental involvement and child discipline.

The findings presented in Table 3 show that:

1) There were significant improvements in parenting behaviours at the post-course and follow-up assessments at the $p<.05$ level. Taking into account the Bonferroni adjustment of $p<.005$, only the 'Poor supervision' measure did not show a significant improvement ($p=.024$).
2) The results of planned comparisons showed that, with the exception of the 'Poor supervision' measure, the pattern of results was the same, with the baseline means being significantly different from both the post-course and follow-up means.
3) With the exception of 'Positive parenting' the post-course and follow-up means were not significantly different from each other.
4) The majority of effect sizes for both the post-course and follow-up comparisons fell into the range of moderate to large ($d>.50$). The exceptions to this were for 'Poor supervision' and 'Positive parenting', where some effect sizes were less than .50.

These findings lead to two general conclusions:

1) With the possible exception of 'Poor supervision', there was consistent evidence of improvements and positive changes in parenting behaviours between the baseline assessment and both the post-course and follow-up assessments, with most effect sizes being moderate to large ($d>.50$).
2) The comparisons between the post-course and follow-up mean scores show that, with the exception of 'Positive parenting', these improvements were sustained over a 30-month follow-up period, with no evidence of diminution of effect size over time. Although there was significant improvement in 'Positive parenting' between baseline and follow-up, some slippage was evident between post-course and follow-up.

Table 3: Comparison of parenting practice scores at baseline, post-course and the 30-month follow-up

Parenting measures[1]	Baseline mean (sd)		Post-course mean (sd)		Follow-up mean (sd)		*p*[5]	Post-course effect size *d (95% CI)*		Follow-up effect size *d (95% CI)*	
Parenting practices[2]											
Poor supervision	6.6 [a]	(1.3)	6.3 [b]	(1.0)	6.4 [a,b]	(0.9)	.024	.22	(-.02 to .46)	.17	(-.07 to .41)
Positive parenting	13.6 [a]	(1.6)	14.4 [b]	(0.9)	14.0 [c]	(1.2)	<.001	.53	(.29 to .77)	.30	(.06 to .54)
Corporal punishment	3.7 [a]	(0.8)	3.2 [b]	(0.5)	3.3 [b]	(0.6)	<.001	.60	(.36 to .84)	.55	(.31 to .79)
Parental involvement	11.8 [a]	(2.0)	12.8 [b]	(1.7)	13.0 [b]	(1.7)	<.001	.55	(.30 to .79)	.62	(.37 to .86)
Inconsistent discipline	7.4 [a]	(2.2)	6.1 [b]	(1.9)	6.2 [b]	(1.8)	<.001	.61	(.37 to .86)	.58	(.33 to .82)
Dealing with misbehaviour[3]											
Lax discipline	8.2 [a]	(2.1)	7.0 [b]	(1.8)	6.9 [b]	(1.8)	<.001	.58	(.33 to .82)	.61	(.36 to .85)
Over-reactive discipline	8.7 [a]	(2.1)	7.1 [b]	(1.7)	7.2 [b]	(1.4)	<.001	.77	(.53 to 1.0)	.72	(.48 to .97)
Hostile discipline	4.4 [a]	(1.1)	3.6 [b]	(0.9)	3.7 [b]	(0.9)	<.001	.69	(.45 to .94)	.63	(.39 to .87)
Total scale[4]	**1.6 [a]**	**(0.3)**	**1.4 [b]**	**(0.2)**	**1.4 [b]**	**(0.2)**	**<.001**	**.88**	**(.63 to 1.1)**	**.82**	**(.57 to 1.1)**

Notes:

1 Sample size N=136.

2 From the Alabama Parenting Questionnaire.

3 From the Arnold-O'Leary Parenting Scale.

4 The 'Total scale' includes questions from the Arnold-O'Leary Parenting Scale that are not included in the Lax, Over-reactive, or Hostile discipline factors, and the 'Total scale' scores are standardised to an average score between 1 and 3.

5 p-values of p<.005 may be considered significant after Bonferroni correction for multiple tests of significance (Bonferroni adjustment to α=.05 for nine tests).

6 The superscripts ('a', 'b', and 'c') on the mean scores denote the statistical differences between mean scores across baseline, post-course, and follow-up measures. Mean scores with different superscripts are significantly different from each other (p<.05), while mean scores with the same superscript are not significantly different (p>.05).

Family relationships at baseline, post-course and follow-up

Table 4 reports on mean scores for a number of measures of family relationships at the baseline, post-course and follow-up assessments. The measures are based on: i) the Parent/Child Conflict Tactics Scale, to assess levels of verbal and physical aggression between the parent and child; ii) the Revised Conflict Tactics Scale, to assess levels of violence between parents; iii) the Parent Problem Checklist, to assess child-rearing disagreement; and iv) the Partner Attachment Scale, to assess partner relationship quality.

The pattern of findings in Table 4 is similar to that seen in Tables 2 and 3, with evidence of reductions in levels of family conflict in both the post-course and follow-up assessments. However, effect sizes were more mixed with the relationship measures than with the measures reported in Tables 2 and 3. For instance, 'Inter-parental violence' saw small effect sizes (d=.27 to .39) albeit significant (p<.004), as did 'Relationship quality' with d=.31 and .40 (p<.001). Effect sizes for 'Conflict between partners' ('other parent') and child were medium (d=.45 to .49, p<.001). All other measures in Table 4 showed medium effect sizes of d>.50 (p<.001).

In general, these findings show evidence of improvements in family relationships at post-course, with these improvements being sustained without diminution of effect size at the 30-month follow-up.

Benefits of IYP for Māori and non-Māori families

An important feature of the evaluation of the IYP programme was the inclusion of sufficient numbers of Māori children to enable comparisons between Māori and non-Māori families. The previous evaluation [25] showed that the IYP programme was effective for both Māori and non-Māori in terms of a range of outcomes. However, it was noted that, while the programme benefited both groups, it had greater benefit for non-Māori for child behaviour outcomes. These differences were small and could have been a consequence of the large number of Māori/non-Māori comparisons made. To examine this issue, comparisons of the behavioural outcomes of Māori and non-Māori children were made at the 30-month follow-up.

Figure 2 presents estimates of the effect size (Cohen's d) and 95% confidence intervals for the baseline/follow-up comparison for the measures of child behaviour. Visual inspection shows the effect sizes were similar and there was a substantial overlap of the findings for Māori and non-Māori children. This was confirmed by a multivariate statistical test across all the measures overall, showing the findings for Māori and non-Māori for the baseline/follow-up comparison were not significantly different ($F(6,129)=0.68$, $p=.670$).

This analysis was repeated for measures of parenting and family relationships, and in all cases the findings showed the programme benefits for Māori were not significantly different from the programme benefits for non-Māori. Collectively, these analyses provide substantial reassurance that the benefits of the IYP programme are similar for Māori and non-Māori families.

Table 4: Comparison of parenting relationship scores at baseline, post-course and the 30-month follow-up

Relationship measures	Baseline mean (sd)		Post-course mean (sd)		Follow-up mean (sd)		*p*[5]	Post-course effect size *d (95% CI)*		Follow-up effect size *d (95% CI)*	
Primary caregiver and child[1]											
Verbal aggression	8.3 [a]	(2.1)	6.9 [b]	(1.9)	7.1 [b]	(1.8)	<.001	.67	(.42 to .91)	.61	(.37 to .85)
Physical assault	11.9 [a]	(1.2)	11.3 [b]	(0.8)	11.3 [b]	(0.7)	<.001	.50	(.26 to .74)	.51	(.27 to .75)
Other parent[2] **and child**											
Verbal aggression	7.5 [a]	(2.2)	6.5 [b]	(1.5)	6.4 [b]	(1.8)	<.001	.45	(.12 to .78)	.49	(.18 to .81)
Physical assault	11.8 [a]	(1.3)	11.2 [b]	(0.8)	11.2 [b]	(0.7)	<.001	.48	(.15 to .81)	.47	(.15 to .79)
Inter-parental violence[3]											
Violence to partner	22.5 [a]	(2.3)	21.9 [a,b]	(1.4)	21.7 [b]	(1.4)	.003	.27	(-.06 to .60)	.33	(.01 to .64)
Violence from partner	22.7 [a]	(2.6)	21.8 [b]	(1.4)	21.7 [b]	(1.5)	.001	.32	(-.01 to .65)	.39	(.07 to .71)
Inter-parental relationship[4]											
Child-rearing disagreement	25.2 [a]	(6.6)	21.6 [b]	(5.0)	20.6 [b]	(4.8)	<.001	.54	(.21 to .87)	.69	(.36 to 1.0)
Relationship quality	31.2 [a]	(4.7)	32.6 [b]	(3.2)	33.1 [b]	(2.8)	<.001	.31	(-.02 to .64)	.40	(.09 to .72)

Notes:

1 Scores are based on the Parent/Child Conflict Tactics Scale; for the primary caregiver, N=136.

2 For the other parent and inter-parental factors (where applicable), N=78 (baseline), 72 (post-course), and 78 (follow-up).

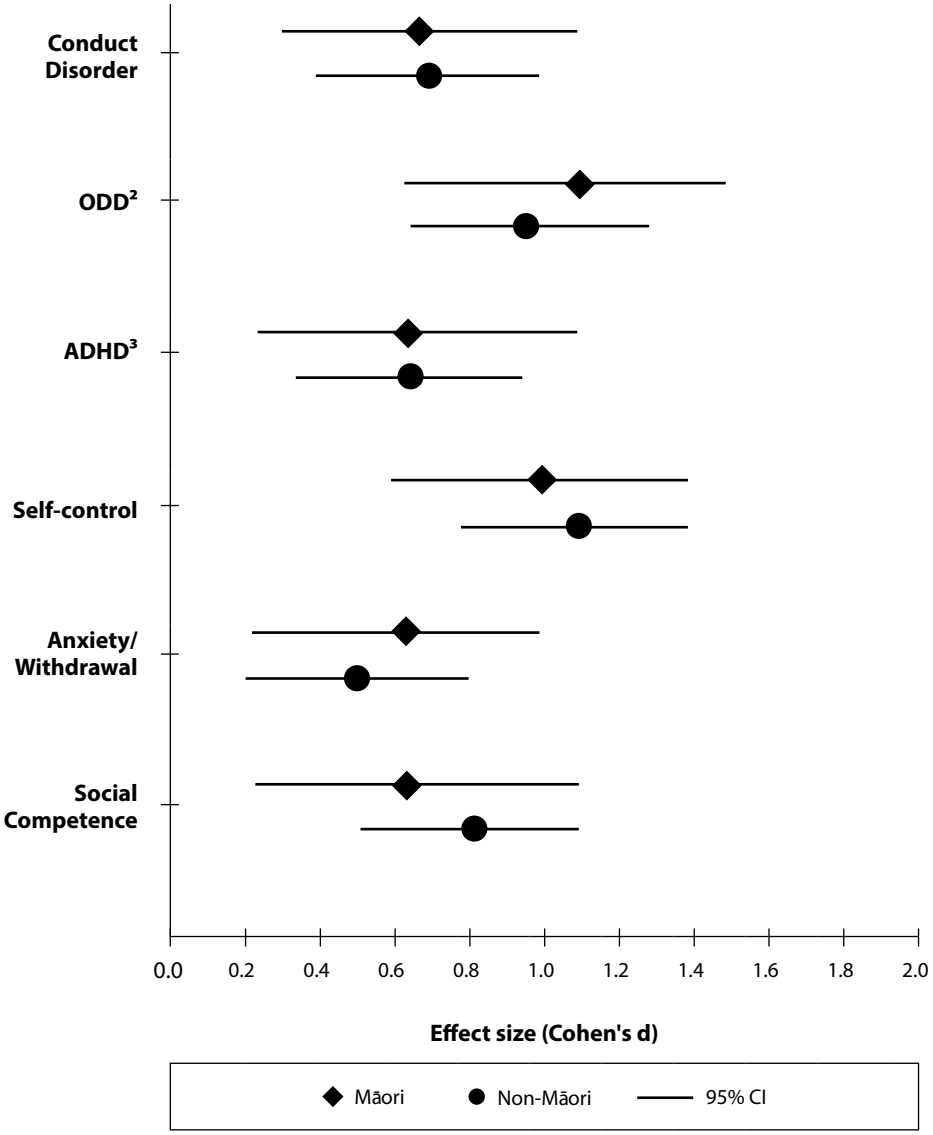
3 Inter-parental violence is from the Revised Conflict Tactics Scale (CTS2).

4 Child-rearing disagreement is from the Parent Problem Checklist, and Relationship quality is from the Partner Attachment Scale.

5 p-values of p<.006 may be considered significant after Bonferroni correction for multiple tests of significance (Bonferroni adjustment to α=.05 for eight tests).

6 The superscripts ('a' and 'b') on the mean scores denote the statistical differences between mean scores across baseline, post-course, and follow-up measures. Mean scores with different superscripts are significantly different (p<.05), while mean scores with the same superscript are not significantly different (p>.05).

Figure 2: Comparison of effect sizes for child behaviours for Māori and non-Māori children at the 30-month follow-up



1 Sample size for Māori is N=48 and for non-Māori is N=88.
2 Oppositional Defiant Disorder.
3 Attention Deficit Hyperactivity Disorder.
4 The 95% confidence interval lines (95% CI) around the estimated effect size represent the range of effect sizes one would expect to see 95% of the time if the experiment was repeated. Any differences between the Māori and non-Māori effect sizes are determined by the p-values from the analysis of ethnicity interactions.

Between-site differences

As noted previously, the IYP programme was piloted at three sites: Canterbury, Mid-Central and Bay of Plenty. In the previous evaluation [25], findings suggested that, while there were significant improvements in all three sites, the Mid-Central site had substantially better results than either the Canterbury or Bay of Plenty sites. These differences were sustained at the long-term follow-up.

Figure 3 provides a graphical comparison of the effect sizes for a series of behavioural outcomes for the three sites. The behavioural measures include Conduct Disorder, Oppositional Defiant Disorder, Attention Deficit Hyperactivity Disorder, Self-control, Anxiety/Withdrawal, and Social Competence. Figure 3 shows that for four outcomes (Conduct Disorder; Oppositional Defiant Disorder; Attention Deficit Hyperactivity Disorder; Anxiety/Withdrawal), the Mid-Central site, represented by the square, had higher effect sizes than the other two sites. Moreover, for all the behavioural measures Mid-Central shows consistently higher effect sizes as verified by a multivariate statistical test of the behavioural measures overall ($F(12,256)=2.69$, $p=.002$).

The reasons for the better results at the Mid-Central site are not known. The study findings do, however, suggest the possibility of substantial between-site variation in the efficacy of the IYP programme that is maintained over time, and the need to monitor sites to ensure the consistency of programme delivery and implementation.

Relationship between clinical status at baseline and outcomes at follow-up

One of the important debates regarding the IYP programme concerns the extent to which it is effective in producing behaviour change in children presenting with clinically significant conduct problems. For example, it could be suggested the programme is more effective for children with mild or sub-clinical problems than it is for children with severe problems, or vice versa. To examine this issue, the relationship between the child's clinical status at baseline and the outcomes at follow-up were compared. Children enrolled in the study were classified into 2 groups:

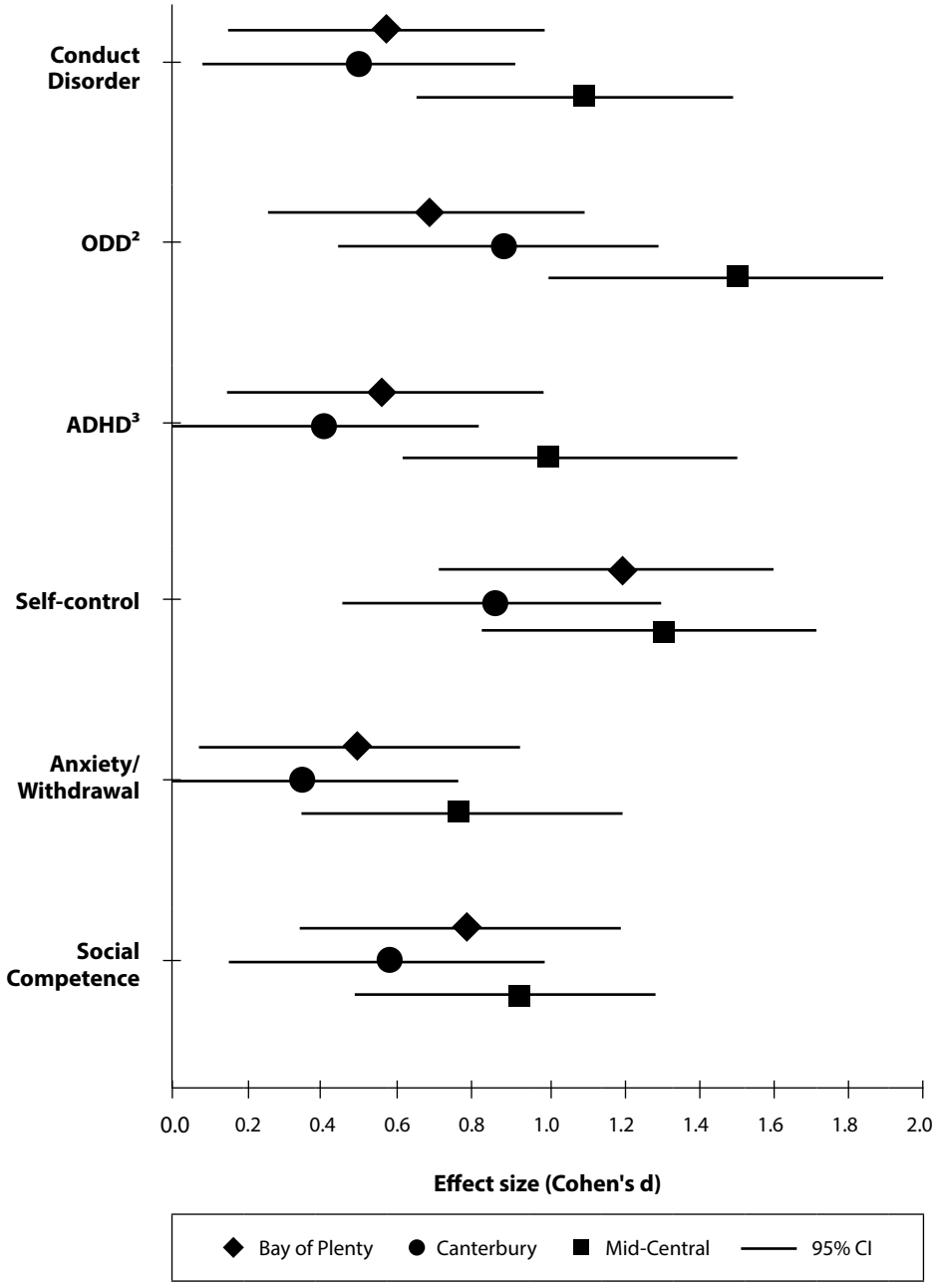
i) those in the clinical range of the Eyberg Child Behaviour Inventory (ECBI) at baseline

ii) those not in the clinical range.

Figure 4 compares the effect sizes on measures of child behaviour for those in the clinical group (N=93) and those in the sub-clinical group (N=33). This shows the programme had greater efficacy for the clinical group in terms of Conduct Disorder ($p=.004$), Oppositional Defiant Disorder ($p=.004$) and Attention Deficit Hyperactivity Disorder ($p=.005$). These findings clearly suggest that, while the IYP programme was effective for both groups, it had greater efficacy for the clinical group. This was confirmed by a multivariate statistical test on all behaviour measures overall ($F(6,119)=4.37$, $p<.001$).

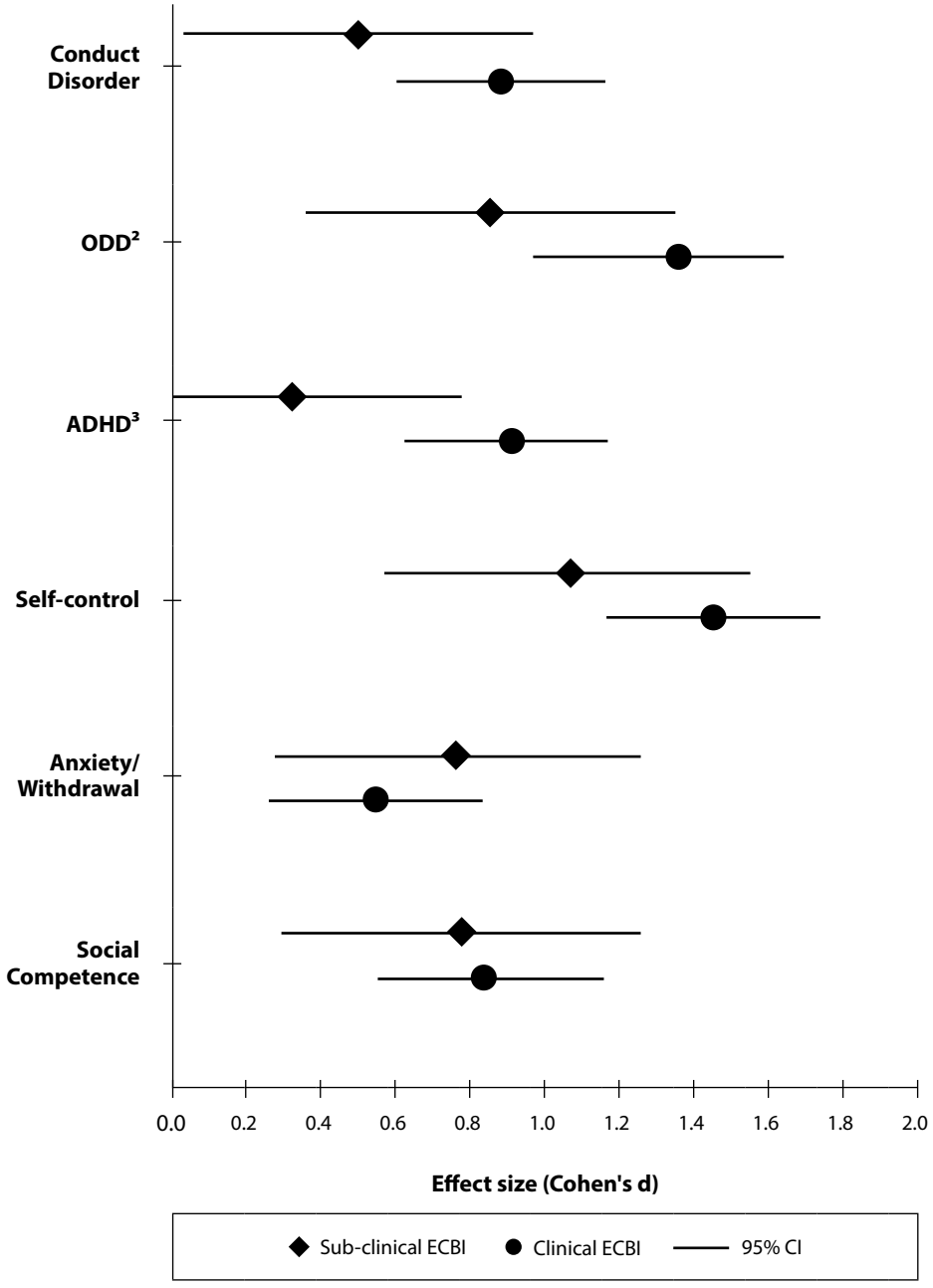
On the basis of these findings, it appears the IYP programme is more suitable for children with clinical levels of conduct problems, suggesting that shorter and less expensive programmes could be used for children with sub-clinical conduct problems. In situations where service provision is limited and where lower intensity courses are available, the greatest return could be obtained by targeting the programme to children whose ECBI scores place them in the clinical range.

Figure 3: Comparison of effect sizes for child behaviour outcomes at the 30-month follow-up, by research site



[1] Sample size for Bay of Plenty is N=44; for Canterbury N=45; for Mid-Central N=47.
[2] Oppositional Defiant Disorder.
[3] Attention Deficit Hyperactivity Disorder.
[4] The 95% confidence interval lines (95% CI) around the estimated effect size represent the range of effect sizes one would expect to see 95% of the time if the experiment was repeated. Any differences between the effect sizes for the research sites are determined by the p-values from the analysis of site interactions.

Figure 4: Comparison of effect sizes for child behaviour outcomes at the 30-month follow-up for sub-clinical and clinical conduct problems[1] classification at baseline



[1] Clinical Conduct Problem is defined by the pre-course ECBI Intensity Score being 60 or greater.
[2] Oppositional Defiant Disorder.
[3] Attention Deficit Hyperactivity Disorder.
[4] Sample size for sub-clinical ECBI children is N=33, and for clinical ECBI children is N=93.
[5] The 95% confidence interval lines (95% CI) around the estimated effect size represent the range of effect sizes one would expect to see 95% of the time if the experiment was repeated. Any differences between the effect sizes for the sub-clinical and clinical groups are determined by the p-values from the analysis of interactions.

Parent satisfaction

In the baseline interview parents were asked to rate their satisfaction with the way the IYP programme was organised using a 3-point scale: not at all, somewhat and very. These same questions were repeated following IYP course completion. In addition, parents were asked to rate their agreement with 13 statements about the IYP course overall after their completion of the 18-week course. These items were based on the Incredible Years Parent Satisfaction Questionnaire, but the scoring was altered from a 7-point to a 3-point scale: not at all, somewhat and a great deal.

Parents rated the same list of items assessing satisfaction with the IYP programme in all five Family Interviews (baseline, mid-course, post-course, 6-month follow-up and 30-month follow-up). They reported high to moderate satisfaction with the IYP programme at the 6-month and 30-month follow-ups. Satisfaction levels at the long-term follow-up remained quite high overall. However, there was a trend for Mid-Central parents to express greater satisfaction with the course and their child's progress than parents in the other two sites. This finding aligns with the relatively greater improvement in behaviour reported by Mid-Central parents compared with Bay of Plenty and Canterbury parents.

In response to the open-ended questions in the long-term follow-up interview, nearly 9 in 10 parents stated that the IYP programme had made a difference for them, their children and their families. Specifically, parents mentioned learning strategies; family being calmer, happier and less stressed; having more confidence as a parent; better understanding of the child; and better communication with child and family. About half (53%) of the parents mentioned things that would help them maintain the IYP strategies, such as referring to the book, having a refresher, catching up with the Group Leader or other parents, and social media to keep parents informed.

Conclusions

This study provides evidence of the maintenance of improvement in child and parent behaviour for New Zealand families, both Māori and non-Māori, receiving an IYP programme. The findings demonstrate that the programme benefits of IYP were sustained for several years with no evidence of any diminution in the size of effect.

The research results provide strong reassurances about the efficacy of the IYP programme over the long term. Specifically, the findings show that:

1) Compared to the baseline values, the findings at post-course showed clear and significant benefits in the areas of child behaviour, parenting, and family relationships. With a small number of exceptions, effect sizes were in the moderate to large range ($d>.50$; $p<.001$).
2) Compared to the baseline values, the findings at the 30-month follow-up showed continued clear and significant benefits in the areas of child behaviour, parenting, and family relationships. As with the baseline/post-course comparisons, most effect sizes were in the moderate to large range ($d>.50$; $p<.001$).
3) 'Positive parenting' and 'Poor supervision' were the only measures where significant differences were found between post-course and follow-up assessments.

These conclusions suggest IYP was an effective programme that showed benefits that persisted in the long term. These findings are generally consistent with the findings of the small number of previous studies that have examined the long-term (1–3 years post-course) benefits of the IYP programme [21, 26, 27, 28, 29].

The analysis addressed three additional issues.

First, an examination was conducted of the extent to which the programme benefits at 30 months were similar for Māori and non-Māori families. Earlier findings suggested child behaviour outcomes for non-Māori were slightly more positive than for Māori. However, the 30-month follow-up showed no differences between Māori and non-Māori on measures of child behaviour, parenting, or family relationships. These findings suggest the IYP programme will be equally effective for Māori and non-Māori. Continuing the use of cultural enhancements [43] to the IYP programme should help with Māori uptake and acceptance of the programme.

Second, the previous report noted some large between-site variations in the efficacy of the IYP programme, with the effect sizes for the Mid-Central site being substantially greater than for the Canterbury and Bay of Plenty sites. These differences were also evident at the 30-month follow-up. The reasons for these differences are not clear, but they could reflect between-site differences in both client recruitment and service provision. These findings raise important questions about the need to monitor the IYP programme at all sites to determine the extent of and reason for between-site differences. Such monitoring will help all IYP providers benefit from examples of best practice.

Third, consideration was given to the extent to which the programme outcomes varied with the severity of the child's behaviour problems at baseline. This analysis showed children whose ECBI scores placed them in the clinical range tended to gain greater benefit than those whose ECBI scores placed them in the sub-clinical range. These findings suggest that, in situations in which the availability of the IYP programme is limited, the best strategy is to focus on the needs of families whose child's score falls within the clinical range of the EBCI. Equally, it is clear children in the sub-clinical range also gained substantial benefits and, under ideal circumstances, the IYP programme should be provided to this group.

These results demonstrate that the pilot implementation of the IYP programme has been highly successful and the programme has been demonstrated to show long-term benefits across a wide range of outcomes.

The evaluation of the IYP programme has been a careful and lengthy process spanning the Pilot Study and the Follow-up Study. It involved longitudinal study of children and families enrolled in the programme, supplemented by single case studies and other forms of data collection. While this data collection process is time-consuming, it does ensure the process of programme implementation is carefully and thoroughly evaluated. It is notable that a recent review of parenting programmes [44] concluded that the IYP programme was one of only two New Zealand-based parenting programmes that have been adequately evaluated. The research model and data collection skills acquired over the course of the evaluation of the IYP programme can be readily transferred to the evaluation of other programmes targeted at providing services to children and families in need of support, mentorship and assistance.

Appendix 1: Source instruments

Child behaviour

In the Family Interview, parents were questioned about their child's behaviour during the past 4 weeks using 111 items based on four recognised scales: the Eyberg Child Behaviour Inventory, the Strengths and Difficulties Questionnaire, the Incredible Years Social Competence Scale and items from the 5-year evaluation of Early Start. For consistency, all scale items were rated on a 3-point rating scale: not at all, somewhat and a great deal.

The Eyberg Child Behaviour Inventory (ECBI) [30, 31] is a 36-item inventory measuring child problem behaviours perceived by the caregiver, and is normed for children aged 2–16 years. The ECBI measures the number of problem behaviours and the frequency with which these behaviours occur. The scale demonstrates good stability, with reliability coefficients from 0.86 (test-retest) to 0.98 (internal consistency) [45]. Good convergent validity is demonstrated by significant correlations with the Child Behaviour Checklist and the Parenting Stress Index [reported in 17]. The ECBI is a well-respected and well used measure for assessing the frequency of conduct problem behaviour that is reliable and valid, and identifies change due to intervention over time [46]. It has been used extensively within the field of parent training intervention including in several studies of the IYP programme [15, 16]. The ECBI has two scales, the Intensity scale and the Problem scale, but the latter was not used in the Family Interview Questionnaire.

The Strengths and Difficulties Questionnaire (SDQ) [32] is a 25-item inventory designed as a behavioural screening measure to assess the occurrence of particular behaviours associated with conduct problems, hyperactivity, emotional symptoms and peer problems in children. The scale has demonstrated good stability as judged by internal consistency (mean Cronbach's alpha=0.73), cross-informant correlation (mean=0.34) and test-retest stability after 4–6 months (mean=0.62) [47].

There are versions for parents and teachers. Both versions contain five subscales: emotional problems, conduct problems, hyperactivity, peer problems and pro-social behaviour. The additional Impact Supplement scale, which measures how the caregiver or teacher perceives the child's behaviour and the impact of the problem on the child's daily life, was not used in the Incredible Years Pilot study. The SDQ is a screening measure and is not as sensitive to clinical change as the ECBI.

The Social Competence Scale – Parent Version (P-COMP), developed by the Conduct Problem Prevention Research Group (Fast Track), consists of 12 items that assess the child's positive social behaviours as perceived by the parent. It includes measures of frustration, tolerance and communication skills. This instrument is also used by the Incredible Years Group Leaders to assess the participating parents' children [33, 48].

Some of the child behaviour items were based on those used in the Early Start evaluation [34]. These items were tested in a New Zealand context and found to provide robust measures of child behaviour.

Confirmatory factor analysis of this item set [35] showed these items measured six correlated dimensions of child behaviour.[1] These dimensions were:

- Conduct Disorder: This dimension was based on a sum of 18 items describing the extent to which the child displayed aggressive and antisocial behaviours. The reliability of the scale assessed by Cronbach's alpha (α) was .87.
- Defiance: This dimension was based on a sum of 15 items describing the extent to which the child showed oppositional, defiant or dishonest behaviours (α=.89).
- ADHD: This dimension was based on a sum of 16 items describing the extent to which the child showed hyperactive, impulsive or inattentive behaviours (α=.92).
- Self-control: This dimension was based on a sum of 15 items describing the extent to which the child showed self-regulatory, flexible or compliant behaviours (α=.87).
- Anxiety/Withdrawal: This dimension was based on a sum of 16 items describing the extent to which the child showed anxious, withdrawn or shy behaviours (α=.78).
- Social Competence: This dimension was based on a sum of 26 items describing the extent to which the child showed helpful, empathetic, respectful, diligent or likeable behaviours (α=.91).

Parenting practices

The parenting practices measures in the Family Interview consisted of two recognised instruments: the Alabama Parenting Questionnaire and the Arnold-O'Leary Parenting Scale. Intact instruments were used, but the scale items were all rated on a consistent 3-point scale: never, sometimes and often for the Alabama; never, less than once a month and once a week or more for the Arnold-O'Leary.

The Alabama Parenting Questionnaire [36] is a 42-item scale designed to tap the parenting dimensions that are risk factors associated with child Conduct Disorder. It loads onto five subscales: parental supervision, positive parenting, corporal punishment, parental involvement and inconsistent discipline (α=.44 to .79, median α=.53). An evaluation with a community sample of Australian children aged 4–9 years [49] showed good internal consistency, validity and test-retest reliability for the measure.

The Arnold-O'Leary Parenting Scale [37] is a 30-item inventory of parenting competencies that measures dysfunctional and/or ineffective parenting practices of parents with younger children. The scale yields an overall score and three revised subscale scores of dysfunctional strategies used by parents tackling problem behaviour (α=.42 to .74, median α=.65). 'Laxness' refers to insufficient monitoring of the child and the child's behaviour, allowing rules to go unenforced or providing positive reinforcement for misbehaviour. 'Over-reactivity' refers to displays of anger, meanness or irritability. 'Hostility' refers to the use of verbal or physical force. The scale has adequate internal consistency and has been found to have good test-retest reliability.

[1] Five of the 111 items were discarded as they did not belong obviously to any of the subscales.

Relationships

The relationships between the caregiver and their child and partner were measured using three instruments. Approval was granted to adapt the Conflict Tactics Scales CTSPC and CTS2 with the payment of copyright fees. The Conflict Tactics Scales [38, 39] have been used for decades to evaluate violence within families and intimate relationships. Two updated versions, the Conflict Tactics Scale Parent/Child (CTSPC) and the Revised Conflict Tactics Scale 2 (CTS2), were used in the Incredible Years Pilot Study. The CTSPC items related to the parents' use of verbal aggression and physical assault in their relationships with their children. Scale items were all rated on a consistent 3-point scale: never, less than once a week, once a week or more. The CTS2 focused on violence between the parents. Scale items were all rated on a consistent 3-point scale: never, sometimes, often.

The Parent Problem Checklist (PPC) was developed as a measure of inter-parental conflict, especially as it relates to the parents' ability to co-operate and to act as a team in performing the executive parenting functions within the family. It contains 16 items measuring the presence or absence of parental disagreement over rules and discipline for child misbehaviour, the occurrence of open conflict over child-rearing issues and whether or not parents undermine each other's relationships with the children. The PPC is a unidimensional measure with moderately high internal consistency and high test-retest reliability [40]. Scale items were all rated on a consistent 3-point scale: not at all, somewhat, a great deal.

The Partner Attachment Scale measures the quality of the relationship between parents. The items used in the study are based on a selected series of items from Braiker and Kelley [41] as used in the Christchurch Health and Development Study (CHDS) 21-Year Interview. Scale items were all rated on a consistent 3-point scale: doesn't apply, somewhat applies, definitely applies.

Appendix 2: Statistical methods

Sustainability of programme outcomes

The sustainability of the IYP programme outcomes was determined by looking for a linear trend of change in the mean scores from the baseline (time #1) through the post-course (time #2) to the follow-up (time #3). The null hypothesis that there was no linear trend of change in the mean scores over time was tested with an Analysis of Variance (ANOVA) for repeated measures. If the p-value for 'time' was small (generally where p<.05) the null hypothesis was rejected, which suggests the programme was effective in changing behaviours.

The general algebraic form of the general linear model can be expressed as below.

$$Y_{ij} = \beta_0 + \beta_1 Time_j + \mu_{ij}$$

Where:

Y = outcome (Conduct Disorder, etc);

i = observations 1 to N; and

j = observations at time points #1 (baseline), #2 (post-course), and #3 (follow-up)

In addition to testing for an overall linear trend of change, the time points were contrasted against each other. First, the post-course mean scores were tested against the baseline mean scores to confirm there were significant differences (real outcomes). Second, the post-course mean scores were tested against the follow-up scores to test the scores had not changed significantly (outcomes were sustained). Third, in the case where outcomes were not sustained, the follow-up mean scores were tested against the baseline scores to test whether the follow-up scores were an improvement over the baseline or not. These tests were represented by the superscripted letters 'a', 'b', and 'c' above each mean score, where letters that are different present mean scores that are different (p<.05). The following table describes how to interpret the superscripted letters.

Mean score superscript	Compared to superscript	Interpretation of mean score comparisons
a	b	mean score 'a' is different from the mean score 'b' (p<.05)
a	c	mean score 'a' is different from 'c' (p<.05)
b	c	mean score 'b' is different from 'c' (p<.05)
b	b	mean score 'b' is the same as 'b' (p>.05)
a	a,b	mean score 'a' is the same as 'a,b' (p<.05) but is different to 'b' (p>.05)
b	a,b	mean score 'b' is the same as 'a,b' (p<.05) but is different to 'a' (p>.05)

Programme effect sizes

In addition to the repeated measures test, the analysis includes effect size estimates to examine the size of the change in behaviours. The effect size calculations used are Cohen's d (expressed as 'd'). Cohen's d is the standardised difference between means of proportions [42]. Cohen suggests an effect size of d=.20 is small, an effect size of d=.50 is medium, and an effect size of d=.80 is large. These interpretations are arbitrary but provide an indication of how large the behaviour change is. A positive (d>0) effect size indicates improved behaviours while a negative (-) effect size (d<0) indicates worsened behaviours.

The effect size for each measure is calculated as the difference between the baseline (time #1) and the comparison mean scores (at times #2 or #3) divided by the standard deviation of the scores at baseline. Because the baseline standard deviation is used, the estimates of effect size will tend to be conservative (smaller) than if the standard deviation of scores at the comparison times were taken into consideration.

Effect size between baseline and post-course results:

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s_1}$$

Effect size between baseline and follow-up results:

$$d = \frac{\overline{x}_1 - \overline{x}_3}{s_1}$$

Where:

$\overline{x}_1$	Mean score at time #1 (baseline)
$\overline{x}_2$	Mean score at time #2 (post-course)
$\overline{x}_3$	Mean score at time #3 (follow-up)
s_1	Standard deviation of scores at time #1 (baseline)

Confidence intervals on the effect sizes

Because the effect sizes are estimates based on samples, 95% confidence intervals (CI) are calculated to provide a range for the estimated effect sizes. The confidence intervals are calculated using a SAS® Macro developed by Hess and Kromrey [50], which calculates the confidence intervals based on three inputs: (1) the effect size, (2) the sample size of the effect size calculation at the comparison interview (treatment), and (3) the sample size of the effect size calculation at baseline (control). The comparison sample sizes used in the calculations are the post-course or follow-up samples instead of the baseline or control sample size. This means the calculated confidence limits will tend to be conservative (slightly wide).

The calculated confidence intervals will be fairly wide for small sample sizes. In some cases of small sample sizes and small effect sizes, the lower confidence limit may be less than zero. The p-values for the tests of a linear trend should be used as evidence of change and not the effect size confidence intervals.

Sustainability of programme outcomes between groups

The programme outcomes between group categories or levels were demonstrated by the effect size charts. Any differences between the group levels were tested by analysing the interaction between the levels and the change in the outcome scores as determined by an Analysis of Variance (ANOVA). Evidence of a significant interaction between the group levels and the change in outcome scores is demonstrated where p<.05. The absence of a significant interaction was taken to imply the relative change was broadly similar across the sub-groups being compared.

In the example of the Māori and non-Māori ethnicity groups, the algebraic form of the general linear model can be expressed as below. The null hypothesis is that the programme outcomes were broadly similar for Māori and non-Māori, and is tested by the p-value for the ethnicity by time interaction.

$$Y_{ij} = \beta_0 + \beta_1 Time_j + \beta_2 Ethnicity_i + \beta_3 (Ethnicity.Time)_{ij} + \mu_{ij}$$

Where:

Y = outcome (Conduct Problem, etc);

i = observations 1 to N; and

j = observations at time points #1 and #2

Corrections to multiple hypothesis testing

The outcomes reported involved multiple hypotheses testing, which requires some consideration of the possibility of low p-values occurring by chance. The following two methods were employed to reduce the chance of false-positive conclusions from hypothesis testing.

First, a Bonferroni adjustment to the α=.05 is applied to the overall programme and sustained outcome tables (below). It was thought a Bonferroni adjustment was sufficient for the overall outcome tables because of their strong findings. The following table shows the Bonferroni adjustments made to the α=.05 level for each of the programme outcome and sustained outcome tables.

Second, interactions between group levels were summarised with a multivariate analysis of variance that tests all of the outcome measures as one outcome. The null hypothesis is that there is no interaction for the multivariate or overall outcome between time and the group levels.

In the example of the Māori and non-Māori ethnicity groups, the general algebraic form of the general linear model is expressed below, where the null hypothesis is that ${\beta_3}^k = 0$ for all k=1 to K where each k is an individual measure such as Conduct Disorder, Oppositional Defiant Disorder, etc.

$${Y_{ij}}^k = {\beta_0}^k + {\beta_1}^k Time_j + {\beta_2}^k Ethnicity_i + {\beta_3}^k (Ethnicity.Time)_{ij} + {\mu_{ij}}^k$$

Where:

Y = outcome (Child Behaviour, etc);

i = observations 1 to N;

j = observations in time points #1 and #2;

k = outcome k (k=1 to K, Conduct Disorder, Oppositional Defiant Disorder, etc)

Number of participants (sample sizes) included for each measure

Tables	Bonferroni adjustment to α=.05	Number of tests
Child behaviours	p-values <.008 are significant	6
Parenting practices	p-values <.005 are significant	9
Conflict tactics/Relationships	p-values <.006 are significant	8

References

1. Fergusson, D M, Boden, J and Hayne, H in P Gluckman (2011) *Improving the Transition*. Wellington
2. Scott, S (2007) Conduct disorders in children. *British Medical Journal*, March, 334:646
3. Ministry of Social Development (2009) *Conduct problems: Best practice report*. Wellington: Ministry of Social Development
4. Lindsay, G, Strand, S, Cullen, M A, Cullen, S, Band, S, Davis, H, Conlon, G, Barlow, J and Evans, R (2011) *Parenting Early Intervention Programme Evaluation*: DFE-RR121A. London: DfE
5. Fergusson, D M, Horwood, L J and Ridder, E (2005) Show me the child at seven: The consequences of conduct problems in childhood for psychosocial functioning in adulthood. *Journal of Child Psychology and Psychiatry*, 2005, 46(8):837–849
6. Feinstein, L and Sabates, R (2006) Predicting adult life outcomes from earlier signals: Identifying those at risk. Centre for Research on the Wider Benefits of Learning, Institute of Education, University of London
7. Odgers, C L, Caspi, A, Broadbent, J M, Dickson, N, Hancox, R J and Harrington, H (2007) Prediction of differential adult health burden by conduct problem subtypes in males. *Archives of General Psychiatry*, 2007, 64:476–484
8. Karoly, L A, Kilburn, M R and Cannon, J S (2005) Early childhood interventions: Proven results, future promise. Santa Monica: RAND Corporation. Retrieved 7 May 2008 from http://www.rand.org/pubs/monographs/2005/RAND _MG341.pdf
9. Belsky, J, Vandell, D L, Burchinal, M, Clarke-Stewart, K S, McCartney, K and Owe, M T (2007) Are there Long-Term Effects of Early Child Care? *Child Development*, March/April, Vol 78, No. 2, 681–701
10. Gluckman, P (2011) Improving the Transition. Wellington
11. Lindsay, G, Strand, S, Cullen, M A, Cullen, S, Band, S, Davis, H, Conlon, G, Barlow, J and Evans, R (2011) *Parenting Early Intervention Programme Evaluation*: DFE-RR121A. London: DfE
12. Webster-Stratton, C and Reid, M J (2010) The Incredible Years Parents, Teachers and Children Training Series: A Multifaceted Treatment Approach for Young Children with Conduct Disorders. In J Weisz and A Kazdin (Eds) *Evidence-based Psychotherapies for Children and Adolescents*, 2nd edition. New York: Guilford Publications
13. Webster-Stratton, C and Hammond, M (1997) Treating children with early-onset conduct problems: a comparison of child and parent training interventions. *Journal of Consulting and Clinical Psychology*, 65(1), 93–109. Supplementary analysis retrieved December 2012 from http://www.incredibleyears.com/library
14. Gardner, F, Burton, J and Klimes, I (2006) Randomised Controlled Trials of a Parenting Intervention in the Voluntary Sector for Reducing Problems in children: Outcomes and Mechanisms of Change. *Journal of Child Psychology and Psychiatry*, 47:11, 1123–1132
15. Scott, S, Knapp, M, Henderson, J and Maughan, B (2001) Multicentre controlled trial of parenting groups for childhood antisocial behaviour in clinical practice. *British Medical Journal,* 323, 194–203
16. Hutchings, J, Bywater, T, Daley, D, Gardner, F, Whitaker, C and Jones, K, et al (2007) Parenting intervention in Sure Start services for children at risk of developing conduct disorder: Pragmatic randomised controlled trial. *British Medical Journal*, 334(7595), 1–7
17. Jones, K, Daley, J, Hutchings, T, Bywater, T and Eames, C (2007) Efficacy of the Incredible Years Basic parent training programme as an early intervention for children with conduct problems and ADHD. *Child: Care, Health and Development*, 33, (6) 749–756
18. O'Neill, D (2009) A Cost-Benefit Analysis of Early Childhood Intervention: Evidence from a Randomised Evaluation of a Parenting Programme. IZA Discussion Paper No. 4518, October

19. Larsson, B, Fossum, S, Clifford, G, Drugli, M B, Handegard, B H and Morch, W-T (2009) Treatment of oppositional defiant and conduct problems in young Norwegian children: results of a randomized controlled trial. *European Child and Adolescent Psychiatry*, 18(1):42–52

20. Kelleher, C and McGilloway, S (2006) The Incredible Years Basic Parenting Programme: The Clondalkin Partnership Study in Ireland. Retrieved December 2012 from http://www.incredibleyears.com/library

21. Posthumus, J A, Raaijmakers, M a J, Maassen, G H, van Engeland, H and Matthys, W (2012) Sustained effects of Incredible Years as a Preventive Intervention in Preschool Children with Conduct Problems. *Journal of Abnormal Child Psychology*, 40(4), 487–500

22. Webster-Stratton, C (1994) Advancing videotape parent training: A comparison study. *Journal of Consulting and Clinical Psychology,* 62(3), 583–593

23. RAND (2007) Promising Practices Network: Programmes that Work: Incredible Years. Retrieved 25 June 2012 from http://www.promisingpractices.net/program.asp?programid=134

24. Ministry of Social Development (2007) *Interagency plan for Conduct Disorder/Severe Antisocial behaviour 2007–2012*. Wellington: Ministry of Social Development

25. Sturrock, F and Gray, D, (2013) Incredible Years Pilot Study Evaluation Report. Wellington: Ministry of Social Development

26. McGilloway, S, Mháille, G N, Furlong, M, Hyland, L, Leckey, Y, Kelly, P, Bywater, T, Comiskey, C, Lodge, A, O'Neill, D and Donnelly, M (2012) The Incredible Years Ireland Study: Parents, Teachers and Early Childhood Intervention. Long-term Outcomes of the Incredible Years Parent and Teacher Classroom Management Training Programmes. Report prepared for Archway, Northern Ireland

27. Jones, K, Daley, J, Hutchings, T, Bywater, T and Eames, C (2007) Efficacy of the Incredible Programme as an early intervention for children with conduct problems and ADHD: long-term follow-up. *Child: Care, Health and Development*, 34, (3) 380–390

28. Reid, M J, Webster-Stratton, C and Hammond, M (2003) Follow-up of children who received the Incredible Years Intervention for Oppositional-Defiant Disorder: Maintenance and prediction of 2-year outcomes. *Behaviour Therapy,* 34, 471–491

29. Webster-Stratton, C (1990) Long-term Follow-up of families with young conduct problem children: From preschool to grade school. *Journal of Clinical Child Psychology*, 19, 144–149

30. Eyberg, S M and Ross, A W (1978) Assessment of child behaviour problems: The validation of a new inventory. *Journal of Clinical Child & Adolescent Psychology*, 7(2), 113–116

31. Eyberg, S M (1980) Eyberg Child Behaviour Inventory. *Journal of Clinical Psychology* 9, 27

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

33. Corrigan, A (2002) Social Competence Scale – Parent Version, Grade 1/Year 2. (Fast Track Project Technical Report). Available from the Fast Track Project Web site, http://www.fasttrackproject.org

34. Fergusson, D M, Horwood, L J, Ridder, E M and Grant, H (2005) *Early Start:* Evaluation report. Retrieved 7 May 2008 from http://www.earlystart.co.nz/pdf/evalreport.pdf

35. Horwood, L J, Gray, D S and Fergusson, D M (2011) The Psychometric Properties of the Child Behaviour Rating Scales used in the Incredible Years Pilot Study (unpublished)

36. Shelton, K K, Frick, P J and Wooton, J (1996) Assessment of parenting practices in families of elementary school-age children. *Journal of Clinical Child Psychology*, 25, 317–329

37. Arnold, P S, O'Leary, S G, Wolff, L S and Acker, M M (1993) The Parenting Scale: A measure of dysfunctional parenting in discipline situations. *Psychological Assessment* 5 (2), 137–144

38. Straus, M, Hamby, S, Boney-McCoy, S and Sugarman, D (1996) The revised Conflict Tactics Scales (CTS2): Development and preliminary psychometric data. *Journal of Family Issues*, 1998, 17(3), 283–316

39. Straus, M A, Hamby, S L, Finkelhor, D, Moore, D W and Runyan, D (1998) Identification of child maltreatment with Parent-Child Conflict Tactics Scales: Development and psychometric data for a national sample of American parents. *Child Abuse and Neglect,* 22(4), 249–270

40. Dadds, M and Powell, M (1991) The relationship of interparental conflict and global marital adjustment to aggression, anxiety and immaturity in aggressive and non-clinic children. *Journal of Abnormal Child Psychology*, 19, 553–561

41. Braiker, H and Kelley, H H (1979) Conflict in the development of close relationships. In R L Burgess and T L Huston (Eds) Social exchange in developing relationships. New York: Academic

42. Cohen, J (1977) *Statistical Power Analysis for the Social Sciences.* New York: Academic Press

43. Werry Centre (2012) *Ngā Tau Mīharo ō Aotearoa*. Auckland: Werry Centre

44. Robertson, J (2014) Effective parenting programmes: A review of the effectiveness of parenting programmes for parents of vulnerable children. Wellington: Families Commission

45. Robinson, E A, Eyberg, S M and Ross, A W (1980) The standardisation of an inventory of child conduct problem behaviours. *Journal of Clinical Child Psychology*, 57, 628–635

46. Rhodes, H (2009). Briefing sheet: Examples of Effective Measuring Tools. Family and Parenting Institute

47. Goodman, R (2001) Psychometric properties of the Strengths and Difficulties Questionnaire (SDQ). *Journal of the American Academy of Child and Adolescent Psychiatry*, 40, 1337–1345

48. Conduct Problems Prevention Research Group (CPPRG) (1995) Psychometric Properties of the Social Competence Scale – Teacher and Parent Ratings. (Fast Track Project Technical Report). University Park, PA : Pennsylvania State University

49. Dadds, M, Maujean, A and Fraser, J A (2003) Parenting and Conduct Problems in Children: Australian Data and Psychometric Properties of the Alabama Parenting Questionnaire. *Australian Psychologist*, Vol 38, No. 3, 238–241

50. Hess, M R and Kromrey, J D (2003) EFFECT_CI: A SAS® Macro for Constructing Confidence Intervals Around Standardized Mean Differences. University of South Florida. Retrieved from http://analytics.ncsu.edu/92 sesug/2003/SD04-Hess.pdf

TE TĀHUHU O TE MĀTAURANGA
Ministry of Education

MINISTRY OF SOCIAL DEVELOPMENT
TE MANATŪ WHAKAHIATO ORA