UNIVERSITY OF CAMBRIDGE
INSTITUTE OF CRIMINOLOGY

Working Paper, July 2007

What Evidence for Prevention?

A Critical Re-examination of the Prevention Study „Empowerment in Family and School (eifas): A Randomized Trial“

Dr Manuel Eisner
Reader in Quantitative Criminology
Institute of Criminology
Sedgwick Site
University of Cambridge
CB3 9DT Cambridge
Tel: 0044 1223 335374
Email: manuel.eisner@crim.cam.ac.uk

Abstract

The ESSKI (Eltern und Schule stärken Kinder) study is a large randomized controlled trial conducted in Switzerland with the goal of preventing problem behaviour and promoting children's development. In entails the implementation of *Triple P* as a parenting programme and of *Fit for Life* as a school-based life-skills programme. A recently published research report concluded that the interventions had been effective in reducing aggression and hyperactivity, in promoting subjective health and in reducing smoking.

In this paper I re-analyse the data published in the project reports. Findings suggest that there are no systematic positive effects of the interventions on core target variables. The conclusion argues for the adoption of better standards for publicly presenting the results of experimental field studies.

Acknowledgments:

I would like to thank Zsolt Lavicza, Denis Ribeaud, Laszlo Erdodi, Friedrich Lösel and Ursula Meidert for comments on earlier drafts of this paper.

Introduction

Programmes aiming at the early prevention of behavioural problems, aggression and substance abuse have become increasingly popular in Switzerland over the past decade. In particular, there is growing interest in interventions that promote parenting skills at the family level as well as programmes that enhance children's social and cognitive skills within school settings. Increasingly such programmes claim scientific evidence for their effectiveness, preferably based on a randomized controlled trial, the "gold standard" of evidence-based prevention. However, the wider public, the media, politicians or administrators can rarely assess the validity of such claims. Especially when communicating outside the scientific community it is therefore crucial to report the findings of experimental studies as accurately and impartially as possible.

In this paper I will critically reanalyze a large randomized trial conducted recently in Switzerland with the goal of preventing problem behaviour and promoting children's development, known as the ESSKI-Study ("Eltern und Schule stärken Kinder"; in English eifas-study for „empowerment in family and school"). The study was conducted by a team of researchers at the University *of Applied Sciences Northwestern Switzerland*, the *University of Fribourg*, the *Zurich University of Teacher Education*, and the *Swiss Institute for the Prevention of Alcoholism and other Addictions*. A scientific report on the study was publicly presented on 18 January 2007 und concluded that the interventions had been effective in reducing aggression and hyperactivity, in promoting subjective health and in reducing smoking.

I show subsequently that these conclusions are not supported by the data reported in the study. Rather, a reanalysis suggests that there were no systematic positive effects of the interventions on the core target variables. The conclusion argues for the adoption of general standards in presenting the results of experimental field studies.

The ESSKI-Study

The design and goals of the ESSKI-Study are described in detail in Schönenberger et al. (2006; 2006). Essentially the study is a randomized field trial that examines the effectiveness of early prevention programmes and that "aims at promoting the health of students, teachers, and parents and at preventing substance abuse, violence, and stress". At baseline the study comprised N = 84 school classes and 1423 children at ages 6 to 12. Classes were randomly allocated to one of four treatment conditions resulting from the factorial combination of two interventions. At the family level, the self-help version of Triple P was implemented in two treatment conditions. Triple P is a parent training programme based on behavioral principles and developed by Matthew Sanders (for an overview see Sanders, Markie-Dadds, & Turner, 2003). The implementation included various self-help materials and was accompanied by telephone support delivered by trained Triple P providers. At the school level the „fit for life – preventing aggression, distress and addiction by improving personality development" programme was implemented. "Fit for life" is a life skills programme developed by Burow, Asshauer and Hanewinkel (1998; 1999). In the ESSKI study it was implemented over a period of 12 weeks.

The main conclusions of the ESSKI study

On 18 January 2007 a media release on the study findings was published in German, French and Italian. Both the media release and the research report can be downloaded from the website of the ESSKI project (http://www.esski.ch/downloads). The media release states: „Children and juveniles who have participated in the ESSKI study are less aggressive and hyperactive, they feel healthier and they smoke less. […] Students in the intervention groups are less aggressive, short-tempered and hyperactive, that they are less likely to feel unhappy and sad, and they report fewer physical conditions. [These results show] "that strategies of health promotion and the prevention of addiction, which start at primary school and involve all actors, i.e. also teachers and parents, are effective." (Eltern und Schule stärken Kinder, 2007, my translation, M.E.). Similar claims are made in the more extensive 50-page research report. For example, the study summary states that "the analyses of the ESSKI study show that the strengths of students have increase both according to teacher and parent assessments, and that weaknesses have decreased" (Schönenberger, Schmid et al., 2006: 2, my translation).

Three reasons led the author to re-exame these findings. First, the highly positive findings of the ESSKI study are surprising when compared to results of similar studies internationally. Unconditionally positive results of universal, general population interventions are extremely rare and are not usually achieved even with considerably more intensive programmes delivered with high implementation fidelity (see, e.g., Lösel & Beelmann, 2003; Wilson & Lipsey, 2006; Wilson, Lipsey, & Derzon, 2003). Secondly, meta-analyses of preventions studies have repeatedly shown that published evaluations yield significantly better results when evaluated by the programme distributors themselves as compared to independent evaluations by third parties (Farrington & Welsh, 2003; Reyno & McGrath, 2006; St Pierre, Osgood, Mincemoyer, Kaltreider, & Kauh, 2006; Wilson et al., 2003). It is unclear how this difference can be explained. However, there is an increased possibility that evaluations conducted by programme promoters are biased towards positive results and should therefore be examined carefully. Thirdly, experimental prevention research in the field of antisocial behaviour has only recently begun in Switzerland (Eisner, Ribeaud, & Bittel, 2006). For this reason there is a lack of established standards for reporting evaluation results and disseminating research findings to the wider public. Scrutiny of new studies may help to identify problematic reporting, which could raise long-lasting public distrust in academic work.

It may be unusual to reexamine study results on the basis of research reports and media releases rather than publications in peer-reviewed academic journals. However, public communication via websites and media releases is becoming increasingly influential, particularly in applied fields of social-science research. In the case of the ESSKI study, for example, the media release was published under the auspices of the well-regarded *Swiss Institute for the Prevention of Alcoholism and other Addictions* to achieve maximum publicity. There can be hardly any doubt that media releases rather than subsequent academic journal publications shape the public and political perception of a prevention study. It therefore is important to examine whether the claims made publicly hold up to closer scrutiny.

Methodological approach

The results section of the ESSKI report displays a series of line charts and interpretations of what the authors believe can be concluded from them. The charts represent means of selected target variables by treatment condition and for pre, post, and follow-up assessments. However, the report fails to clarify exactly how the data were analysed and how the conclusions were arrived at. The only remark on the statistical analysis is a note that "all results were significant" (Schönenberger, Lattmann et al., 2006: 20). Output displayed in the appendix suggests that an analysis of variance was conducted, but no further details are given. I will discuss below why the authors' methodological approach is inadequate for examining intervention effects. At this stage, however, I would only like to highlight two omissions that contradict widely held criteria for the descriptive validity of evaluation reports (Farrington, 2003; Lösel & Koferl, 1989). First, the research report presents findings on a *subset of target dimensions* only. In particular, comparison between the instruments mentioned in the report and scales actually shown revealed that some outcome measures are not discussed in the results section. There is wide agreement, however, that adequate reporting requires full coverage of all relevant target dimensions, since intentional or unintended omissions can misguide readers into erroneous conclusions. Secondly, the report fails to show any data on unstandardized or standardized effect sizes, their confidence intervals or significance levels. However, not least the American Psychological Association recommends in the 2001 Publication Manual that authors routinely report magnitude-of-effect measures in conjunction with significance levels (American Psychological Association, 2001).

In reaction to these shortcomings the subsequent re-analysis is based on three principles. First, I examine those outcomes that the authors describe as their *main target variables*. Secondly, I examine *all* subdimensions that represent the respective outcome rather than limiting myself to a selection of outcome measures. Third, I use generally accepted approaches for estimating effect sizes with published data.

Identifying target outcomes

In their reports the authors of the study emphasize three domains, in which they claim to have achieved positive outcomes (see quotes above). These are: *more strengths* (i.e. more prosocial behaviour) and *fewer difficulties* (i.e. problem behaviours) of students; *increased life satisfaction* and *fewer physical health problems*; and a *lowered risk of smoking*. For each domain the authors distinguish several subdimensions and/or data from different informants. In particular, the scales and subdimensions are:

Strengths and Difficulties Questionnaire

The Strengths and Difficulties Questionnaire is a brief measure of prosocial behavior and psychopathology amongst 3-16-year-olds (Goodman, 2001). It consists of 25 items. In the ESSKI-study it was administered to parents and teachers. While the authors of the questionnaire distinguish five sub-scales with five items each, the authors of the ESSKI study only distinguish a total "strengths" and a total "difficulties" dimension. The difficulties subscale comprises emotional problems, behavior problems and hyperactivity. The strengths subscale is designed to capture prosocial behavior. The authors thus present a total of four outcome measures, namely each subscale according to the parent assessment and according to the teacher assessment.

Children's Quality of Life

This instrument was administered to the children and is based on the Kindl-R Questionnaire by Ravens-Sieberer and Bullinger (Bullinger, von Mackensen, & Kirchberger, 1994; Ravens-Sieberer, 1998). The original instrument comprises six sub-dimensions, namely general health, fun and laughter, self-esteem, good relations with parents, good relations with friends, and having fun at school. Each sub-dimension is assessed with four items. In the reports available to me the authors of the ESSKI-study only report data on the last five dimensions, but not on general health. I have no indication whether this subdimension was not assessed or whether data are not reported.

Smoking and intentions to smoke

This instrument includes three single questions in the child questionnaire, namely whether the child agrees that it will become a smoker in later life, whether it thinks smoking is cool, and whether it currently is a smoker.

I subsequently use all instruments as implemented by the authors, assuming acceptable psychometric properties. Also, I will not discuss issues of validity in any detail. It may be noted, however, that the Strengths and Difficulties Questionnaire only includes two items on physical aggression, namely “often has temper tantrums or hot tempers” and “often fights with other children or bullies them”. It seems questionable to assume that aggression – an outcome repeatedly emphasized in the report and the media release - can be measured validly with only two items.

The appendix of the Schmid et al. (2007) report includes a full documentation of data on all pre, post, and follow-up measures, including those not presented in the published German report.[1] The subsequent analyses will re-examine change between the pre and the follow-up assessments, which were conducted five months after the intervention. No analysis will be done of change between pre and post measures. The reason is, first, that intervention effects are of limited practical significance if they can’t be maintained over at least five months. Secondly, the study authors explicitly argue that they have demonstrated long-term effectiveness of the interventions. Hence the methodological question simply becomes: *Was the change between pre and follow-up assessments significantly better in one of the intervention groups as compared to the control group?*

The reports include all necessary information to compute effect sizes for the *strengths and weaknesses* domain and the children’s *quality of life* domain. The situation is different for smoking and smoking intent. For all respective items the available documents only display data for the post and the follow-up assessments of the children. No data are given for the pre-intervention assessment, meaning that no pre-post comparisons can be conducted. Yet the whole logic of randomized experimental studies is based on comparing change in the experimental and the control conditions *during* the period of the intervention (Shadish, Cook, & Campbell, 2002). In contrast, change between post and follow-up assessments cannot be unambiguously related to an intervention, especially if treatment conditions differ in respect of age-structure and possibly other covariates of the outcomes. The report does not explain why the authors believe that causal effects can be established using post to follow-up comparisons. I therefore decided to exclude this domain from subsequent analyses.

Computing effect sizes

The two research reports present *means*, *Ns*, and *standard deviations* for each target outcome at each point of measurement (i.e. pre, post, and follow-up). This is all the information needed to compute between-group effect sizes, which represent the extent to which the intervention group did better or worse than the control group (Cohen, 1988). Essentially it is a two-step procedure. In a first step, within-group effect sizes are computed as …

$$\text{Cohen's d} = \frac{M_{\Pr e} - M_{Follow-up}}{\sqrt{(SD^2_{\Pr e} + SD^2_{Follow-up})/2}} \qquad \textit{Equation 1}$$

These within-group effect sizes show the standardized gain score, within each treatment condition, between pre and follow-up. In line with common practice I show all standardized effects so that a positive value (> 0) represents an effect in the desired direction. Effect sizes

[1] I am grateful to the study authors for providing me with this unpublished report.

shown in the tables are Hedge's corrected effects, but differences to simple effects are minimal (Hedges & Olkin, 1985).

To compute the standard error of the effect size I used the formula given by Hedges and Olkin …

$$S.E._{Cohen's\,d} = \sqrt{\frac{N_{pre} + N_{follow-up}}{N_{pre} \times N_{follow-up}} + \frac{d^2}{2(N_{pre} + N_{follow-up})}} \qquad \textit{Equation 2}$$

In a second step, the between-group effect size is computed as the difference in standardized gain scores, i.e.

$$\mathrm{ES}_{\text{between groups}} = \mathrm{ES}_{\text{within(Treatment)}} - \mathrm{ES}_{\text{within(control)}} \qquad \textit{Equation 3}$$

This procedure yields the same results as the approach suggested by Lipsey and Wilson (2001: 179) for computing standardized differences in gain scores between treatment and comparison groups. In order to assess the statistical significance of the effect size I examined whether the confidence intervals of the respective within-group effect sizes overlap. Many researchers believe that this can be done by examining the overlap of 95% confidence intervals. However, as Payton et al. (2003) show, non-overlap of 95% confidence intervals corresponds to a significance level far more restrictive than $p < 0.05$. If standard errors of the compared parameters are roughly equal, non-overlap of approximately 83-84% confidence intervals is equivalent to a test of significance at $p < 0.05$ (Payton et al., 2003).[2]

Results

Appendix 1 reports all relevant data and analyses. Means, standard deviations and Ns are those reported in the published reports. To minimize the risk of erroneous calculations of effect sizes and associated confidence intervals all calculations were done independently by two researchers.

Table 1 summarizes the results shown in appendix 1. It displays analyses for each comparison between an intervention group and a control group for each subdimension of the strengths and weaknesses and the quality of life questionnaires.

Column 3 shows Cohen's d, a standardized measure for the size of intervention effects. According to general conventions (Cohen, 1988) a Cohen's d of 0.2 is considered a "small" effect and a value of 0.5 is considered a moderate effect. Column 4 shows whether the control group or the intervention group has improved more, irrespective of statistical significance. Column five shows whether the between-groups effect size can be assumed to be significant at a $p<0.05$ level of confidence.

The table includes a total of 27 comparisons. The mean value of all effect sizes is d = -

[2] Several effect-size calculators are available on the web. I used the excel-spreadsheets provided by the „Curriculum, Evaluation and Management Centre" of the University of Durham to compute effect sizes, associated standard errors, and confidence intervals. (http://www.cemcentre.org/renderpage.asp?linkID=30325017)

0.04 suggesting that, overall, the control group did slightly (but not significantly) better than the average of all treatment conditions. 15 comparisons favor the control group, meaning that there was more change in the desirable direction in the control group as compared to the intervention group. 9 comparisons favored the treatment group. In two comparisons the effect sizes were identical.

However, almost all between-group effect sizes are small and not statistically different from zero. Only two comparisons yield significant results at $p < 0.05$. This is approximately the number of significant effects that can be expected by chance amongst 27 effect size coefficients. One comparison favors the control group and one comparison favors the intervention group.

Table 1 Intervention Effects in the ESSKI-study, Summary

(1) Examined sub-dimension

(2) Comparison with control group

(3) Between-Group Effect size Cohen's d = Within-Group ES_{TG} – Within-Group ES_{KG}. Within-group effect sizes are shown in Appendix 1. Positive values represent better change in the treatment group.

(4) Better change in Control-group or treatment group based on sign of (3).

(5) Significant effect at $p < 0.05$ level, based on confidence intervals of within-group ES.

(1)	(2)	(3) Effect size Cohen's d	(4) Better change in CG or TG		(5) Difference significant at p < 0.05
Strengths, teacher assessment*	Fit for Life	-0.20		CG	NO
	Triple P	-0.11		CG	NO
	Combined	-0.34		CG	YES
Difficulties, teacher assessment	Fit for Life	-0.15		CG	NO
	Triple P	-0.01		CG	NO
	Combined	+0.02	TG		NO
Strengths, parent assessment	Fit for Life	-0.02		CG	NO
	Triple P	-0.02		CG	NO
	Combined	+0.10	TG		NO
Difficulties, parent assessment	Fit for Life	+0.03	TG		NO
	Triple P	+ 0.38	TG		YES
	Combined	+0.16	TG		NO
Q of life: Parent child relationship	Fit for Life	+0.01	TG		NO
	Triple P	+0.00	=	=	NO
	Combined	+0.01	TG		NO
Q of life: Self.esteem	Fit for Life	+0.03	TG		NO
	Triple P	+0.07	TG		NO
	Combined	-0.10		CG	NO
Q of life: Have fun at school	Fit for Life	-0.09		CG	NO
	Triple P	-0.15		CG	NO
	Combined	-0.21		CG	NO
Q of life: Fun and laughter*	Fit for Life	-0.05		CG	NO
	Triple P	-0.09		CG	NO
	Combined	+0.04	TG		NO
Q of life: Relationship to friends*	Fit for Life	-0.21		CG	NO
	Triple P	0.00	=	=	NO
	Combined	-0.15		CG	NO

* Data for these sub-dimensions were not published in the German report and were compiled from the English version of the report.

Possible reasons for the divergent conclusions

All in all the reanalysis thus suggests that the ESSKI study shows no evidence in support of positive effects of the interventions. It is hence important to examine why the research team may have arrived at diametrically opposite conclusions. Unfortunately, as mentioned above, the published report does not provide full information about how the data were analysed except stating that the results were "significant" (Schönenberger, Lattmann et al., 2006: 20). However, the unpublished English research report includes a paragraph on the data analysis (Schmid et al., 2007). It explains that "for teachers, effects of group and time of measurement were tested applying a two way analysis of variance with one repeated factor (ANOVAR). Our main hypothesis, that Condition 3 (Fit For Life + Triple P) is the most effective in terms of health promotion for participants can be tested through the interaction between group and time. The same analysis was also applied on the individual level for parents as well as for students, using the general linear modeling approach. In order to control for the clustering in school classes, however, we also report results from a mixed model analysis where we tested for significant variation of the intercepts. The test represents differences between school classes and the analysis controls for possible differences in the primary sampling unit of school classes." (Schmid et al., 2007: 9)

These sentences suggest a misinterpretation of the results of the analyses of variance that were conducted. In particular, the GLM outputs shown in the appendices (Schönenberger, Schmid et al., 2006: 44-50) suggests that two equations were estimated for each outcome, namely…

Equation 1 Outcome = Group + Time + (Group*Time)

Equation 2 Outcome = Group + Time + (Group*Time) + Intercept

The authors therefore estimated two main effects and one interaction effects for each target outcome. The explanations given in the report (see quote above) fail to clarify entirely which of these components they consider evidence for the intervention effects. It is therefore worthwhile to explore what the components of the equations actually test.

The factor GROUP measures the main effect of treatment group membership across all three assessments at t_{pre}, t_{post}, and $t_{follow-up}$. Its F-value indicates whether the means of the four treatment conditions differ in any way. No ex-post tests were conducted to examine which group differs in what direction from any of the others. However, overall group differences – especially when averaged across three assessments - are not relevant for assessing treatment effects.

TIME measures whether there are any differences, averaged across treatment conditions, between Pre, Post, and follow-up assessments. Since the authors did not specify a linear effect, it is impossible to establish in what direction these differences go. A significant effect may mean a declining trend, an increasing trend, or any kind of zig-zag pattern. However, time effects are also irrelevant for assessing treatment effects.

The only relevant factor for assessing intervention effects is the GROUP*TIME interaction. However, a significant interaction only tells us that two or more groups differed in different ways over time. It does not tell us where, i.e. between which groups, and when, i.e. between which time points, differences existed. Neither does a significant F value tell us anything about the direction of these differences. For example, it could easily be that a significant effect results from a better development in the control group in comparison to all other groups.

To demonstrate treatment effects the authors would have needed to conduct at least two additional analyses. First, having identified a significant TIME*Group effect they would have to try to localize more precisely where the effect originates. A variety of so-called ex-post tests are available to explore this issue. However, a significant ex-post test does not tell in what direction the significant difference occurred. In a last step, therefore, the researcher needs to compare means in order to establish whether the control-group actually changed less in the desired direction as compared to the treatment group. It is only after there two additional steps that a positive treatment effects has been demonstrated at a given level of statistical significance.

INTERCEPT measures deviations of class-level means from the grand mean and is only relevant as a control variable that increases the efficiency of the estimates of the substantive variables.

Conclusions and Limitations

The reanalysis conducted in this paper suggests that the interventions of the ESSKI study did not have positive effects in two main targeted domains. Also, there is no evidence that any one of the treatment conditions systematically differed from the control group. Neither did the Triple P intervention nor the “fit-for-life” social skills programme result in significant change of behaviour outcomes. For the third domain, smoking and intent to smoke, the nature of the data rends analyses of causal effects impossible.

This reanalysis has limitations. For example, I have not examined further target measures that may be relevant as study outcomes. I hence do not know whether the study had positive effects on domains that were not emphasized in the media release. Also, I limited this analysis to an examination of effects between the pre and the follow-up measures. It is therefore possible that short term effects were achieved. As mentioned, examining effects at follow-up is justified in light of the authors’ emphasis on the long long-term preventive effectiveness of the interventions. Also, one may note that the four groups were not fully equivalent at t_0. Accordingly more detailed analyses should probably include covariates. Finally, one may note that the ESSKI study used a cluster randomized design with classes as randomization units. One would therefore wish to conduct multi-level (see, e.g. Cook, 2005). However, models that account for group-randomization usually yield fewer significant effects than the simple between-group effect sizes reported here.

In any case, effectiveness is associated, for practitioners and a wider public, with the idea that the treated group improves relative to the control group. As shown above, this is not the case in more than 50 percent of the comparisons conducted in this paper (i.e. 15 out of 27 comparisons). It may well be that the authors of the study still conclude that the interventions had the desired effects. In this case, however, special efforts should be made to explain the chosen methodological approach.

The findings of this paper were presented to the authors of the ESSKI study. The authors maintain that their findings have been achieved using appropriate methods and that the statements in the media-release adequately reflect the empirical findings. Our opinions differ on this point.

There are more general conclusions that can be drawn from this reanalysis. Most importantly, Switzerland and Germany have seen a considerable increase in the number of randomized controlled experiments that aim at identifying effective programmes for preventing mental health problems or behavioural problems amongst children and adolescents. This is a positive development that will hopefully contribute to more effective prevention strategies in the future. However, as yet there are hardly any quality-control mechanisms that provide guidance about how to report and present evaluation findings in a way that is transparent to practitioners and decision-makers. It would be highly desirable, therefore, to make increased efforts for disseminating methodological standards for conducting and presenting evaluation research. Such standards are available. Farrington (2003), for example, has recently suggested a comprehensive set of methodological quality standards for evaluation research that can be understood and easily used by scholars, practitioners, policy makers, the mass media, and systematic reviewers. This includes, for example, a list of minimum elements required in evaluation reports such as a full description of the statistical methods used, the effect sizes and their confidence intervals, possible threats to internal validity, and a full documentation of possible conflicts of interests. Farrington (2003: 55) also suggests that professional associations and funding agencies should get together to develop a checklist of items that must be included in all research reports on impact evaluations. Such standards would greatly help to establish trust, outside the scientific community, in the results of evaluation studies.

Appendix 1 Within-Group Effect Sizes - Detailed Tables

Note: All means, standard deviations and Ns are based on the tables provided by the project authors.
* The effect size in the treatment group is different from the effect size in the control group at $p < 0.05$.

Table 1 Children's Strengths, Teacher Assessment

	Mean (Standard Deviation)			Effect Size	84 % Confidence Interval	
	Pre	Follow-Up	Change	Cohens d	Lower	Upper
Control Group (N = 303)	2.41 (0.40)	2.52 (0.36)	+0.11	**+0.29**	+0.17	+0.40
„Fit for Life" (N = 317)	2.40 (0.44)	2.44 (0.41)	+0.04	**+0.09**	-0.02	+0.21
Triple P Group (N = 357)	2.40 (0.40)	2.47 (0.37)	+0.07	**+0.18**	+0.08	+0.29
Combined Group (N = 265)	2.49 (0.38)	2.47 (0.37)	-0.02	**-0.05***	-0.18	+0.07

Children's strengths have declined in the combined group and increased in the other three groups. The change in the combined group was significantly worse than the change in the control group.

Table 2 Children's Difficulties, Teacher Assessment

	Mean (Standard Deviation)			Effect Size	84 % Confidence Interval	
	Pre	Follow-Up	Change	Cohens d	Lower	Upper
Control Group (N = 303)	1.42 (0.34)	1.35 (0.31)	-0.07	**+0.21**	+0.10	+0.33
„Fit for Life" (N = 317)	1.42 (0.33)	1.40 (0.33)	-0.02	**+0.06**	-0.05	+0.17
Triple P Group (N = 357)	1.37 (0.30)	1.31 (0.29)	-0.06	**+0.20**	+0.10	+0.31
Combined Group (N = 265)	1.38 (0.32)	1.31 (0.30)	-0.07	**+0.23**	+0.11	+0.35

Difficulties have decreased in all four groups. There are no significant differences between groups at $p < 0.05$.

Table 3 Children's Strengths, Parent Assessment

	Mean (Standard Deviation)			Effect Size	84 % Confidence Interval	
	Pre	Follow-Up	Change	Cohens d	Lower	Upper
Control Group (N = 140)	2.51 (0.30)	2.56 (0.28)	+0.05	**+0.17**	-0.00	+0.34
„Fit for Life" (N = 166)	2.44 (0.33)	2.49 (0.34)	+0.05	**+0.15**	-0.01	+0.30
Triple P Group (N = 126)	2.54 (0.28)	2.58 (0.25)	+0.04	**+0.15**	-0.03	+0.33
Combined Group (N = 127)	2.46 (0.29)	2.54 (0.29)	+0.08	**+0.27**	-0.10	+0.45

Children's strengths have increased in all four conditions according to parent assessments. There are no significant differences between treatment groups and the control group.

Table 4 Children's Difficulties, Parent Assessment

	Mean (Standard Deviation)		Change	Effect Size Cohens d	84 % Confidence Interval	
	Pre	Follow-Up			Lower	Upper
Control Group (N = 140)	1.40 (0.29)	1.38 (0.32)	-0.02	**+0.07**	-0.10	+0.23
„Fit for Life" (N = 166)	1.45 (0.30)	1.42 (0.30)	-0.03	**+0.10**	-0.05	+0.25
Triple P Group (N = 126)	1.48 (0.30)	1.36 (0.23)	-0.12	**+0.45***	+0.27	+0.63
Combined Group (N = 127)	1.45 (0.30)	1.38 (0.30)	-0.07	**+0.23**	+0.06	+0.41

Children's Difficulties have decreased in all groups. The decline in the Triple P only group is significantly better than the decline in the control group.

B) Children's Quality of Life

Table 5 Parent-Child Relationship, Child Assessment

	Mean (Standard Deviation)		Change	Effect Size Cohen's d	84 % Confidence Interval	
	Pre	Follow-Up			Lower	Upper
Control Group (N = 254)	2.64 (0.56)	2.69 (0.56)	+0.05	**+0.09**	-0.04	+0.21
„Fit for Life" (N = 244)	2.71 (0.54)	2.76 (0.47)	+0.05	**+0.10**	-0.03	+0.23
Triple P Group (N = 241)	2.67 (0.55)	2.72 (0.54)	+0.05	**+0.09**	-0.04	+0.22
Combined Group (N = 165)	2.65 (0.56)	2.70 (0.48)	+0.05	**+0.10**	-0.06	+0.25

Parent-child relationships have increased in all four groups. There are no differences between groups.

Table 6 Self-esteem, Child Assessment

	Mean (Standard Deviation)		Change	Effect Size Cohen's d	84 % Confidence Interval	
	Pre	Follow-Up			Lower	Upper
Control Group (N = 253)	2.51 (0.58)	2.58 (0.58)	+0.07	**+0.12**	-0.00	+0.25
„Fit for Life" (N = 247)	2.55 (0.56)	2.63 (0.53)	+0.08	**+0.15**	+0.02	+0.27
Triple P Group (N = 243)	2.47 (0.55)	2.59 (0.57)	+0.12	**+0.21**	+0.09	+0.34
Combined Group (N = 167)	2.60 (0.52)	2.61 (0.51)	+0.01	**+0.02**	-0.13	+0.17

Self-esteem has increased in all four groups. There are no significant differences between groups.

Table 7 Enjoying School, Child Assessment

	Mean (Standard Deviation)		Change	Effect Size Cohen's d	84 % Confidence Interval	
	Pre	Follow-Up			Lower	Upper
Control Group (N = 254)	2.45 (0.65)	2.53 (0.59)	+0.08	**+0.13**	-0.00	+0.25
„Fit for Life“ (N = 244)	2.58 (0.61)	2.59 (0.58)	+0.01	**+0.02**	-0.11	+0.14
Triple P Group (N = 221)	2.58 (0.60)	2.57 (0.59)	-0.01	**-0.02**	-0.15	+0.12
Combined Group (N = 165)	2.73 (0.49)	2.69 (0.53)	-0.04	**-0.08**	-0.23	+0.08

Enjoying school has increased most in the control group and decreased most in the combined group. However, differences fail to be significant at the $p < 0.05$ level.

Table 8 Fun and Laughter, Child Assessment

	Mean (Standard Deviation)		Change	Effect Size Cohen's d	84 % Confidence Interval	
	Pre	Follow-Up			Lower	Upper
Control Group (N = 253)	2.59 (0.58)	2.58 (0.58)	-0.01	**-0.02**	-0.11	+0.09
„Fit for Life“ (N = 246)	2.67 (0.55)	2.63 (0.53)	-0.04	**-0.07**	-0.05	+0.20
Triple P Group (N = 242)	2.65 (0.55)	2.59 (0.57)	-0.06	**-0.11**	-0.02	+0.23
Combined Group (N = 167)	2.66 (0.52)	2.67 (0.51)	+0.01	**+0.02**	-0.17	+0.13

There are no differences in change scores between groups.

Table 9 Good Relations to Friends, Child Assessment

	Mean (Standard Deviation)		Change	Effect Size Cohen's d	84 % Confidence Interval	
	Pre	Follow-Up			Lower	Upper
Control Group (N = 251)	2.71 (0.56)	2.78 (0.48)	+0.07	**+0.13**	+0.01	+0.26
„Fit for Life“ (N = 242)	2.81 (0.46)	2.77 (0.49)	-0.04	**-0.08**	-0.21	+0.04
Triple P Group (N = 239)	2.69 (0.55)	2.76 (0.49)	+0.07	**+0.13**	+0.01	+0.26
Combined Group (N = 165)	2.78 (0.49)	2.77 (0.49)	-0.01	**-0.02**	-0.17	+0.13

„Good relations to friends” have decreased most in the „Fit for Life“ Group. They have increased most in the Control group and the Triple P group. The difference between the “Fit for life” group and the control group fails to be statistically significant.

References

American Psychological Association. (2001). *Publication Manual of the American Psychological Association: fifth Editions*: APA.

Bullinger, M., von Mackensen, S., & Kirchberger, I. (1994). KINDL - Ein Fragebogen zur Erfassung der gesundheitsbezogenen Lebensqualität von Kindern. *Zeitschrift für Gesundheitspsychologie, 1*, 64-77.

Burow, F., Asshauer, M., & Hanewinkel, R. (1998). *Fit und stark fürs Leben. 1. und 2. Schuljahr. Persönlichkeitsförderung zur Prävention von Aggression, Rauchen und Sucht*. Leipzig: Ernst Klett Grundschulverlag.

Burow, F., Asshauer, M., & Hanewinkel, R. (1999). *Fit und stark fürs Leben. 3. und 4. Schuljahr. Persönlichkeitsförderung zur Prävention von Aggression, Rauchen und Sucht*. Leipzig: Ernst Klett Grundschulverlag.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cook, T. D. (2005). Emergent principles for the design, implementation and analysis of cluster-based experiments in social science. *Annals of the American Academy of Political and Social Sciences, 599*, 176-198.

Eisner, M., Ribeaud, D., & Bittel, S. (2006). *Prävention von Jugendgewalt: Wege zu einer evidenzbasierten Gewaltprävention*. Bern: Eidgenössische Ausländerkommission.

Eltern und Schule stärken Kinder. (2007). ESSKI-Projekt macht Kinder stark (Media release, 18 January 2007, available at www.esski.ch, last accessed on 16 March 2007).

Farrington, D. (2003). Methodological Quality Standards for Evaluation Research. *Annals of the American Academy of Political and Social Sciences, 587*, 49-68.

Farrington, D., & Welsh, B. (2003). Family-based Prevention of Offending: A Meta-analysis. *Australian and New Zealand Journal of Criminology, 36*(2), 127-151.

Goodman, R. (2001). Psychometric Properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry, 40*(11), 1337-1345.

Hedges, L., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. New York: Academic Press.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-Analysis*. Thousand Oaks: Sage.

Lösel, F., & Beelmann, A. (2003). Effects of Child Skills Training in Preventing Antisocial Behavior: A Systematic Review of Randomized Evaluations. *The ANNALS of the American Academy of Political and Social Science, 587*(1), 84-109.

Lösel, F., & Koferl, P. (1989). Evaluation Research on Correctional treatment in West Germany: A Meta-Analysis. In H. Wegener, F. Lösel & J. Haisch (Eds.), *Criminal Behavior and the Justice System: Psychological Perspectives*. New York: Springer.

Payton, M. E., Greenstone, M. H., & Schenker, N. (2003). Overlapping Confidence Intervals or Standard Error Intervals: What do They Mean in Terms of Statistical Significance. *Journal of Insect Science, 3*(34), 34.

Ravens-Sieberer, U. (1998). Assessing health-related quality of life in chronically ill children with the German KINDL: first psychometric and content analytical results. *Quality of Life research, 7*(5), 399-407.

Reyno, S. M., & McGrath, P. J. (2006). Predictors of parent training efficacy for child externalizing behavior problems - a meta-analytic review. *Journal of Child Psychology and Psychiatry, 47*(1), 99-111.

Sanders, M. R., Markie-Dadds, C., & Turner, K. T. (2003). Theoretical, Scientific and Clinical Foundations of the Triple Positive Parenting Program:A Population Approach to the Promotion of Pa renting Competence. *Parenting Research and Practice Monograph, 1*, 1-21.

Schmid, H., Anliker, S., Bodenmann, G., Cina, A., Fäh, B., Kern, W., et al. (2007). *Empowerment in family and schools (eifas): A randomized controlled trial (unpublished report).*

Schönenberger, M., Lattmann, U. P., Fäh, B., Schmid, H., Bodenmann, G., Cina, A., et al. (2006). Eltern und Schule stärken Kinder" (ESSKI); Konzept eines mehrdimensionalen Forschungs- und Entwicklungsprojektes im Bereich psychosoziale Gesundheit in Schule und Elternhaus. *abhängigkeiten*(3).

Schönenberger, M., Schmid, H., Fäh, B., Bodenmann, G., Lattmann, U. P., Cina, A., et al. (2006). *Projektbericht "Eltern und Schule stärken Kinder" (ESSKI); Ein Projekt zur Förderung der Gesundheit bei Lehrpersonen, Kindern und Eltern und zur Prävention von Stress, Aggression und Sucht - Ergebnisse eines mehrdimensionalen Forschungs- und Entwicklungsprojekts im Bereich psychosoziale Gesundheit in Schule und Elternhaus*. (report available at http://esski.ch/html/download.htm).

Shadish, W. R., Cook, T. D., & Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

St Pierre, T. L., Osgood, D. W., Mincemoyer, C. C., Kaltreider, D. L., & Kauh, T. J. (2006). Results of an Independent Evaluation of Project ALERT Delivered in Schools by Cooperative Extension. *Prevention Science, 6*(4), 305-317.

Wilson, S. J., & Lipsey, M. W. (2006). The Effects of School-based Social Information Processing Interventions on Aggressive Behavior, Part I: Universal Programs (Campbell Collaboration Systematic Review, http://www.campbellcollaboration.org/doc-pdf/wilson_socinfoprocuniv_review.pdf).

Wilson, S. J., Lipsey, M. W., & Derzon, J. H. (2003). The Effects of School-Based Intervention Programs on Aggressive Behavior: A Meta-Analysis. *Journal of Consulting and Clinical Psychology, 71*(1), 136-149.