The Relationship Between Short-term Outcomes and Long-term Outcomes Produced by Outdoor Adventure Programmes:

A Rapid Review

April 2016

Alison O'Mara-Eves[1], Caroline Fiennes,[2] Sandy Oliver[1]

[1] EPPI-Centre, UCL Institute of Education
[2] Giving Evidence

About The Authors

Giving Evidence is a consultancy and campaign, promoting charitable *giving* based on sound *evidence*.

Through consultancy, Giving Evidence helps donors and charities in many countries to understand their impact and to raise it. Through campaigning, thought-leadership and meta-research, we show what evidence is available and what remains needed, what it says, and where the quality and infrastructure of evidence need improving.

Giving Evidence was founded by Caroline Fiennes, a former award-winning charity CEO, and author of *It Ain't What You Give*. Caroline speaks and writes extensively about these issues, e.g., in the Stanford Social Innovation Review, Freakonomics, and the Daily Mail. She is on boards of The Cochrane Collaboration, Charity Navigator (the world's largest charity ratings agency) and the US Center for Effective Philanthropy.

The Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) is part of the Social Science Research Unit (SSRU), UCL Institute of Education, University College London.

The EPPI-Centre was established in 1993 to address the need for a systematic approach to the organisation and review of evidence-based work on social interventions. The work and publications of the Centre engage health and education policy makers, practitioners and service users in discussions about how researchers can make their work more relevant and how to use research findings.

Founded in 1990, the Social Science Research Unit (SSRU) is based at the UCL Institute of Education, University College London. Its mission is to engage in and otherwise promote rigorous, ethical and participative social research as well as to support evidence-informed public policy and practice across a range of domains including education, health and welfare, guided by a concern for human rights, social justice and the development of human potential.

***

The views expressed in this work are those of the authors and do not necessarily reflect the views of the EPPI-Centre or the funder. All errors and omissions remain those of the authors.

The study was guided by an Advisory Group drawn from the Sail Training sector (listed in Appendix 1), and we are grateful to them for their input. It was conducted from July 2015 to February 2016.

Contact:
Caroline Fiennes
Director, Giving Evidence
+44 7803 954512
caroline.fiennes@giving-evidence.com

Table of Contents

Boxes

Box 1: What is Sail Training?

Box 2: What are primary studies and systematic reviews?

Box 3: Why we look at the method used in research and its rigour

Box 4: What it means for an outcome to be improved or maintained

Box 5: The impact of non-cognitive skills on outcomes for young people

Summary

Introduction

This study reviewed the existing literature aiming to **'identify key short-term outcomes that, if changed by an intervention [with young people], have a beneficial impact on longer-term outcomes such as attendance and/or attainment in education, employability, and mental health'.**

Our interest was specifically in outdoor learning programmes that aim to improve young people's life trajectory. This review should therefore be relevant to anybody interested in outdoor or adventure programmes for young people. We sought to identify short-term outcomes which are observable in the short-term which turn out to be good 'markers' for longer-outcomes.

Activity (the intervention)	→	**Short-term outcomes** (interim)	→	**Longer-term outcomes** (end goal)
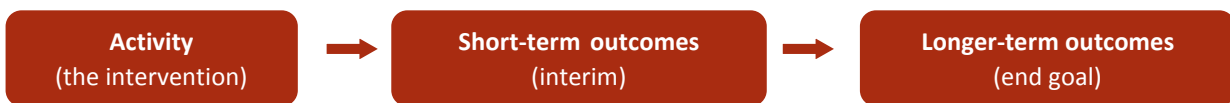
Method

This was a rapid review of the existing literature. We focused on 16 primary studies. Those we found by searching for relevant systematic reviews, of which we found 21. We harvested primary studies from them, and from other methods (such as looking for articles that cite sources already identified, known as 'forward citation chasing'). This produced 42 primary studies, of which 16 proved to be relevant.

We used them to answer our research question, of course, but also pulled from them insights about: the research literature and how it might usefully develop; how evaluations are run in this arena; how intervention effectiveness can be improved; harms sometimes created by these interventions.

Findings

We found five outcomes which are improved by such outdoor programmes and where the improvement is maintained or increases after the programme. Those are: behavioural conduct; academic performance; coping skills; locus of control; and weight loss. Nine other outcomes show an improvement after these programmes which is either maintained or increased thereafter – but only when measured on *some scales*; measured on other scales, those improvements are not maintained. We found one outcome which is improved after the intervention which is correlated with improvement in a *different* outcome: weight-loss after the programme seems to have a significant relationship with improved self-reported physical appearance some time later.

However, the literature is quite weak. Few studies (four) investigated the relationship between one outcome and different outcomes at later points. Only half of the studies had any kind of comparison group (to compare the effect of the intervention with, say, the normal maturation process), and only two comparison groups were randomised (i.e., in only two studies were the groups which got the programme and which got something else likely to be similar in all other respects). Almost every study had major problems collecting data on participants some time after the programme: this 'attrition' was worse amongst people who had been on the adventure programme, and affects the reliability of the conclusions. Several study authors wondered whether everybody in the programme actually received the same intervention. None of the studies was conducted in the UK, and, oddly, fewer than half were published in academic journals.

Some studies found harms to some participants, such as reduced self-esteem, and some even found harms quite broadly across the sample. These seem to be abated by high instructor-to-participant ratios, and by tailoring the programme to the participants.

Implications

If Sail Trainers (and providers of other outdoor adventure programmes) all measure their performance on the five outcomes which are maintained or improved over time, they will be able to accumulate their findings and strengthen the evidence used for decisions. It may help their case to funders if they use the measurement scales used in the studies here, since they are well-validated, and consistency between providers will enable comparison. Since there are so many unanswered questions about the effects of outdoor programmes including Sail Training, the Sail Training community may do well to engage with researchers to persuade and support them to address questions of greatest interest to practitioners.

Introduction

This study reviewed the existing literature aiming to **'identify key short-term outcomes that, if changed by an intervention [with young people], have been demonstrated to have a beneficial impact on longer-term outcomes such as attendance and/or attainment in education, employability, and mental health'.**

It was commissioned by the Association of Sail Training Organisations (ASTO) and Sail Training International (STI) (see Box 1). One aim of Sail Training is to improve young people's life trajectory, moving them closer to training or employment; ASTO and STI hypothesise that Sail Training, and other activities with young people, can produce short-term outcomes, some of which foreshadow important longer-term outcomes. However, though there is anecdotal evidence for this, there is little research which provides strong scientific evidence of it. Reasons for this might be the cost of measuring and evaluating an intervention long after it has finished, or the difficulty in getting responses from participants who might be hard to reach. Measuring short-term outcomes—i.e., immediately after the completion of the intervention—is more practical, but does not tell us about the maintenance or changes in intervention effects further in the future.

Sail Training itself has a limited research literature. However Sail Training is only one type of outdoor learning that aims to improve young people's life trajectory. Other types include adventure training, ropes courses, challenge courses and experiential education. Therefore, **we took a broader scope, including residential, or experiential education programmes that involve outdoor adventure** (hereafter referred to as adventure programmes) since this was likely to capture greater evidence about outcomes relevant to Sail Training and providers of other interventions with young people. We sought to identify short-term outcomes that either last for months or years, or are associated with important long term outcomes: this would pave the way for Sail Trainers and others to investigate through primary research whether and when and for whom their interventions produce those short term 'marker' outcomes.

| Activity (the intervention) | → | Short-term outcomes (interim) | → | Longer-term outcomes (end goal) |
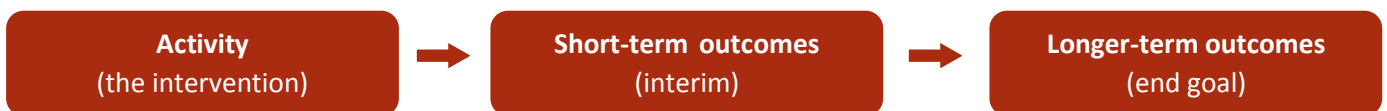
This review should therefore be relevant to anybody interested in the effects of outdoor adventure programmes for young people.

Giving Evidence, which works to enable charitable *giving* and charitable activity to be based on sound *evidence*, and the EPPI-Centre at University College London, which is committed to informing policy and professional practice with sound evidence, joined forces to conduct this review.

Box 1: What is Sail Training?

Sail Training provides challenge and adventure for young people. It takes them into a different environment where leadership, teamwork and comradeship are vital. It can be a valuable experience for young people from troubled backgrounds. More information is on http://www.asto.org.uk under 'Sail to Adventure'.

The Association of Sail Training Organisations (ASTO) is the umbrella charity for Sail Training in the UK. Its 30 members include the Tall Ships charities, including the Jubilee Sailing Trust, the Sea Cadets, and the Ocean Youth Trusts. ASTO provides funding for voyage fees, runs Sail Training races and promotes Sail Training in the UK.

Sail Training International is the international umbrella body for Sail Training with members all over the world. It works closely with the world's Sail Training Tall Ships, national Sail Training organisations and host ports, to help young people benefit from the Sail Training experience. It is the world's leading provider of races and events, conferences and seminars, publications, research and other services for the international Sail Training community.

However, though this research aimed to *inform* Sail Training, on which the literature is still young, we looked at similar studies where young people participated in residential programmes which involved a degree of risk taking out of doors, in other words outdoor adventure activities more broadly.

The research aims and question

Aims

To identify short-term outcomes that are linked to long-term outcomes after the completion of an adventure programme. By extension, to develop a short list of short-term outcomes that are likely to be indicative of future maintenance or growth of intervention effects.

Research question

Which outcomes show maintenance of adventure programme effects or increases in effects over time? A secondary research question: are any short-term outcomes linked to different outcomes at later measurement points?

Method

This was a rapid review of the research evidence found in studies that:

- Included (a) young people not in education, training or employment; (b) identified by the primary study authors as young people from disadvantaged, deprived, or low socio-economic groups or geographical areas; or (c) young people with experience of engagement with the justice system.
- Addressed sail training or similar activities (adventure programmes, ropes courses, and wilderness programmes) with an overnight stay.
- Measured outcomes from young people with and without engagement in one of these activities (for more detail see Appendix 2).
- Reported empirical data in 1995 or later, in the English language.

We attempted to be systematic in terms of transparency (through a clearly documented approach to locating evidence) and minimising bias (by not 'cherry-picking' evidence that supports a particular hypothesis), but cannot make claims about comprehensiveness. Whilst we endeavoured to locate as many studies as possible in the time and resources available, the search strategy (described below) is not as extensive as would be conducted in a full systematic review.

Box 2: What are primary studies and systematic reviews?

A primary study is a study of people. It can involve questionnaires, surveys or interviews, or other measurements about people such as their income or height.

A systematic review is a study of studies. It is a structured investigation to find, critically appraise and synthesise all the relevant primary research on a specific topic. Systematic reviews are stronger than non-systematic 'literature reviews' in that they: (i) can reconcile differences in the conclusions of different studies by looking across a larger set of participants; (ii) identify gaps to inform further research; (iii) are more transparent and hence can be reproduced by other researchers in future; and (iv) are less prone to bias, as science writer and doctor Ben Goldacre (2012) explains:

"Instead of just mooching through the research literature consciously or unconsciously picking out papers that support [our] pre-existing beliefs, [we] take a scientific, systematic approach to the very process of looking for evidence, ensuring that [our] evidence is as complete and representative as possible of all the research that has ever been done.

Thus a systematic review is more likely to be *accurate* and hence useful to practitioners for informing research and programme design than non-systematic literature. It is also more *credible* and hence useful in terms of convincing funders and policy-makers.

Each systematic review defines a *scope* (the topics, geography and timescale of interest) and the way that it will search for studies with that *remit* (the 'search strategy'). Most set some threshold for the *quality* of the primary studies they include in their analysis (the importance of quality of primary studies is discussed in Box 3). This is significant because the systematic review process is not magic: if the primary studies on which a systematic review is based are unreliable, the review's results will be unreliable. As a Yale cardiologist wrote recently on Twitter (Krumholz 2015): '*You can't just combine weak evidence and pretend that when mushed together it is strong. [Rather] it is meta-mush.*'

Searching for studies

Identifying studies in this review involved three successive stages:

1. First, we identified relevant systematic reviews. This yielded 21 systematic reviews;
2. Second, we found primary studies that appeared relevant from the systematic review authors' description. This yielded 42 primary studies;
3. Third, we inspected the full text of the reports of those 42 primary studies to check their relevance. We found that 16 of them were relevant, and 26 were not relevant to include.

The method for the search is described in more detail in Appendix 2. The included systematic reviews are listed in Appendix 3; the included primary studies are listed in Appendix 5; the reasons for the exclusions of the 26 primary studies are given in Appendix 4.

Figure 1: summarises the process through which we identified the primary studies
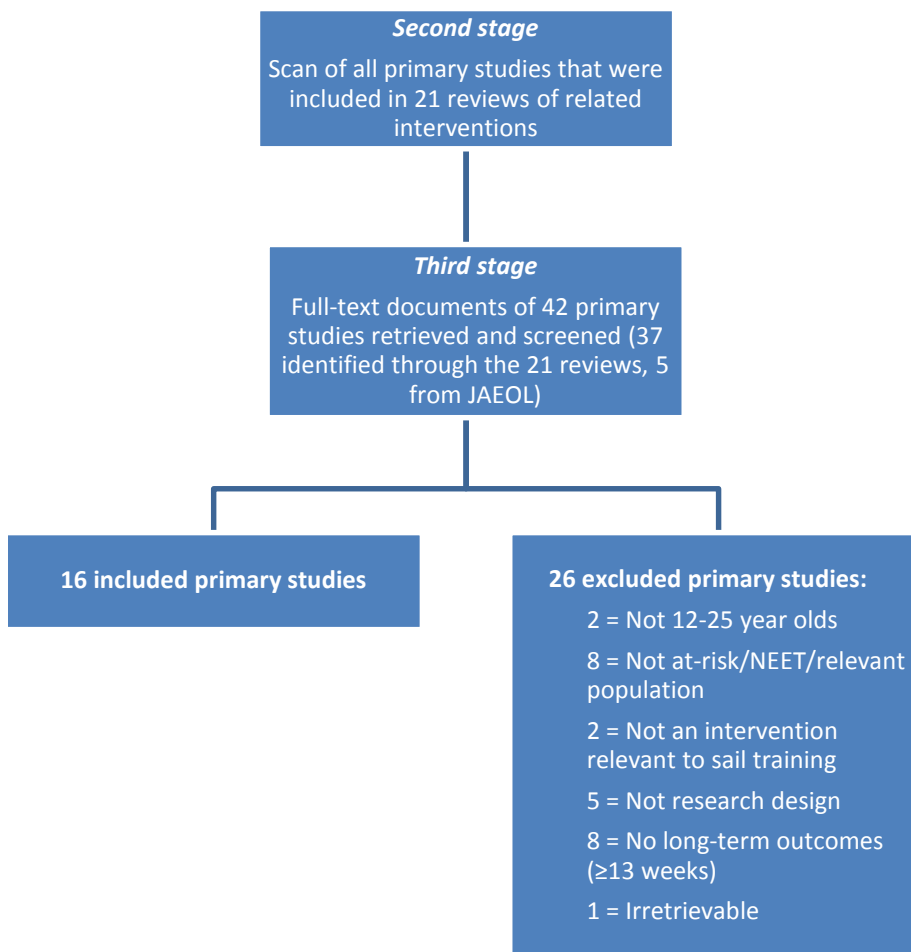
Note: *JAEOL = Journal of Adventure Education and Outdoor Learning; NEET = Not in education, employment, or training*

Characteristics of the included studies, their interventions, and their participants

Populations

Most studies looked at programmes with both genders. Three studies were of male-only programmes and two did not specify gender.

Table 1: Gender of participants in programmes studied

Gender	# studies
All male	3
Mixed	11
Unclear	2

Most studies looked at young people with some risk factor, most commonly that they were identified as 'at risk':

Table 2: Risk categorisation of participants in programmes studied

Identified focus	# studies *
At-risk	8
Behaviour-disordered	4
Mental health	2
Emotionally-disturbed	2
Substance abuse	2
Educationally-disengaged	2
Physical	1
Adjudicated youth	1
* Some studies included more than one focus type	

We found no relevant primary studies conducted in the UK. Most were in the US, with the remainder from Australia and Canada.

Table 3: Country of programmes studied

Country	# studies
USA	11
Australia	3

Canada	2
UK	None

Interventions

The duration of the interventions studied varied enormously, with one programme at several months, and the most common duration was 21 days. This is significant in relation to translating the literature to Sail Training, where most programmes are five days or a week.

Table 4: Length of programmes studied

Programme length (days)	# studies
2	1
3	1
8	1
10	1
13	2
21	4
28	1
45	1
49	1
57	1
70	1
112	1

The mix of programme models is as per Table 5 below. The definitions here are taken from Bowen & Neill's (2013) review of adventure programmes:

1. Expedition = programme is conducted primarily in wilderness on expeditions

2. Multiple = mix of programme components

3. Ropes, challenge, or adventure-based = ropes course, challenge course, or adventure-based therapy

4. Base camp = outdoor base camp supplemented by wilderness expeditions. These programmes may be dissimilar to Sail Training in that participants return to the base-camp every night, so there is less immersion in the unfamiliar.

Table 5: Programme model categories

Programme model	# studies
Expedition	5
Multiple	5
Ropes, challenge, or adventure-based	4
Base camp	2

Funding for the interventions studied (not necessarily for the studies themselves) was most commonly from the non-government sources

Table 6: Funding of interventions studied

Funding of the intervention	# studies
Private (non-government)	13
Public (government)	3

Comparison

The 16 included primary studies were evenly split between those which had, and which did not have, a control group

Table 7: Numbers of studies with/without control groups

Design	# studies
Comparison (with control group)	8
Time-series/longitudinal (no control group)	8

Of the eight studies which had a control group, the most common comparator was nothing, i.e., the control group received no programme. Two studies involved a 'waitlist', in which all participants get the same programme but some wait longer for it.

Table 8: Types of control group

Control group type (from 8 studies)	# studies *
Inactive (no treatment)	4
Waitlist	2
Alternative intervention	2
"Normal population" reference group	1
Unclear (no access to full report)	1
* Some studies included more than one comparison group	

Only two of the studies had control groups chosen at random (which increases the similarity between the group which gets the programme and the group which doesn't, and thus increases the chance that any differences in outcomes between them is caused by the programme rather than other factors).

Table 9: Selection method for control groups

Randomisation (from 8 studies)	# studies
Not randomised	5
Randomised	2
Unclear (no access to full report)	1

A major problem in almost all of the included studies was researchers' inability to gather data about the participants 'at follow-up' i.e., some time after the intervention finished. For only five of the 16 studies was the response to questions at follow-up equal to or above 70%. Worse, several studies stated that non-response at follow-up was more pronounced amongst people who had been on the programme than from people in the control groups, which made a fair comparison impossible.

Table 10: Follow-up rates

Follow-up rate	# studies
>=70%	5
<70%	7
Unclear amount, but stated problem with missing data	2
Unclear	2

Notes on follow-up:

- The time from the end of the intervention to the last follow-up point ranged from three months to 21 months
- The average time from the end of the intervention to the last follow-up was seven months
- Even for the five studies that had ≥70% response rate at follow-up, there were sometimes more non-responders in the treatment group
- Getting responses at follow-up was problematic even when researchers used various approaches and incentives to improve response rate.

Study authors and publication

Very unusually, the set of included studies contained fairly few which were published in academic journals. Six of the 16 were theses by Masters or PhD students.

Table 11: Nature of the studies

Publication type	# studies
Journal article	7
Thesis	6
Other (report, data set)	3

Box 3: Why we look at the method used in research and its rigour

i) Because different research methods give different answers

"Two men say they're Jesus: One of them must be wrong" (Dire Straits lyric!)

The table below shows the effect of a reading programme in India measured using several research methods[i]. These methods all used the same outcome measures, but the experimental designs were different.

The answers vary widely: some suggest that it works well, others show it to be detrimental. Clearly there can only be one correct answer! All the other answers are incorrect: and could mislead donors or practitioners to implement this programme at the expense of another which might be better.

Method	Impact Estimate
(1) Pre-post	26.42*
(2) Simple Difference	-5.05*
(3) Difference-in-Difference	6.82*
(4) Regression	1.92
(5)Randomized Experiment	5.87*

*: Statistically significant at the 5% level

The answers vary because research methods vary in how open they are to biases (i.e., systematic errors). For instance, suppose that a medical trial involves giving patients a drug for two years. Suppose that that drug has horrible side-effects such that during the two years, some patients can't stand taking it so they drop out of the trial (or worse, perhaps the drug kills some of them). If the trial only collects data on patients who are still in the trial after two years, it will systematically miss the important insights about those side-effects. This 'survivor bias' will make the drug look more effective than it really is.

Somebody reading the trial results without knowing that detail wouldn't be able to distinguish the actual effect of the drug from that of this survivor bias. Similarly, if a study only looks at the outcome (in the example above, it's reading level) before the programme and then afterwards (i.e., is a pre-post study), it won't be possible to distinguish whether any improvement in reading levels was due to the programme or just to the fact that children learn over time anyway.
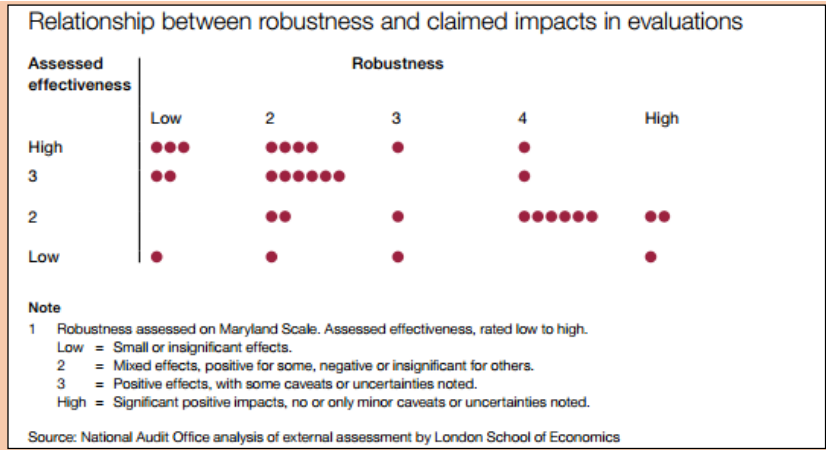
The tables below include a statement of the degree of bias in the studies used for this review, and the biases are discussed further in Appendix 2.

{As an aside, contrary to popular myth, it is not invariably the case that robust research is more expensive than unreliable research, nor that randomised controlled trials (the most reliable design for a single primary study) are invariably terribly expensive: many are cheap or free.}

ii) Because weaker research methods allow for more positive findings

The UK National Audit Office searched for literally every published evaluation of a UK government programme (National Audit Office 2013). Of those, it chose a sample, and ranked on one hand, the quality of the research method ('robustness' on the x axis, i.e., how insulated the study is from bias), and on the other, the positive-ness of the programme ('claimed impact').

The trend line on the resulting graph below would slope diagonally downwards. It shows that more robust research only allows for modest impact claims whereas weak research allows much stronger claims.

Relationship between robustness and claimed impacts in evaluations

Assessed effectiveness	Robustness				
	Low	2	3	4	High
High	●●●	●●●●	●	●	
3	●●	●●●●●●		●	
2		●●	●	●●●●●●	●●
Low	●	●	●		●

Note

1   Robustness assessed on Maryland Scale. Assessed effectiveness, rated low to high.
     Low   =   Small or insignificant effects.
     2   =   Mixed effects, positive for some, negative or insignificant for others.
     3   =   Positive effects, with some caveats or uncertainties noted.
     High   =   Significant positive impacts, no or only minor caveats or uncertainties noted.

Source: National Audit Office analysis of external assessment by London School of Economics

Bad research can be persuaded to say almost anything, and won't allow researchers to distinguish the effects of a programme from other factors (e.g., the passage of time, the mindset of participants, other programmes) nor from chance.

Most social interventions have a small effect and a reliable research method will show what that is: bad research is likely to overstate it. The highest estimate for the reading programme above is from the pre-post study which is a weak study design.

This relationship between weak research methods and positive findings has been shown in many fields including in education and in medical research.

Box 4: What it means for an outcome to be improved or maintained
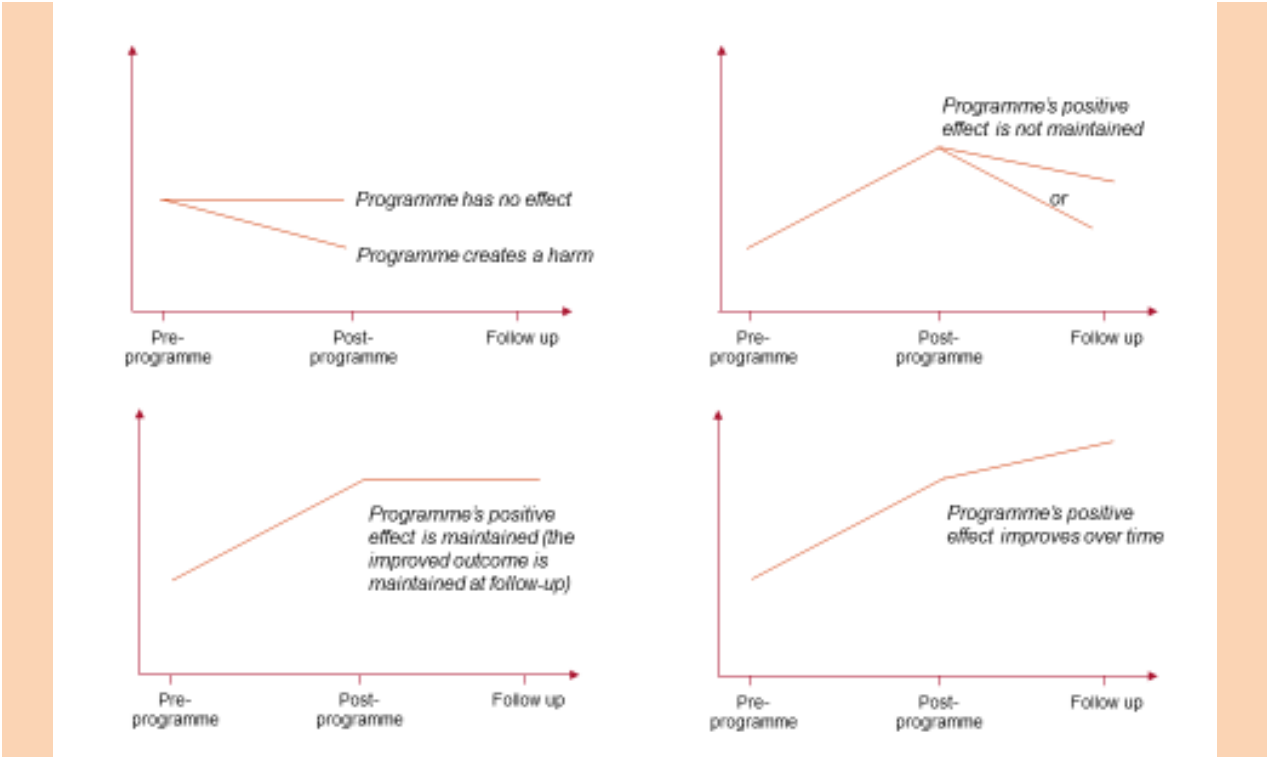
Outcomes studied

We identified 24 categories of outcomes, many of which also had sub-domains. We identified studies which found support for a link between the short-term and long-term measurement of each outcome. From this long-list, we made three short lists: a list of outcome links that had more promising support as defined by the strength of evidence (**Table 12**); a list of outcome links that had support when measured on some subscales but not all (**Table 14**); and a list of outcome links that did not have support (**Table 15**). **Table 13** breaks down the outcomes in **Table 12** to show which studies supported the link and some of the study characteristics. {Note that the range of times when 'long-term follow-up measures' were taken makes direct comparisons difficult.}

Table 12: Outcomes that were maintained or increased from post-test to long-term follow-up in at least one low risk of bias study or multiple substantial risk of bias studies

Outcome	Overall strength of evidence	Improved or maintained?
Behavioural conduct	Moderate strength from 9 studies: 6 with some risk of bias that support the link and 3 that did not support the link	Improved/maintained
Academic performance	Low strength from 2 studies with some risk of bias that support the link	Improved
Coping skills	Low strength from 2 studies with some risk of bias that support the link	Improved/maintained
Locus of control	Low strength from 3 studies: 2 with some risk of bias that support the link and 1 that did not support the link	Improved/maintained
Weight loss	Low strength from 1 study with low risk of bias that supports the link	Maintained

Table 13: The outcomes in Table 12 in detail: the studies reporting the outcomes, their follow-up periods, and their sample sizes

Outcome	Overall strength of evidence	Study	Maintained or improved?	Sample size
Behavioural conduct	Moderate (6 studies with risk of bias; 3 studies didn't find an improvement)	Russell 2002	Improvements in some scales at 12 months	858
		Lewis 2012	Improved disruptive behaviour at 12 months	190
		Brand 1998	Improved on some scales, maintained others at 21 months	190
		Parker 1995	Improved on one scales, not maintained on others at 6 months	84
		Garst 2001	Improved at 4 months	58
		Harper 2007	Maintained at 12 months	221
		Grayson 2001	No gain at 3 months	176
		Pommier 1995	Not maintained at 3 months	79
		McGarvey 2004	No gains at 7 months	44

Outcome	Overall strength of evidence	Study	Maintained or improved?	Sample size
Academic performance	Low (2 risk of bias studies)	Grayson 2001	Improved at 3 months	176
		Harper 2007	Improved at 12 months	221
Coping skills	Low (2 risk of bias studies)	Crisp 2003	Improved at 24 months	52
		Mohr 2001	Maintained at 3 months	102
Locus of control	Low (2 risk of bias studies; 1 didn't find an improvement)	Brand 1998	Improved at 21 months	190
		Martinez 2002	Maintained at 9 months	223
		Parker 1995	No gains at 6 months	84
Weight loss	Low (1 low risk of bias study)	Jelalian 2006	Maintained at 6 months	76

Table 14: Short-list of outcome links that had support for some subscales but not all (inconsistent evidence)

Outcome	Overall strength of evidence
General Self-esteem	Low (6 studies, mixed risk of bias; 7 didn't find an improvement)
School self-esteem	Low (2 studies, high risk of bias; 3 didn't find an improvement)
Social confidence	Low (2 studies, high risk of bias; 2 didn't find an improvement)
Social relationships	Low (1 study, high risk of bias; 1 didn't find an improvement)
Parental relationships, family functioning	Low (4 studies, high risk of bias; 1 didn't find an improvement)
Substance abuse	Low (2 studies, high risk of bias; 2 didn't find an improvement)
Suicide risk	Low (1 study, high risk of bias; 1 didn't find an improvement)
Criminogenic outcomes (e.g., attitudes towards police)	Low (1 study, high risk of bias; 1 didn't find an improvement)
Mental health (e.g., general; somatic symptoms)	Low (3 studies, high risk of bias)

Table 15: List of outcome links that did not have support, and the number of studies that reported those outcomes (i.e., where there was no significant improvement between baseline (pre-test) and follow-up)

Outcome	# of studies
Athletic competence	2
Community involvement	1
Depression	3
Empathy	1
Goals/aspirations	2
Group cohesion	2
Independence	1
Peer relationships	4
Self-efficacy	2
Social competence/acceptance	9

Cross-outcome analyses

Only four studies did any cross-outcome analysis, i.e., compared changes between different outcomes.

A significant relationship was found between weight loss at post-test and improved self-reported physical appearance at follow-up.

Most studies looked at maintenance of outcomes, rather than predicting later outcomes from different outcomes measured earlier. Beyond the one above, we can't say much about using different measures[1].

Discussion

Summary of the findings

We found five outcomes which are improved by such outdoor programmes and where the improvement is maintained or increases after the programme. Those are: behavioural conduct, academic performance, coping skills, locus of control, and weight loss. Nine other outcomes show an improvement after these programmes which is either maintained or increased thereafter – but only when measured on *some scales*; measured on other scales, those improvements are not maintained. We found one outcome which is improved after the intervention which is correlated with improvement in a *different* outcome: weight-loss after the programme seems to have a significant relationship with improved self-reported physical appearance some time later.

Strengths and limitations of the study

As the interest was in Sail Training in particular, and this literature is very small, we included other programmes which share key characteristics, such as young people engaging with a degree of risk, out of doors, and being away overnight. We drew on studies that made efforts to make fair comparisons between young people who had been offered or experienced adventure programmes and other very similar young people. We may not have found all the existing studies that have done this.

---

[1] The four studies were:

Jelalian (2006) – 'Adventure therapy' combined with cognitive-behavioral treatment for overweight adolescents. This is the only study to report significant cross-outcome findings. It reported that 'End of treatment weight loss' was associated with improvement in self-reported physical appearance in the intervention group but not the control.

Freedman (1996) – The effects of a therapeutic wilderness experience for emotionally disturbed adolescents. The Gross Cohesiveness Scale was not significantly correlated with the Goal Attainment Survey or the Coopersmith Self-esteem Inventory at posttest or follow-up.

Lewis (2012) – Examining changes in substance use and conduct problems among treatment-seeking adolescents.

Correlations between conduct and substance abuse were not significant between prior conduct and subsequent substance abuse at any time point, or vice versa.

Martinez (2002) – A wilderness therapy program for a diverse group of at risk adolescent boys: Changes in self-esteem and locus of control and their relationship to group affiliation

Looked at changes in self-esteem and locus of control as they related to perception of group affiliation, but analysis at was at pretest and posttest only (not follow-up).

Providers of outdoor learning and adventure (and other) programmes sometimes claim that their interventions improve young people's non-cognitive ('soft') skills, and that that has an effect on 'hard' skills. The UK Cabinet Office and the Education Endowment Foundation commissioned a literature review[ii] in 2013 of the above title to investigate. It defined 'non-cognitive skills' as "attitudes, behaviours, and strategies which facilitate success in school and workplace, such as motivation, perseverance, and self-control…factors termed 'non-cognitive'" and looked for evidence of their relationship to 'hard' outcomes such as literacy, numeracy, educational attainment, employment, health, well-being, engagement, employability, civic participation, and voting.

It found that "Despite increasing evidence that non-cognitive skills are strongly _correlated_ with success, the evidence seems to be less clear that there is a _causal_, robustly measurable relationship between such factors and [the 'hard' outcomes'].

"Factors such as self-control and school engagement are correlated with academic outcomes, financial stability in adulthood, and reduced crime. However, robust evidence of a _causal_ relationship is limited…"

It found that "No single non-cognitive skill predicts [all the meaningful] long-term outcomes."

In terms of the state of the literature, it found that: "Many studies define and measure non-cognitive skills in disparate ways, assess them in isolation, and focus on short-term outcomes. Priorities for future research should be to explore how skills can be transferred between areas of a young person's life, and how far changes can be sustained in the long term…"

The three non-cognitive skills for which it found appreciable evidence (see table below) that they have an appreciable effect on 'hard' outcomes are:

1. Self-efficacy: ""an individual's belief that they have the capability to succeed at a particular task in the future. Whereas self-concept of ability assesses how an individual feels about their past performance in relation to others, self-efficacy measures an individual's expectations about whether or not they can successfully perform a specific task at a later point in time. In practice, self-efficacy focuses on the successfully mastering of a specific task, while self-concept of ability is concerned with the affective appraisal of one's performance in an academic domain, relative to others."
2. Expectancy-value theory: ""motivation to achieve, consisting of (1) students' expectations of success and (2) their perception of the overall value of the activity or task", and
3. Meta-cognition: ""goal-oriented efforts to influence one's own learning behaviours and processes by focusing awareness on thinking and selecting, monitoring, and planning strategies that are most conducive to learning. [e.g.] setting goals, planning and problem-solving, being aware of one's strengths and weakness, monitoring one's progress and understanding, and knowing when and why to use certain strategies."

	Quality of measurement	Malleability	Effect on other outcomes	Strength of Evidence
1. Self-Perceptions				
Self-Concept of Ability	High	Medium	Not available	Medium
Self-Efficacy	High	High	High	Medium
2. Motivation				
Achievement Goal Theory	High	Medium	Low to medium	Medium
Intrinsic Motivation	High	Medium	Low to medium	High
Expectancy-Value Theory	Medium	Not available	Medium to high	Medium
3. Perseverance				
Engagement	Medium	Not available	Not available	Low
Grit	Medium	No evidence	No evidence	Low
4. Self-Control	Medium	Low to medium	Low	Medium
5. Meta-Cognition	Medium	Medium to high	Medium to high	High
6. Social Competencies				
Leadership Skills	Low	Not available	No evidence	Low
Social Skills	Medium	Medium to high	Low to medium	High
7. Resilience and Coping	Medium	High	Low	Medium
8. Creativity	Medium	Not available	No evidence	Low

Harms

Four studies discussed harms that were measured or identified (i.e., areas where participants performed worse after the intervention). In some studies, these harms were seen for some individual participants; importantly, **some studies found harms quite broadly across the sample**.

These were sometimes reduced self-esteem, loneliness, problems with other participants, "inappropriate relationships" with other participants, sexual activity. One study mentioned how a spike in self-esteem (an apparent improvement) for one participant might have been explainable through "inappropriate relationships" of that participant with other participants (Bocarro 1995).

Generally, harms were identified through interviews and/or qualitative research alongside the quantitative data collected.

The studies identify various ways to reduce harms, which include:

- high instructor to participant ratios
- having experienced and peer instructors
- individualised / tailored components to meet the needs of the participants
- consideration of group composition: catering to participants with very diverse needs was generally seen as a bad thing.

Clearly, future evaluation – and implementation – should not only try to protect against harms, but ensure that measures are in place to identify them if they do occur.

Running programmes

While extracting data from the studies, we found some potentially useful information about how the intervention and/or the evaluation was implemented. We did not set out to look at these topics, and therefore did not pre-specify categories for them but rather simply noted the information in the way given in the primary studies. It is therefore not a comprehensive list of all implementation issues to consider, nor is it 'objective'—in most cases, the list represents the views or observations of the primary researchers (usually noted in the discussion / limitation sections of the reports) rather than 'hard data' gathered as part of a rigorous process evaluation. We capture this information because it might provide useful guidance when planning a similar intervention or evaluation.

The general themes of the implementation issues are presented in Table 17, with the number of studies that mentioned something related to the theme. We then briefly describe the key points for each theme.

Table 17: Implementation issue noted by the primary study authors and the number of studies that discussed the issue

Implementation issue noted by the primary study authors	# studies
Connecting to real life; support after the intervention ends	10
Missing data; drop outs	10
Intervention deliverer	5
Domain specificity of outcomes	5
Emphasis on individual variation	5
Measuring harms	4
Alternative explanations through mixed methods	4
Evaluation planning	4
Duration	4
Group composition (e.g., mixed gender; mixed problems)	3
Expectation management	2
Intra-services cooperation	2

Connecting to real life and support after the intervention ends. Ten studies mentioned something relating to the need to make the intervention relevant to life after the intervention and/or the importance of developing support programmes (e.g., group reunion barbecues; ongoing telephone support) after the intervention ends.

Intervention deliverer. Five studies referred to the importance of the quality or appropriateness of the person delivering the intervention. In general, this involved reflecting on the need for people trained in youth work or fostering inclusion of the participants—that is, training or experience in skills beyond the adventure component. One study suggested that improvements seen in white male participants might have been attributable to the fact that the people running the programme were also white males (Bocarro 1995).

Group composition. Three studies reflected on the role that group composition might play in an intervention's effectiveness. Perhaps the most interesting suggestion here was that having a very diverse group might reduce an intervention's impact: "The implications for a treatment group with very diverse problem areas is that it can make effective processing of the adventure activities less relevant to each population" (McGarvey 2004).

Expectation management. Garst 2001 noted that "Providing participants with clear behavioral expectations (Priest & Gass, 1997) and enforceable consequences yields positive results. Participants suggested that expectations, rules, and consequences influenced positive behavioral changes during the trip, which may have influenced behavior once participants returned home." Michalski (2003), in contrast, noted that "one in ten parents expressed some concerns about the camp or conveyed a sense that their expectations had not been met", highlighting the need to specify what is achievable at the outset of the intervention.

Intra-service cooperation. Whilst this was only explicitly mentioned in two studies, the cooperation between multiple services (e.g., the adventure programmers, schools, and youth workers) is likely to be an important issue in ensuring the intervention and its participants are supported to actively engage in the programme.

Duration. Three studies noted that a minimum intervention length appeared to be necessary to affect change (Garst 2001; McGarvey 2004; Michalski 2003), while two studies suggested that the time commitment was the reason for many of the drop-outs in that study (Jelalian 2006).

Running evaluations: Problems in the evidence base

Missing data, drop outs and attrition. These were problems in almost all studies. Several studies mention problems getting responses even after extensive attempts to reach people and using multiple strategies (phone and mail, etc). Ten studies explicitly mentioned the challenges that they faced. Response rates were often lower for participants than for people in the control groups.

Three studies suggested reasons for this attrition. Bocarro (1995) noted that a lack of respect for the research meant that intervention staff did not administer the evaluation materials appropriately; Brand (1998) noted that scales which were lengthy or difficult to read probably increased non-response rates; and Grayson (2001) mentioned several barriers to evaluation including staff changes and lack of prioritisation of the research by schools.

Fidelity to intervention (i.e., whether all participants receive the same experience) was also noted as a problem. Several studies noted having to change the programme as it went along, usually due to organisational / logistical difficulties.

Variance in individual responses. There were benefits for some participants, neutral for others, and harms for others. This suggests / requires that the studies be interpreted with reduced emphasis on means / averages, and more exploration than normal on why it works for some people and not others, e.g., match between administrator and participant; interest or sense of burden; connecting to real life experiences. This also suggests the usefulness of qualitative data to help explain the variation.

Several studies used **mixed research methods**, i.e., they measured both quantitative outcomes and collected qualitative data. In four of the studies, the researchers elaborated on some of the observed trends in the quantitative data by exploring the qualitative data. For example, one study noted that "the children who did well (favorable changes) were agreeable and liked camp. The children who did consistently poorly had far fewer positive comments (especially considering the greater number of such children) and considerably more comments that reflected behavioral and emotional problems" (Grayson 2001). Qualitative data can add depth to the understanding of quantitative data, and is particularly helpful in exploring 'outliers' and harms. Interestingly, Michalski (2003) noted inconsistencies between quantitative data and qualitative data that highlight the need to assess impacts through different ways.

Domain specificity of outcomes. Five studies discussed whether having global measures (e.g., general self-esteem) might not be able to detect the benefits of an intervention, and suggested that more specific outcomes (e.g., task or sub-domain orientated measures) might be more appropriate for assessing the intervention.

Appendices

Appendix 1: Composition of Advisory Group

The Advisory Group met once at the beginning of the process to discuss scope and once when the research was complete and we were starting to draw conclusions and draft this report. It comprised:

James Stevens (Chairman, Association of Sail Training Organisations)

Lucy Gross (Association of Sail Training Organisations)

Paul Bowen (Sea Change Sailing Trust)

Eric Fletcher (Ocean Youth Trust North)

Alex Coakley (Morning Star Trust)

Julie Fellows (Island Trust)

Roger Crabtree (Tectona Trust)

Richard Thornton (UKSA, the UK Sailing Academy)

Leonie Back (The Cirdan Trust)

Nick Fleming (Ocean Youth Trust Scotland)

Rob Kelvey (RAKTrain)

Murray Henstock (Sail Training International)

Appendix 2: Research method

The following details the process for finding primary studies, identifying which primary studies to include, and the data extracted from them[2 iii].

A.   Searching for studies

1.   Identifying systematic reviews

We identified systematic reviews through searching various websites and databases; specifically, Web of Knowledge, ERIC through EBSCOhost, and Google Scholar. We used search terms related to adventure programmes (e.g., Sail Training, outdoor learning, adventure training, ropes course, challenge course, experiential education) and systematic reviews (e.g., systematic review, meta-analysis, literature review). The aim of this step was to capitalise on the systematic searches that have already been carried out for other reviews by identifying relevant primary studies included in those reviews.

2.   Finding primary studies

Having identified a list of 21 systematic reviews that appeared to be related to the topic (see Appendix 3), we examined the reviews to see whether they included relevant interventions, reported long-term outcomes, *and* had studies with our population of interest. From this investigation, ***we identified five reviews that might contain relevant studies*** (Bowen & Neill, 2013; Durlak et al., 2010; Gillis & Speelman, 2008; Hattie et al., 1997; Rickinson et al., 2004).

To capture more recent research, we searched for papers citing the five apparently relevant systematic reviews (**forward citation chasing**) using Scopus, which claims to be the largest abstract and citation database of peer-reviewed literature, covering scientific journals, books and conference proceedings (https://www.elsevier.com/solutions/scopus).

We also contacted the authors of the five apparently relevant systematic reviews to determine (a) whether they have updated their research on this topic since publication, and (b) whether they can provide information on the studies included in their reviews, especially the population characteristics and presence of follow-up measures. Responses were received from all five review teams.

To maximise the likelihood of identifying evaluations of sailing-related interventions, we searched a **database of research on the topic of outdoor learning** that had been identified for a previous review conducted by members of this team[3]. Searching this resource involved scanning titles in the database. Studies must have mentioned Sail Training, tall ships, or related seafaring programmes to be considered for full-text retrieval. The studies in this database are the product of systematic searches and have already been systematically classified; they thus represent a valuable shortcut to evidence.

We conducted **hand searching of the Journal of Adventure and Experiential Outdoor Learning** over the years 2003 to the current issue (and searched issues from the journal's launch in 2000 to 2002). This journal is known to be particularly relevant to the topic and was searched to identify any relevant, recent Sail Training interventions.

---

[2] The full protocol is published at http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3477

[3] More detail is at www.giving-evidence.com/outdoor-learning

The above search processes identified a corpus of studies that appeared to be relevant. Helpfully, authors of two of the aforementioned systematic reviews (Bowen & Neill, 2013 and Gillis & Speelman, 2008) provided information about follow-up data and the population for each of their included studies, so we were able to pre-screen many of the studies using the reviewers' descriptions of the studies. From the five systematic reviews mentioned above, **we identified 37 primary studies that appeared to meet our criteria** and for which full-text documents should be retrieved.

An additional five primary studies on adventure programmes (including one specifically on Sail Training) which we found from searching the database of outdoor learning and the Journal of Adventure and Experiential Outdoor Learning, were also retrieved for full-text screening.

3.   Sifting primary studies

Primary studies were screened on the full-text documents. To be included in our review, primary studies needed to meet the following criteria:

1.   Must be published in English language.
2.   Must be published in 1995 or later.
3.   Must be primary research of an empirical nature (not a literature review or opinion piece).
4.   Must include the population: 12 to 25 year olds. Where a wide age range is included in the study, then only studies whose mean age falls within 12-25 years, or that reports outcomes separately for one or more subgroups that fall within our range of interest, will be included.
5.   Must include the population: (a) young people not in education, training or employment; (b) identified by the primary study authors as young people from disadvantaged, deprived, or low socio-economic groups or geographical areas; or (c) young people with experience of engagement with the justice system.
6.   Must include the intervention: related to Sail Training, adventure programmes, ropes courses, and wilderness programmes. It must include an overnight component.
7.   Must report outcomes measured for *either (or both)* of the following:
    1.   both immediate post-test (at the completion of the intervention) *and* long-term follow-up (at least 13 weeks after the completion of the intervention) for *both* a treatment and a comparison group. (The value in limiting to studies with a comparator is that we can be more confident that any changes in the outcome over time are due to the effects of the intervention, rather than maturational or developmental changes.) These studies will be analysed using a "difference-in-differences" approach, which tests whether there are differences in the changes across time between the two groups. And / or ...
    2.   three time points (baseline, immediate post-test, and long-term follow-up). These studies will be analysed using a "time-series" approach, to see whether intervention effects are maintained, increased, or decreased at follow-up relative to post-test, taking into account the score at baseline.

For the purposes of identifying studies and for the analyses, ***we did not limit the outcome type***. Any outcome that is measured at both immediate post-test and long-term follow-up (i.e., at least 3 months after the intervention ends) is potentially eligible for inclusion. This is because the purpose of this review is to identify *which* outcomes might be useful to measure in future research, rather than establish whether a particular outcome is benefited by such interventions.

For the purposes of identifying studies, initially we did not exclude studies without a comparator group. However, we stated in the research protocol (published before work began) that if we found sufficient studies with comparator groups, we will subsequently exclude studies without a comparator group. The problem with excluding non-comparator studies at the outset is that we might not have sufficient includable studies, plus there is likely to be some useful information to be gained from time-series designs (which do not have a comparison group). For the time-series designs, we would need three time points, so that we could establish whether the intervention had any immediate benefits and then whether the scores persisted, increased, or decreased at follow-up measurement.

The 16 included studies are listed in Appendix 5.

B. Data extraction

Thanks to the helpfulness of Daniel Bowen, co-author of the Bowen and Neill (2013) review, we were provided with a dataset that comprises most of the studies that we included in our review. This dataset includes Bowen and Neill's data extractions on a range of variables (see Appendix 7). We manually extracted our own data on the variables in Appendix 8, which were seen as particularly important to the purpose of our review. (We therefore had access to the extracted data for the variables in Appendix 8 *plus* the data manually extracted by our team, below.)

We additionally collected data on the outcomes measured and results pertaining to the outcomes, plus implementation notes (i.e., points raised by the primary study authors about either the intervention or the evaluation that might be informative for future work in this area). Because we had no advance criteria for which outcomes to include or what implementation issues might be raised in the primary studies, these were 'rolling lists' that were continually added to as new outcomes or issues were reported in the primary studies.

In all cases, the final follow-up measure was used, as we were more interested in long-term outcomes. We had anticipated that we would calculate effect sizes to facilitate the analysis but, as we explain in the analysis section, this was considered not to be appropriate for this review.

Figure 1 below shows the three main types of quantitative data that we considered and how this relates to the inclusion criteria.

Figure 1: Types of quantitative data that we considered and how this relates to the inclusion criteria
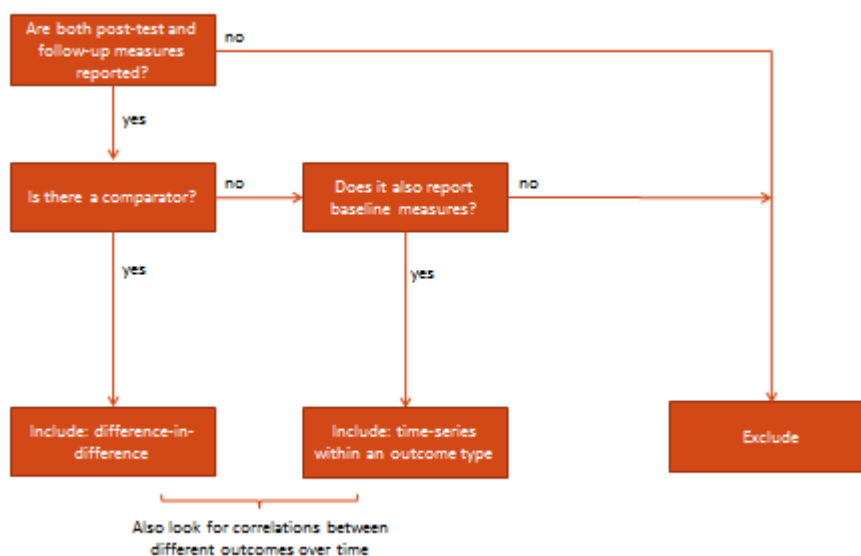
We had initially anticipated using the Methodological Quality Rating Scale (MQRS) as used in the Bowen and Neill (2013) meta-analysis for assessing the methodological rigour of the included studies. However, it became apparent from interacting with the literature that this was not an appropriate tool for either the aims of this review or the literature that was ultimately included. Our review did not aim to assess the magnitude of effect, and therefore indicators of precision were not as critical establishing whether changes in outcome were caused by the intervention itself versus other factors (e.g., the participants maturing, or receiving attention or another factor). Whilst almost all of the included studies rated at least as "moderate" on the MQRS, our assessment indicated that the risk of bias in terms of the causal claims that we were hoping to establish was greater than this.

In other words, in a meta-analysis that deals in magnitude of effects, the critical appraisal should emphasise risks to precision (i.e., how confident we are about the estimated magnitude). In a narrative review that focuses on significant relations, then critical appraisal must emphasise risks to causal attributions.

We therefore adapted the MQRS tool by selecting the indicators of most relevance to this study and adding some new indicators:

Critical appraisal indicators

- Overall score MQRS [Scores assigned by Bowen & Neill (2013): "The MQRS contains 12 items for rating a study, including design, quality control, follow-up length, and attrition. Ratings closer to 0 indicate poorer methodological quality, while ratings closer to 16 indicate better quality."
    - Low score (1-5)
    - Moderate (6-11)
    - High Score (12-16)
- Control group type
    - Waitlist control
    - Other adventure programme
    - Alternative intervention
    - Inactive control
    - "Normal population" reference group
    - No comparator
    - Unclear
- Randomisation (for studies with control group)
    - Randomised
    - Not randomised
    - No comparator
    - Unclear
- Groups (for studies with control group)
    - Equivalent at baseline
    - Not equivalent at baseline (specify)
    - Not relevant (no comparator)
    - Unclear
- Follow-up rate at last measurement
    - >=70%
    - <70%
    - Unclear amount, but stated problems with missing data
    - Unclear
- Dropouts Discussed

- Appropriate analysis
- Fidelity to intervention
  - Changes made throughout programme
  - Unstated/ unclear
- No long-term follow-up results for control

Studies were then considered to have '**substantial risk of bias'** or **'low risk of bias'** depending on the following factors:

- Design: studies without a control group (n = 8) were automatically considered to be at risk of bias for addressing our research question. This was because, without a control group, causal claims are much more tenuous.
- For controlled trials: studies with high attrition (i.e., people not completing the final follow-up measure) or differences between the intervention and the control group before the intervention begins were considered to be at higher risk of bias. This is because having non-equivalent groups can affect whether an intervention effect is observed or not.

D. Analysis

Given that this review aimed to identify outcomes that might be of interest to measure at the short-term because of their association with longer-term benefits, meta-analysis was deemed unnecessary (and also would have been unreliable on the literature that we found, because so few outcomes were reported in multiple studies). Instead, *we produced a table that reports the outcomes that have a demonstrable beneficial link between the short-term and the long-term measurement alongside the strength of evidence supporting that link.*

We adopted a similar approach to the NICE Public Health Methods Guidance[iv] for determining the overall strength of evidence for each outcome. The overall strength of evidence encompassed three elements:

1. The volume of research supporting the link between outcomes (quantity)

2. The potential risk of bias in studies supporting the link (quality)

3. The volume of research that did not support the link (consistency)

The possible overall strength of evidence for each outcome (across all studies reporting that outcome) could be classified as one of the following[4]:

- **low strength** – no strong evidence / conclusions from low risk of bias studies, with supportive evidence / conclusions from a small number of studies rated as having a substantial risk of bias;
- **moderate strength** – tentative evidence / conclusions from low risk of bias studies or supportive evidence / conclusions from numerous studies rated as having a substantial risk of bias;
- **strong evidence** – clear conclusions from multiple studies with low risk of bias that are not contradicted by other low risk of bias studies.

---

[4] NB. These three categories are quite different from those in the NICE (2009) methods manual. The threshold was lowered slightly for this study because the purpose is not to identify whether an intervention is effective, but rather to get ideas about which outcomes might be useful to measure.

Bartůněk, D., Neuman, J., Martin, A.J. (2008). Applying meta-analysis to Czech outdoor research. *Horizons. 43*, pp. 29–31.

Bowen, D.J., Neill, J.T. (2013). A Meta-Analysis of Adventure Therapy Outcomes and Moderators. *The Open Psychology Journal. 6*, pp. 28-53

Cason, D.R., Gillis, L.H. (1994). A meta-analysis of outdoor adventure programming with adolescents. *Journal of Experiential Education. 17*(1), pp. 40-7.

Coalter, F., Dimeo, P., Morrow, S., Taylor, J. (2010). *The benefits of mountaineering and mountaineering related activities: a review of literature.* Stirling, UK: University of Stirling.

Cooley, S., Burns, V., Cumming, J. (2015). The role of outdoor adventure education in facilitating groupwork in higher education. *Higher Education*. 69, pp. 567-582.

Durlak, J.A., Weissberg, R.P., Pachan, M. (2010). A meta-analysis of after-school programs that seek to promote personal and social skills in children and adolescents. *American Journal of Community Psychology. 45*, pp. 294–309.

Gillis, L.H., Speelman, E. (2008). Are challenge (ropes) courses an effective tool? A meta-analysis. *Journal of Experiential Education. 31*(2), pp. 111-35.

Hans, T.A. (2000). A meta-analysis of the effects of adventure programming on locus of control. *Journal of Contemporary Psychotherapy. 30*(1), pp. 33-60

Hattie, J., Marsh, H.W., Neill, J.T., Richards, G.E. (1997). Adventure Education and Outward Bound: Out-of-class experiences that have a lasting effect. *Review of Educational Research. 67*, pp. 43-87.

Higgins, S., Katsipataki, M., Kokotsaki, D., Coe, R., Major, L.E., Coleman, R. (2013). The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit: Technical Appendices.

Laidlaw, J.S. (2000). *A meta-analysis of outdoor education programs.* Unpublished doctoral dissertation. Greeley, CO: University of Northern Colorado.

McKenzie, M.D. (2000). How are adventure education program outcomes achieved?: A review of the literature. *Australian Journal of Outdoor Education*. 5, pp. 19-28.

Neill, J.T., Richards, G.E. (1998). Does outdoor education really work? A summary of recent meta-analyses. *Australian Journal of Outdoor Education. 3*, pp. 1-9.

Neill, J.T. (2003). Reviewing and benchmarking adventure therapy outcomes: Applications of meta-analysis. *Journa of Experiential Education. 25*(3), pp. 316-21.

Neill, J. (2008). Meta-Analytic Research on the Outcomes of Outdoor Education. 6th Biennial Coalition for Education in the Outdoors Research Symposium, Bradford Woods, IN, 11-13 January, 2002.

Neill, J.T. (2009). Outdoor education meta-analyses. [Online] http://wilderdom.com/research/ResearchReviewsMetaanalysis.html [Accessed: 22.03.16].

Reddrop, S. (1997). *Outdoor programs for young offenders in detention: an overview.* Hobart, Australia: National Clearing House.

Rickinson, M., Dillon, J., Teamey, K., Morris, M., Choi, M.Y., Sanders, D., Benefield, P. (2004). *A review of research on outdoor learning.* Published by Field Studies Council: OP87, ISBN 1 85153 893 3.

SMCI Associates (2013). 'Living Wild - Chance for Change' Highland LEADER Programme: Evaluation Report. East Lothian: SMCI Associates.

Stott, T., Allison, P., Felter, J., Beames, S. (2013). Personal development on youth expeditions: A literature review and thematic analysis. Leisure Studies.

Wilson, S.J., Lipsey, M.W. (2000). Wilderness challenge programs for delinquent youth: A meta-analysis of outcome evaluations. *Evaluation and Program Planning. 23*(1), pp. 1-12.

Appendix 4: Reasons for excluding primary studies at the second stage (based on full-text documents) (n = 26)

Exclusion criterion	Number of studies
Participants: Not 12-25 year olds	2
Participants: Not population (at-risk, NEET, etc.)	8
Intervention: Not adventure programme	2
Research design: Not controlled trial or longitudinal	5
Follow-up measures: Not at least 13 weeks/ 3 months after the intervention ends	8
Irretrievable	1
Total	*26*

Bocarro, J. (1998) *An examination of the efficacy of an adventure-based counseling program on at-risk youth*. Master's thesis: Dalhousie University.

Brand, D. (1998) *A longitudinal study of behaviour disordered adolescents and the effects on them of a wilderness-enhanced program*. PhD: University of Wollongong .

Crisp, S.J.R., Hinch, C. (2003) *SWATRAD & GO WEST Master March 2003 [Data file].* Melbourne: Unpublished dataset, Neo Psychology.

Freedman, E.B. (1996) *The effects of a therapeutic wilderness experience for emotionally disturbed adolescents*. PhD: California School of Professional Psychology.

Garst, B., Scheider, I., Baker, D. (2001) Outdoor adventure program participation impacts on adolescent self-perception. *Journal of Experiential Education*. 24(1), pp 41-9.

Grayson, R.C. (2001) *Summer camp as an intervention for atrisk youth*. PhD: Claremont Graduate University .

Harper, N.J., Russell, K.C., Cooley, R., Cupples, J. (2007) Catherine Freer Wilderness Therapy Expeditions: An exploratory case study of adolescent wilderness therapy, family functioning, and the maintenance of change. *Child Youth Care Forum*. 36, pp. 111-29.

Jelalian, E., Mehlenbeck, R., Lloyd-Richardson, E.E., Birmaher, V., Wing, R.R. (2006) 'Adventure therapy' combined with cognitive-behavioral treatment for overweight adolescents. *International Journal of Obesity*. 30(1), pp. 31-9.

Lewis, S.F. (2012) Examining changes in substance use and conduct problems among treatment-seeking adolescents. *Child and Adolescent Mental Health*. pp. 1-6.

Martinez, M. (2002) *A wilderness therapy program for a diverse group of at risk adolescent boys: Changes in self-esteem and locus of control and their relationship to group affiliation*. PhD: Alliant International University.

McGarvey, A.L. (2004) *An evaluation of a ropes course: Efficacy for at-risk youth with externalizing versus internalizing symptoms*. PhD: University of Connecticut .

Michalski, J.H., Mishna, F., Worthington, C., Cummings, R. (2003) A multi-method impact evaluation of a therapeutic summer camp program. *Child and Adolescent Social Work Journal*. 20(1), pp. 53-76.

Mohr, P., Heseltine, K., Howells, K., Badenoch, D., Williamson, P., Parker, A. (2001) *Evaluation of Operation Flinders wilderness adventure program for youth at risk*. Adelaide: The Forensic and Applied Psychology Research Group.

Parker, M.W., Stoltenberg, C. (1995) Use of adventure experiences in traditional counseling interventions. *Psychological Reports*. 77(32), pp. 1376-8.

Pommier, J.H., Witt, P.A. (1995) Evaluation of an Outward Bound School plus Family Training Program for the Juvenile Status Offender. *Therapeutic Recreation Journal*. 29, pp. 86-103.

Russell, K.C. (2002) *Longitudinal assessment of treatment outcomes in outdoor behavioral healthcare*. Moscow: University of Idaho.

Appendix 6: Characteristics of included studies

Study author and year	Total Sample Size	Country	Length of intervention	Intervention type	Length of follow-up	Population focus	Study design
Bocarro 1998	25	Canada	21 days	*Multiple:* Adventure-based counselling	4.5 months	At-risk	Longitudinal
Brand 1998	190	Australia	10 days	*Expedition:* wilderness-enhanced	21 months	Behaviour-disordered	Comparison
Crisp & Hinch 2003	52	Australia	70 days	*Multiple:* Wilderness Adventure Therapy	Clinical groups: 3 months, 24 months; At-risk groups: 3 months	Mental health	Longitudinal
Freedman 1996	61	USA	13 days	*Expedition:* therapeutic wilderness experience	3 months	Emotionally-disturbed	Comparison
Garst et al. 2001	58	USA	3 days	*Expedition:* Outdoor adventure program	4 months	At-risk	Longitudinal
Grayson 2001	176	USA	21 days	*Base camp:* Summer camp	3 months	At-risk	Comparison
Harper et al. 2007	221	USA	21 days	*Expedition:* Catherine Freer Wilderness Therapy Expeditions	12 months	Behaviour-disordered; Emotionally-disturbed; Substance abuse	Longitudinal
Jelalian et al. 2006	76	USA	112 days	*Ropes-/Challenge-/Adventure-based:* 'Adventure therapy' combined with cognitive-behavioral treatment	6 months	Physical	Comparison
Lewis 2012	190	USA	57 days	*Multiple:* Outdoor behavioral healthcare	12 months	Substance abuse	Longitudinal
Martinez 2002	223	USA	13 days	*Base camp:* wilderness therapy program	9 months	At-risk	Longitudinal
McGarvey 2004	44	USA	49 days	*Ropes-/Challenge-/Adventure-based:* Ropes course	7 months	At-risk; Behaviour-disordered; Educationally-disengaged	Comparison
Michalski et al. 2003	96	Canada	21 days	*Ropes-/Challenge-/Adventure-based:* therapeutic summer camp program	6-8 months	Mental health; Educationally-disengaged	Longitudinal
Mohr et al. 2001	102	Australia	8 days	*Expedition:* Operation Flinders wilderness adventure program	3 months	At-risk	Comparison
Parker & Stoltenberg 1995	84	USA	2 days	*Ropes-/Challenge-/Adventure-based:* Adventure experiences in traditional counseling interventions	6 months	At-risk; Behaviour-disordered	Comparison

Study author and year	Total Sample Size	Country	Length of intervention	Intervention type	Length of follow-up	Population focus	Study design
Pommier & Witt 1995	79	USA	28 days	*Multiple:* Outward Bound School plus Family Training Program	3 months	Adjudicated youth	Comparison
Russell 2002	858	USA	45 days	*Multiple:* Outdoor behavioral healthcare	12 months	At-risk	Longitudinal

Appendix 7: Data extracted by Bowen & Neill (2013)

1. Publication year
2. Publication type
3. Mean age and age category
4. Sample sizes (total, and separately for intervention and control group)
5. Gender (number of participants of each)
6. Country
7. Race
8. Identified population (at-risk, clinical, adjudicated)
9. Identified issue
    1. Abuse Victims (Physical, Emotional or Sexual)
    2. Adjudicated Youth
    3. Behaviour Disordered
    4. Disabilities
    5. Educationally Disengaged
    6. Emotionally Disturbed
    7. Families
    8. Mental Health
    9. Mixed
    10. Physical (e.g., Brain Injury, Weight-Loss, etc.)
    11. Substance Abuse
    12. Welfare
10. Funding type (private or public)
11. Use of adventure as the therapeutic mode (primary or adjunctive)
12. Program delivery (continuous or intermittent)
13. Group structure (closed or open)
14. Placement type (private of adjudicated)
15. Program type (contained, continuous-flow, base-camp, residential, mixed)
16. Program model (base camp, expedition, residential, outpatient, multiple, ropes / challenge / adventure-based)
17. Daily duration
18. Program length
19. Methodological Quality Rating Scale (MQRS)
    1. Study Design
    2. Follow-up rate
    3. Collaterals Interviewed
    4. Objective Verification of Self-Report Data
    5. Dropouts Discussed
    6. Appropriate Analysis

POPULATION

- Mean age known
- Gender
  - All male
  - All female
  - Mixed (add details if known)
  - Unclear
- Race
  - 60% Caucasian
  - 60% Minority
  - Mixed, No Race > 60%
  - Unclear
- Identified focus
  - At-risk
  - Behaviour-disordered
  - Mental health
  - Emotionally-disturbed
  - Physical
  - Substance abuse
  - Educationally-disengaged
  - Adjudicated youth

INTERVENTION

- Funding
  - Private
  - Public
- Programme model
  - Base camp
  - Expedition
  - Outpatient
  - Multiple
  - Ropes / Challenge / Adventure-based
- Programme length (days)

DESIGN

- Comparison
- Longitudinal

STUDY CHARACTERISTICS

- Publication type
  - Thesis
  - Journal article
  - Other
- Country
  - Australia
  - USA
  - Canada
- Time to follow-up

Journal analysis

Thomas, G., Potter, T., Allison, P. (2009) A Tale of three journals: An investigation of the development and futures of AJOE, JAEOL and JEE. *Australian Journal of Outdoor Education.* 13, pp. 16-29.

UK surveys

Glackin, M., Jones, B. (2012) Park and Learn: Improving Opportunities for Learning in Local Open Spaces. *School Science Review*. 93, pp. 105-113.

Taylor, C., Power, S., Rees, G., Taylor, C. (2009) *Out-of-school learning: the uneven distribution of school provision and local authority support*. United Kingdom.

Tilling, S. (2004) Fieldwork in UK Secondary Schools: Influences and Provision. *Journal of Biological Education*. 38, pp. 54-58.

Waite, S. (2011) *Teaching and learning outside the classroom: personal values, alternative pedagogies and standards*. United Kingdom.

Systematic reviews

Bowen, D.J., Neill, J.T. (2013) A Meta-Analysis of Adventure Therapy Outcomes and Moderators. *The Open Psychology Journal.* 6, pp. 28-53.

Cason, D., Gillis, H. (1994) A meta-analysis of outdoor adventure programming with adolescents. *Journal of Experiential Education.* pp. 40-47.

Coalter, F., Dimeo, P., Morrow, S., Taylor, J. (2010) *The Benefits of Mountaineering and Mountaineering Related Activities: A Review of Literature: A Report to the Mountaineering Council of Scotland. Mountaineering Council of Scotland*. British Mountaineering Council. Department of Sports Studies, University of Stirling.

Cooley, S., Burns, V., Cumming, J. (2015) The role of outdoor adventure education in facilitating groupwork in higher education. *Higher Education.* 69, pp. 567-582.

Davies, D., Jindal-Snapeb, D., Collier, C., Digbya, R. (2013) Creative Learning Environments in Education – a systematic literature review. *Thinking Skills and Creativity.* 8, pp. 80-91.

Gill, T. (2011) *Children and nature. A Quasi-systematic review of the empirical evidence*. Greater London Authority.

Gillis, L.H., Speelman, E. (2008) Are Challenge (Ropes) Courses an Effective Tool? A Meta-Analysis. *Journal of Experiential Education*. 31, pp. 111-135.

Hattie, J., Marsh, H. W., Neill, J.T, Richards, G.E. (1997) Adventure education and outward bound: Out-of-class experiences that make a lasting difference. *Review of Educational Research*. 67, pp. 43-87.

Higgins, S., Katsipataki, M., Kokotsaki, D., Coe, R., Major, L.E., Coleman, R. (2013) *The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit: Technical Appendices*.

Jill Dando Institute (2015) *Wilderness Challenge Programmes*.

Lock, R. (2010) Biology Fieldwork in Schools and Colleges in the UK: An Analysis of Empirical Research from 1963 to 2009. *Journal of Biological Education*. 44, pp. 58-64.

Neill, J. (2008) Meta-Analytic Research on the Outcomes of Outdoor Education. *6th Biennial Coalition for Education in the Outdoors Research Symposium*, Bradford Woods, IN, 11-13 January, 2002.

Neill, J. (2008) *Enhancing life effectiveness. The impacts of outdoor education programs*. University of Western Sydney.

Rickinson, M., Dillon, J., Teamey, K., Morris, M., Choi, M.Y., Sanders, D., Benefield, P. (2004) *A review of research on outdoor learning*. National Foundation for Educational Research.

SMCI Associates (2013) *'Living Wild - Chance for Change' Highland LEADER Programme: Evaluation Report.* East Lothian: SMCI Associates.

Stott, T., Allison, P., Felter, J., Beames, S. (2015) Personal development on youth expeditions: A literature review and thematic analysis. *Leisure Studies* 34(2), pp. 197-229.

UK primary studies

Amos, R., Reiss, M. (2012) The Benefits of Residential Fieldwork for School Science: Insights from a Five-Year Initiative for Inner-City Students in the UK. *International Journal of Science Education*. 34, pp. 485-511.

Barry, M. (2010) *An Evaluation of Venture Trust's Moving On – Willow Partnership Project*. University of Strathclyde, UK.

Beames, S. (2004) Overseas youth expeditions with Raleigh International: a rite of passage*? Australian Journal of Outdoor Education*. 8, pp. 29-36.

Beames, S. (2005) Expeditions and the social construction of the self. *Australian Journal of Outdoor Education.* 9, pp. 14-22.

Birdwell, J., Wybron, I. (2014) *Scouting for Skills*. Demos, London.

Bowey, L., McGlaughlin, A. (2006) The Youth Crime Reduction Video Project: An Evaluation of a Pilot Intervention Targeting Young People at Risk of Crime and School Exclusion. *Howard Journal of Criminal Justice*. 45, pp. 268-283.

Bowker, R. (2004) Children's Perceptions of Plants Following Their Visit to the Eden Project. *Research in Science and Technological Education*. 22, pp. 227-243.

Brewer, J., Sparkes, A.C. (2011) The Meanings of Outdoor Physical Activity for Parentally Bereaved Young People in the United Kingdom: Insights from an Ethnographic Study*. Journal of Adventure Education and Outdoor Learning.* 11, pp. 127-143.

Brown, D., Neale, N., Francis, R. (2011) Peak into the Past: An Archaeo-Astronomy Summer School*. School Science Review.* 93, pp. 83-90.

Christie, B., Higgins, P., McLaughlin, P. (2014) 'Did you enjoy your holiday?' Can residential outdoor learning benefit mainstream schooling? *Journal of Adventure Education & Outdoor Learning*. 14, pp. 1-23.

Cooley, S., Holland, M., Cumming, J., Novakovic, E., Burns, V. (2014) Introducing the use of a semi-structured video diary room to investigate students' learning experiences during an outdoor adventure education groupwork skills course. *Higher Education*. 67, pp. 105-121.

Cramp, A. (2008) Knowing Me Knowing You: Building Valuable Relationships outside the Classroom. *Education* 3-13. 36, pp. 171-182.

Crosbie, J. (2014) *The value of outdoor education for people with disabilities: An in-depth case study of the Calvert Trust.* Doctoral/PhD. University of Edinburgh.

Davis, B., Waite, S. (2005) *Forest Schools: an evaluation of the opportunities and challenges in Early Years Final Report*. University of Plymouth.

Dillon, J., Morris, M., O'Donnell, L., Raid, A., Rickinson, M., Scott, W. (2005) *Engaging and Learning with the Outdoors – The Final Report of the Outdoor Classroom in a Rural Context Action Research Project*. National Foundation for Education Research, Slough.

Dismore, H., Bailey, R. (2005) "If Only": Outdoor and Adventurous Activities and Generalised Academic Development. *Journal of Adventure Education and Outdoor Learning*. 5, pp. 9-19.

Dourneen, J.(2010) *Setting free the spirit of English in the Brecon Beacons: the value of learning outside the classroom*. United Kingdom.

Fabian, H. (2005) *Outdoor learning environments: easing the transition from the foundation stage to Key Stage One*. United Kingdom.

Fielding, S., Jones, M. (2014) Children Go Bonkers about Bugs. *Primary Science.* pp. 23-27.

Fox, P., Avramidis, E. (2003) An evaluation of an outdoor education programme for students with emotional and behavioural difficulties. *Emotional & Behavioural Difficulties*. 8, pp. 267-283.

Gerret, K. (2014) *Cedar Road Primary School Aberdovey Analysis* 2013 / 2014.

Halocha, J. (2005) Developing a Research Tool to Enable Children to Voice Their Experiences and Learning through Fieldwork. *International Research in Geographical and Environmental Education*. 14, pp. 348-355.

Heeley, S. (2013) *The Value of Outdoor Education For Young People With Learning Difficulties: A Case Study of Arthog Outreach And Southall School*. [dissertation for BA Honours degree].

Hinds, J. (2011) Woodland Adventure for Marginalized Adolescents: Environmental Attitudes, Identity and Competence. *Applied Environmental Education and Communication*. 10, pp. 228-237.

Horseman, L. (2010) *Reflections and Evaluation on Forest School projects 2010.* Kindling.

Humberstone, B., Stan, I. (2009) Well-being and outdoor pedagogies in primary schooling: the nexus of well-being and safety. *Australian Journal of Outdoor Education*. 13, pp. 24-32.

Humberstone, B., Stan, I. (2011) Outdoor Learning: Primary Pupils' Experiences and Teachers' Interaction in Outdoor Learning. *Education* 3-13. 39, pp. 529-540.

Humberstone, B., Stan, I. (2012) Nature and Well-Being in Outdoor Learning: Authenticity or Performativity. *Journal of Adventure Education and Outdoor Learning*. 12, pp. 183-197.

Loynes, C. (2010) Journeys of Transition: The Role of Narrative within the Stoneleigh Project. *Journal of Adventure Education and Outdoor Learning*. 10, pp. 127-145.

Maynard, T., Waters, J. (2007) Learning in the Outdoor Environment: A Missed Opportunity?. *Early Years: An International Journal of Research and Development*. 27, pp. 255-265.

Maynard, T., Waters, J., Clement, J. (2013) Child-Initiated Learning, the Outdoor Environment and the "Underachieving" Child. *Early Years: An International Journal of Research and Development*. 33, pp. 212-225.

McArdle, K., Harrison, T., Harrison, D. (2013) Does a Nurturing Approach That Uses an Outdoor Play Environment Build Resilience in Children from a Challenging Background?. *Journal of Adventure Education and Outdoor Learning*. 13, pp. 238-254.

McManus, J. (2012) The Thurston Family Project: Working with families through outdoor activities and resiliency training. *Psychology of Education Review*. 36, pp. 40-45.

Newman, M., Higgins, P., McLaughlin, P., Allison, P. (2004) Outdoor Experiential Education Programmes for Socially Excluded Young People An Evaluation of the Venture Trust's 10-day Pilot Programme.

North Lanarkshire Council and the Outward Bound Trust (2014) *North Lanarkshire Council and the Outward Bound Trust Interim Evaluation. 2013-14*. The Outward Bound Trust.

O'Brien, L. (2009) Learning Outdoors: The Forest School Approach. *Education* 3-13. 37, pp. 45-60.

O'Brien, K. (2014) *Developing a Growth Mindset through Outdoor Personal Development: Can an intervention underpinned by psychology increase the impact of an outdoor learning course for young people?* [MSc dissertation]

PACEC (2011) *Impact assessment evaluation of The Scout Association*. 84.

Peacock, A. (2006) *Changing Minds The lasting impact of school trips. A study of the long-term impact of sustained relationships between schools and the National Trust via the Guardianship scheme*. University of Exeter.

Price, A. (2015) Improving school attendance: Can participation in outdoor learning influence attendance for young people with social, emotional and behavioural difficulties? *Journal of Adventure Education and Outdoor Learning*. 15, pp. 110-122.

Heaven, A., Browne, A. (2008) *A small scale evaluation of existing practice, monitoring systems and outcomes for vulnerable children.* Lindley Educational Trust.

Riley, K., Ellis, S., Weinstock, W., Tarrant, J., Hallmond, S. (2006) Re-engaging disaffected pupils in learning: insights for policy and practice. *Improving Schools*. 9, pp. 17-31.

Kenny, R. (2010) *Forest School and the Early Years Foundation Stage – An explorative case study*.

Sandford, R.A., Duncombe, R., Armour, K.M. (2008) The Role of Physical Activity/Sport in Tackling Youth Disaffection and Anti-Social Behaviour. *Educational Review*. 60, pp. 419-435.

Sandford, R.A., Armour, K.M., Stanton, D.J. (2010) Volunteer Mentors as Informal Educators in a Youth Physical Activity Program. *Mentoring & Tutoring: Partnership in Learning*. 18, pp. 135-153.

Scrutton, R.A. (2015) Outdoor adventure education for children in Scotland: Quantifying the benefits. *Journal of Adventure Education and Outdoor Learning*. 15, pp. 123-137.

SMCI (2013) *'Living Wild – Chance for Change' Highland LEADER Programme*. SMCI Associates.

Sharpe, D. (2014) Independent Thinkers and Learners: A Critical Evaluation of the "Growing Together Schools Programme". *Pastoral Care in Education*. 32, pp. 197-207.

Slade, M., Lowery, C., Bland, K. (2013) Evaluating the Impact of Forest Schools: A Collaboration between a University and a Primary School. *Support for Learning*. 28, pp. 66-72.

Social Finance, Adviza Partnership (2014) E*nergise Innovation Limited a Year in Review October 2012 – October 2013*.

Sproule, J., Martindale, R., Wang, J., Allison, P., Nash, C., Gray, S. (2013) Investigating the Experience of Outdoor and Adventurous Project Work in an Educational Setting Using a Self-Determination Framework*. European Physical Education Review*. 19, pp. 315-328.

Stott, T., Hall, N. (2003) Changes in aspects of students' self-reported personal, social and technical skills during a six week wilderness expedition in Arctic Greenland. *Journal of Adventure Education and Outdoor Learning.* 3(2), pp. 159-169.

Waller, T. (2007) The Trampoline Tree and the Swamp Monster with 18 Heads: Outdoor Play in the Foundation Stage and Foundation Phase. *Education* 3-13. 35, pp. 393-407.

White, R. (2012) A Sociocultural Investigation of the Efficacy of Outdoor Education to Improve Learner Engagement. *Emotional & Behavioural Difficulties.* 17, pp. 13-23.

Williams, R. (2013) Woven into the Fabric of Experience: Residential Adventure Education and Complexity. *Journal of Adventure Education and Outdoor Learning.* 13, pp. 107-124.

Williams, R. (2013) *Report to The Outward Bound Trust on an Evaluation of The Classic Course.*

York Consulting (2015) *Paul Hamlyn Foundation Evaluation of Learning Away: Final Report.*

Other References

[i] JPAL. Impact Evaluation: Why randomize? Slide 36. [Online] http://ocw.mit.edu/resources/res-14-002-abdul-latif-jameel-poverty-action-lab-executive-training-evaluating-social-programs-2011-spring-2011/lecture-notes/MITRES_14_002S11_lec3.pdf [Accessed: 22.03.16].

[ii] Morrison, L.M., Schoon, I. (2013) *The impact of non-cognitive skills on outcomes for young people Literature review*. Education Endowment Foundation. [Online] https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Non-cognitive_skills_literature_review.pdf [Accessed: 22.03.16].

[iii] O'Mara-Eves, A., Fiennes, C., Oliver, S. (2016) Long-term outcomes of adventure, outdoor, and experiential education: Protocol for a review of existing evidence. [Online] http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3477 [Accessed: 22.03.16].

[iv] NICE (2009) *Methods for the development of NICE public health guidance (second edition)*.

Coe, R. (2012) It's the Effect Size, Stupid: What effect size is and why it is important [Online] http://www.leeds.ac.uk/educol/documents/00002182.htm [Accessed: 22.10.2015].

Goldacre, B. (2012) *Bad Science*. Fourth Estate.

Hoffmann, T.C., et al. (2014) Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *The British Medical Journal*. 348.

Kent County Council. (2014) HeadStart Kent [Online] http://www.kelsi.org.uk/__data/assets/powerpoint_doc/0010/25588/Programme-Board-Main-Presentation-19-11-14-AF-UN.pptx [Accessed: 22.10.15].

Krumholz, H. (@hmkyale) "You can't just combine weak evidence and pretend that when mushed together it is strong. It is meta-mush." 30 August 2015. Tweet. [Online] https://twitter.com/hmkyale/status/638013905055444992 [Accessed: 22.10.15].

The Lancet (2014). *Research: increasing value, reducing waste* [Online] www.thelancet.com/series/research [Accessed 22.10.15].

National Audit Office (2013) *Evaluation in Government.*

Research Rundowns. *Effect Size*. [Online] https://researchrundowns.wordpress.com/quantitative-methods/effect-size/ [Accessed: 22.10.15].

Yusuf, S., Collins, R., Peto, R. (1984) Why do we need some large, simple randomized trials? *Statistical Medicine* 3(4), pp. 409-22.