Journal of Experiential Education
2015, Vol. 38(1) 8–25
© The Authors 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1053825913514730
jee.sagepub.com
SAGE

# Measuring the Unmeasurable: Upholding Rigor in Quantitative Studies of Personal and Social Development in Outdoor Adventure Education

**Roger Scrutton[1] and Simon Beames[1]**

**Abstract**

Outdoor adventure education (OAE) has a long history of being credited with the personal and social development (PSD) of its participants. PSD is notoriously difficult to measure quantitatively, yet stakeholders demand statistical evidence that given approaches to eliciting PSD are effective in their methods. Rightly or wrongly, many stakeholders want evidence in numbers and not anecdote. This article examines 22 recent studies and 6 meta-analyses in the peer-reviewed literature that have made use of quantitative methods to measure PSD in OAE. The investigation revealed three recurring problem areas that are effectively weakening the body of literature: the sample, the questionnaire and its use, and management of variables. Examples of good practice are reviewed and concrete suggestions made to uphold methodological rigor.

**Keywords**

quantitative, development, adventure, outdoor

Outdoor adventure education (OAE) has a long and rich history of being credited with the personal and social development (PSD) of its participants. PSD is particularly difficult to measure quantitatively in any field (indoors or out, adventurous or not),yet stakeholders of all kinds—not least funders, politicians, and program planners—

[1]University of Edinburgh, UK

**Corresponding Author:**
Roger Scrutton, Moray House School of Education, University of Edinburgh, Edinburgh EH8 8AQ, UK.
Email: roger.scrutton@ed.ac.uk

understandably demand "proof" in the form of statistical analysis that given approaches to eliciting PSD are effective in their methods.

This article's rationale lies in our suspicion that over recent years there has been a growing body of rich qualitative literature reporting strong outcomes from OAE programs, while at the same time, there has been doubt over the robustness of evidence coming from the quantitative domain. To more deeply and systematically examine questions surrounding the objective measurement of what can be loosely termed "personal growth," we decided to critique a number of journal papers that used quantitative methods to do just this. This investigation, therefore, aims to consider ways in which quantitative measures of benefit can be conducted rigorously and, thus, be of greater service to the field of OAE. We do this by highlighting the difficulties of defining PSD, considering recurring aspects of problematic research, and offering constructive suggestions for upholding the rigor of quantitative measurements of PSD through OAE.

## PSD in the Outdoor Adventure Context

PSD is an idea that has evolved out of an early 20th-century tradition of "character building," which is a rather bullish name for something that might be interpreted as "pointlessly unpleasant or difficult" (Brookes, 2003, p. 49). In the context of OAE, the term "personal and social development" seems to have emerged in the late 1980s (Hopkins & Putnam, 1993), although there had been growth in OAE in schools, youth organizations, and outdoor centers to promote PSD for some time before that. Nowadays, PSD describes a process that is arguably much broader and more sensitive to the needs of individuals and society in general.[1]

It would be most helpful to have a universally accepted definition of PSD for the purposes of measuring it, but it is very difficult to find one. More than 30 years ago, Pring (1982), writing on PSD in the school curriculum, noted that "it seems impossible to make any coherent sense of it as an area" (p. 4). Much more recently, a review of the role of PSD in education and employment concluded that it is a "slippery" concept with "soft boundaries" that means slightly different things to different people (Young Adult Learners Partnership, 2003). Nevertheless, out of a perceived need to assess PSD within the school curriculum, Buck and Inman (1998) defined it as having three parts: knowledge and understanding, skills development, and attitudinal development.

Despite the multi-faceted and elusive nature of PSD, it appears to comprise fundamental elements which OAE is seen to benefit. Allison and von Wald (2010), writing on expeditions, note that "PSD is often understood as developing confidence, cooperation, trust and teamwork. Talking with people in outdoor education suggests that self-esteem is regularly identified as central to PSD" (p. 57). The review of research in outdoor learning by Rickinson and colleagues (2004) suggests that outdoor adventure impacts positively on a set of personal and social skills that include locus of control and coping strategies (e.g., overcoming fear; p. 6). A comprehensive list of possible PSD outcomes from OAE, according to the English Outdoor Council (2005), includes

the more commonly cited parameters of confidence, resilience, self- and social-awareness, cooperation, trust, and teamwork, but also identifies a positive attitude to challenge and adventure, environmental awareness, the acquisition of outdoor skills, health and fitness, and increased motivation toward learning.

Thus, there are certain elements of PSD, as it is understood today in its broader form, that are commonly identified as the stated outcomes from courses of outdoor adventure activities, and it is these elements that require trustworthy and dependable measurement. As measurement is commonly made on the basis of participants' self-assessment (usually by completing a questionnaire before and after an intervention), dependable measurement relies on both researchers and participants having a common understanding of PSD.

## Methodological Criticism

Criticism of the application of quantitative research methods in outdoor education is not new. In a major review of research by Barrett and Greenaway published in 1995, it was concluded that the efforts in the United Kingdom to evaluate the effectiveness of outdoor adventure in PSD tended to be "isolated," "inconclusive," "over-ambitious," "uncritical," "not of a high standard," and "difficult to locate" (p. 53). This damning verdict on U.K. research was followed by an equally damning verdict on research outside the United Kingdom. Although their review was primarily concerned with understanding the *processes* of PSD through OAE, they emphasized that before getting to that point, researchers had to be capable of measuring *outcomes*; measurement methodology was a specific target for their negative criticism.

Numerous authors have made similar, if not quite so negative, criticisms in the intervening years. In particular, Rickinson et al. (2004) recommended improving "the methodological rigour of outdoor learning research and evaluation" as well as the "research-based understandings of the outdoor learning process" (p. 56). In reviewing meta-analyses of quantitative measures of the impact of outdoor education programs on PSD, Neill (2008) noted a number of methodological weaknesses (see next section). More robust research design, richer and more precise measurement of dependent variables, and a greater emphasis on analyzing and reporting the effects of (assumed) independent variables were mentioned specifically. While it is healthy for researchers to be aware of limitations in their work, this widespread and ongoing negative criticism in a field of research is unusual and needs to be addressed—particularly if it means that other fields are looking at our body of literature and rendering similar judgments. In our field's defense, however, we note that the vast majority of evaluations made over 50 years of research reveal a picture of impact that is consistent with the wider fields of psychology and education in general (Figure 1).

Figure 1 also shows us that, when compared with wider fields with more confidence in quantitative research, the volume of published research in OAE—represented here by the number of meta-analyses—is small. It is probably premature to bemoan our inability to understand program processes using quantitative data (Barrett & Greenaway, 1995) when the literature on outcomes is so sparse and the number of variables in any one study of the impact of OAE on PSD is so great. Still, we are left

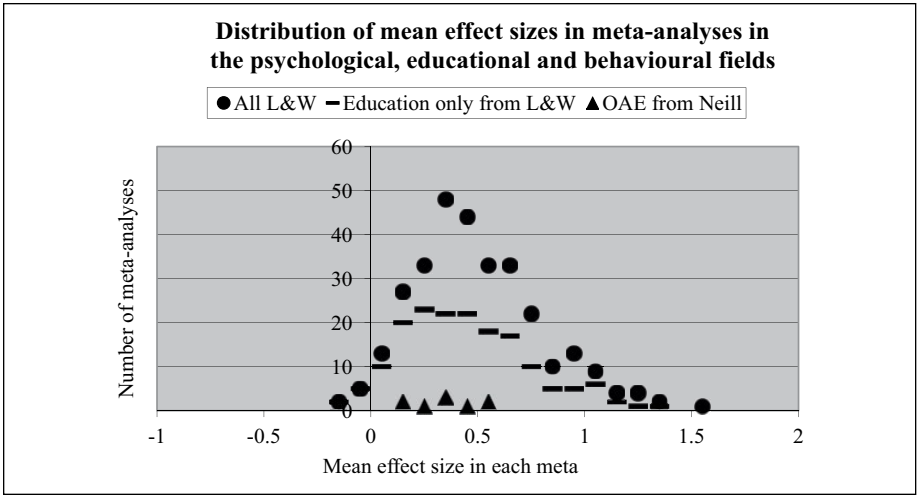


**Figure 1.** The distribution of mean Cohen effect sizes in meta-analyses.
*Note.* Dots—psychological, educational, and behavioral interventions (Lipsey & Wilson, 1993); lines—educational only (Lipsey & Wilson, 1993); triangles—outdoor adventure education (Neill, 2008). Cohen's (1988) nomenclature for effect sizes is: "small," 0.2-0.3; "medium," c.0.5; "large," c ≥ 0.8. OAE = Outdoor adventure education.

with circumstances that demand research of a high standard, but which by our own admission often produce work of low quality and risk tarnishing the reputation of the field of OAE in general.

The aforementioned elusiveness of universal agreement on what PSD actually is (e.g., Buck & Inman, 1997; Pring, 1982), combined with perceptions that empirical research to measure PSD through OAE is unconvincing (e.g., Neill, 2008; Rickinson et al., 2004), has left a key element of our work (PSD) "under proven" by the dominant research tradition (positivism). In the following sections of this article, we identify those aspects of the quantitative measurement of benefits that have attracted most criticism and discuss how the criticisms might be addressed.

## Common Problems for Quantitative Researchers

Notwithstanding the evidence in Figure 1, Rickinson et al. (2004) commented on the diversity of findings in OAE research varying widely between different types of programs and different participants. This lack of convergence contributed to their call for an improvement in methodological rigor. In another, more detailed review, Neill (2008) identified the following common problems and limitations in quantitative research in OAE:

- the usage of inappropriate self-report questionnaires,
- low statistical power,
- over-reliance on inferential statistics,

- a lack of control and comparison groups,
- a lack of longitudinal data, and
- a lack of investigation of independent variables.

Researchers, being aware of these potential weaknesses, commonly include in their publications a statement on the limitations of their methodology. To access this information, and thus pursue our aim of considering ways in which quantitative measures of benefit might be conducted rigorously, we have undertaken a survey of quantitative methodology in articles published in issues of the *Journal of Adventure Education and Outdoor Learning* and the *Journal of Experiential Education* during the last 12 years. We found 22 publications, listed in the Appendix, that reported measures of the outcomes of adventure education courses for aspects of PSD. Table 1 summarizes the limitations of each piece of research as acknowledged by the researchers themselves, as well as possible additional limitations identified by us. The table enabled us to identify the most commonly identified limitations and, in bringing these together under a few headings, we found the common problems closely matched those identified by Neill (2008).

The majority of studies in Table 1 concerned adolescents (sometimes in therapy) and college students. Two studies were concerned with school children on multi-activity outdoor adventure programs, one was on families and another on adults. Despite this diversity, there is a strong commonality among authors in identifying limitations associated with the recruitment of the sample, the use of questionnaires, including testing for their validity and reliability, survey design, and the difficulty of controlling for parameters emerging as independent variables that were not anticipated or simply ignored. In addition, there are instances where there are limitations on reporting on the distribution of scores in the data sets, such as the normal distribution of data in small samples that influences the choice of statistical tests, ceiling and floor effects, and descriptive statistics. It is likely that researchers have carried out these "methodological essentials," but we would encourage their publication as a matter of course.

Along with the 22 papers considered, we also undertook a review of published meta-analyses in the field of OAE to identify the key criteria that have been used to judge the methodological rigor of studies included in each meta-analysis (Table 2). Again, there is considerable commonality between these key criteria and the methodological limitations identified in the 22 papers in Table 1 and by Neill (2008).

## Overcoming Recognized Limitations in Quantitative Research on PSD in OAE

We now look at the three aspects of quantitative methodology in which limitations cluster and make suggestions for mitigating or overcoming these limitations.

### *Size and Composition of the Sample*

In theory, it is easy to define the sample suited to a particular study, but in practice it is often very difficult to obtain. The total sample size below which authors begin to see

**Table 1.** Limitations of Each of the 22 Studies Reviewed as Acknowledged by the Researchers Themselves (X), With Some Additional Limitations Identified by Us (0).

| | Criterion | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sample | | | Questionnaire | | Survey design | | | Presentation of results | | |
| Publication | Inadequate sample size | No use of control group | Poor composition | Inappropriate for the study | Poor or no reporting of V&R | No pre-test | Poor timing of tests | No longitudinal component | No presentation of data | Poor descriptive statistics | Other variable problems |
| Bell (2006) | | | X | | | N/A | | | 0 | | |
| Bobilya, Akey, and Mitchell (2011) | X | | | | | | | 0 | | | |
| Clark, Marmol, Cooley, and Gathercoal (2004) | | N/A | X | | | | 0 | | | | |
| Cross (2002) | X | | 0 | X | | | 0 | 0 | | | |
| Ewert and Yoshino (2011) | X | | X | | | | | | | | |
| Freeman (2002) | X | X | X | | | | 0 | | | 0 | |
| Gassner and Russell (2008) | | N/A | | 0 | 0 | N/A | | | 0 | | X |
| Gustafsson, Szczepanski, Nelson, and Gustafsson (2012) | | | X | | | | | | | | 0 |
| Hatch and McCarthy (2005) | X | X | X | | | | | | 0 | | |
| Harper (2009) | X | X | | X | 0 | | N/A | | | | |
| Kaly and Heesacker (2003) | | 0 | 0 | | | | 0 | 0 | 0 | | |
| Larson (2007) | 0 | | X | | | | 0 | X | 0 | | |
| Martin and Leberman (2005) | X | | X | | | | 0 | | | 0 | |
| Mygind (2009) | X | 0 | | | | | | | | | |
| O'Connell (2010) | X | | X | 0 | | | | | | | |
| Orren and Werner (2007) | X | | X | | | | | | | | X |
| Passerelli, Hall, and Anderson (2010) | X | X | | X | 0 | 0 | | | | | |
| Sheard and Golby (2006) | | | 0 | 0 | | | | | 0 | | |
| Stott and Hall (2003) | | N/A | | X | | | 0 | | | | |
| Sugerman (2001) | | | | X | | | | | | | |

*(continued)*

**Table 1. (continued)**

| | Criterion | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sample | | | Questionnaire | | Survey design | | | Presentation of results | | |
| Publication | Inadequate sample size | No use of control group | Poor composition | Inappropriate for the study | Poor or no reporting of V&R | No pre-test | Poor timing of tests | No longitudinal component | No presentation of data | Poor descriptive statistics | Other variable problems |
| Wang, Ang, Teo-Koch, and Kahlid (2004) | | | **0** | | | | | | | | |
| Whittington and Mack (2010) | | **0** | **0** | | | | **0** | **X** | | | |

*Note.* **Sample**
*Inadequate size*: <100 participants in total is the criterion for too small, based on power 0.8, effect size 0.5, significance criterion 0.05. Where subsamples of <100 are used they are also included here, as well as samples suffering from attrition in longitudinal studies.
*No control group.* Equally relevant to study design. N/A is used where a control group was impossible or inappropriate for the study design.
*Poor composition.* A wide range of limitations: The use of volunteer or convenience samples, non-random selection, poor representation of the general population, lack of equivalence between treatment and control groups, selection bias.

**Questionnaire**
*Inappropriate for the study.* A wide range of limitations: Inappropriate for the age group or poor psychometric properties, modified for the study and not testing strongly, overlong or out-of-date, use of scale that encourages ceiling or floor effects.
V&R. Validity and reliability.

**Survey design**
*No pre-test.* N/A is used where a pre-test was impossible or inappropriate for the study design.
*Poor timing of tests.* Most common limitation here is conducting tests too close in time to the intervention.

**Presentation of results**
*No presentation of data*. This refers to the lack of information on tests for normal data distribution, ceiling or floor effects, and justification for using certain statistical methods.
*Poor descriptive statistics*. In some cases descriptive statistics are not adequately reported to support the inferential statistics or their interpretation.

**Other variable problems**
A range of issues in addition to those created by not having a control group or testing too close to an intervention: For example, lack of knowledge of the activities of the control group, the emergence of non-response bias in questionnaire returns, the emergence of a variable related to sample demographics, use of parents as proxies to score tests.

**Table 2.** Meta-Analyses in the Field of OAE: Criteria Used to Ensure the Methodological Rigor of Studies Included in Each Meta-Analysis.

| | Criterion | | | | | |
|---|---|---|---|---|---|---|
| | Sample | | | Questionnaire | Project and survey design | |
| Publication | Adequate sample size and/or representation of the population | Use of control group | Random assignment of participants to groups | Quality of questionnaire | Pre-test as well as post-test(s) | Clear description of project design |
| Cason and Gillis (1994) | | X | X | | X | |
| Hattie, Marsh, Neill, and Richards (1997) | X | X | | X | | X |
| Zelezny (1999) | | X | X | X | X | |
| Wilson and Lipsey (2000) | | X | X | | | |
| Hans (2000) | | | | | X | |
| Gillis and Speelman (2008) | | X | | | X | |

*Note.* OAE = outdoor adventure education.

as a limitation is around 80 to 100. However, there are ways and means by which small samples can be the basis of worthwhile research.

First, because sample size is related to the three parameters of statistical power, significance criterion, and effect size (see Cohen [1988] or Social Research Methods [2013] for discussions of this relationship), some compromise in the values of these other parameters will permit useful work. For example, relaxing the significance criterion, $p$, from .01 to .05 or accepting that only medium to large intervention effects will be detected with statistical significance are possible compromises. Naturally, claims made on the basis of the results would have to be tempered accordingly. Whether or not useful research can be done at subsample level (e.g., for gender or age group differences), and in a longitudinal study, where attrition of sample size is common, is a further consideration if sample size is small. Second, authors often comment on how their small sample size limits the generalizability of their results. Certainly, small size might limit the extent to which the sample is representative of a much larger and/or heterogeneous population, especially if the sample is also a convenience or self-selected sample. Nevertheless, good research can be done if claims of generalization are made in proportion to the representative nature of the sample.

The lack of a control group was seen as a limitation of the study in about one third of the entries in Table 1. Recruiting a large-enough sample from which to select or allocate equivalent (in size and composition) treatment and control groups is often a significant challenge. The use of an equivalent control group provides the opportunity to control for potential variables other than the intervention itself. However, even without a control group it is possible to conduct good research. A quasi-experimental research design that uses a pre-test and post-test in effect uses the treatment group as the control by testing it before it undertakes the intervention. This raises questions over when the pre-test should be conducted, but it is nevertheless an option if only a small sample size is available. Alternatively, a non-equivalent group of participants might be available to act as a control, and, again, claims made on the basis of the results would have to be tempered accordingly.

## *Self-Report Questionnaires: Suitability, Scales, and Administration*

Sixteen of the 22 studies in Table 1 used what might be called established questionnaires that had been designed, tested, published, and used elsewhere. The other 6 studies used tailor-made questionnaires that presumably were tested and judged to be valid and reliable within the context of the study, although this was not always reported. In total, 21 different questionnaires or sets of questionnaires were used. A more comprehensive survey by Sibthorp (2000) revealed a picture of a few commonly used instruments but many tailor-made ones. This finding leads to the following question: Should researchers be encouraged to use only "established" questionnaires to increase comparability between outcomes? Given the diversity of research objectives in OAE, it may not be possible for only a few established questionnaires to accommodate this diversity while permitting comparability. There is a danger of an instrument being used that is inappropriate for the research objectives of a project and/or the participants. If time

and resources permit, a better approach would be the rigorous design and testing of questionnaires that are specifically suited to each study. Careful reporting of the design and testing is essential to convince the research community the work is rigorous. In addition to numerous textbooks covering rigorous self-report questionnaire design and testing (e.g., Loewenthal (2001)), Sibthorp (2000) gives a brief, readily accessible set of suggestions for adventure education purposes.

Despite the availability of guidance on questionnaire design, an aspect that is rarely addressed in the OAE literature is the measurement scale (Carifio & Perla, 2007). Virtually, all self-report questionnaires assume that PSD can be measured using a Likert-type scale. Although the linearity of scores on such scales is debatable, issues more commonly revolve around the range of possible responses on individual items, whether there should be a mid-point, and the wording that is attached to each response. For example, research shows that with children, scales with few responses, perhaps four or five, of clearly different meaning should be used (Borgers, Hox, & Sikkel, 2004). By far the most common response range in our survey, across all age groups, was indeed four or five points. Neill's (2007) Life Effectiveness Questionnaire (LEQ) with an 8-point range was used in one study on a sample of adolescents and young adults, for which the questionnaire was specifically designed. Although short response ranges are popular, they are more susceptible to floor and ceiling effects, particularly if there is a participant response bias toward one end of the scale (McBee, 2010). In the pre-test–post-test approach to testing, this limits how much a participant can score down or up in the post-test, thus reducing any measure of effect. There is, therefore, an argument for using a longer scale if participants are able to understand the more subtle differences between responses: four or five points are suitable for children, but for adolescents, eight might be more appropriate.

An important decision to make during survey design concerns the timing of tests. Typically, a quantitative study will include a pre-test, a post-test, and, for longitudinal studies, a delayed-test some weeks or months afterwards. With respect to both pre-test and post-test, there is potential for the emotions experienced by participants in the days around joining or leaving an intervention to distort their scores (Ewert & Sibthorp, 2009; Hattie, Marsh, Neill, & Richards, 1997). It is advisable, therefore, to administer tests at least a week before and after the intervention, although there are many studies that cannot (or do not) do this and use the questionnaire a few days before and after or even on arrival at and departure from the residential center (e.g., Martin & Leberman, 2005). Of the 22 studies we reviewed, 15 conducted their pre-test and post-test within a week of the beginning and end of the intervention and, of these, 7 were on the first and last day of the intervention. On the other hand, there are examples in the literature of very good survey designs and test timings. In the next section of this article, we review a recent publication by Hayhurst, Hunter, Kafka, and Boyes (2015) in which they adopted a two pre-test, two post-test design, the first and last of these being one month removed from the intervention, to detect any impact on scoring of intervention-related emotions.

Occasionally, demonstrating rigor is simply a matter of making sure the process of survey data inspection is fully reported. Half of the articles we reviewed failed in this

area in some way or other. A common pitfall is poor reporting on data distributions, and whether they are normal or skewed, before going on to use an appropriate suite of statistical tests. It is very rare to see reports on ceiling or floor effects in data distributions (only 1 of the 22 papers mentioned this possibility), and even descriptive statistics are sometimes missing. It is easy to include this basic information, which creates a useful stepping stone from data acquisition to informed statistical analysis.

## *Managing Variables: Research Design*

Residential OAE programs are unusually rich experiential environments in which PSD can be readily elicited. One reason for this richness is OAE's multi-dimensional nature and the combined effect of these dimensions (McKenzie, 2003; Williams, 2013). While there may be a certain amount of "magic" in the ways in which the participants, the natural environment, the course activities, the instructors, and the wider social circumstances combine, these multiple dimensions and the many variables they create also present researchers with considerable obstacles, both within studies and when comparing the results of one study with those of another (Nichols, 2000). Commonly, the effect of many variables operating is to increase the variance on the chosen dependent variable (usually the effect of the intervention), which in turn reduces measured effect size and reduces the chance of a statistically significant result. It is important to manage the many variables so that their impact is reduced or, ideally, eliminated.

A particular interaction problem is the role of confounding variables. Ewert and Sibthorp (2009) identify a number of potential confounding variables in the wider field of experiential education. Only in 1 of the 22 papers we surveyed was the issue of the impact of multiple variables explicitly addressed. Clark, Marmol, Cooley, and Gathercoal (2004) note that "because we sought . . . such a holistic perspective, the sheer number of . . . variables examined could have potentially confounded our results" (p. 230). We suspect that this problem is much more widespread than is reported. Because in any one research study it is impractical to measure all potential variables, the normal practice would be to use a research design that reduces the likelihood of uncontrolled variables being present. The use of a control group, ideally an equivalent control group, provides a means by which some unwanted variables associated with the sample are controlled. A second means of controlling for unwanted variables at the research design stage is to use a testing schedule that will identify a confounding variable, such as intervention-related emotions in participants, as previously discussed.

It is clear that managing other variables that have the potential to influence the effect of the chosen independent variable is best done at the research design stage. However, if a researcher is lucky enough to have a large sample that can be subsampled and still permit statistically robust analysis, it is possible to isolate the effect of unexpected variables at the data analysis stage. In a recent study with 10- to 12-year-old school pupils attending a residential adventure course, data inspection revealed the possibility of socio-economic background affecting pupils' responses to the questionnaire (Scrutton, 2014). The sample was large enough—470 in total—to subsample and

compare scores between groups of pupils from different socio-economic backgrounds and thus understand the effect of this variable.

## Examples of Methodological Good Practice

Among the 22 articles that we reviewed, there is evidence of methodological good practice, although it is often sitting beside methodological limitations in the same article. Examples of good practice include the following: the random assignment of participants to the treatment and control groups from larger cohorts, thus providing good equivalence of gender, age, and race (Larson, 2007); the research, design, and testing of a tailor-made questionnaire for the specific needs of the project (Whittington & Mack, 2010); the succinct reporting of data analysis and results (Bobilya, Akey, & Mitchell, 2011); and the use of a time series of four tests to compensate for the lack of a control group to identify intervention-related variables such as pre-intervention anxiety (Hatch & McCarthy, 2005). The article by Gustafsson, Szczepanski, Nelson, and Gustafsson (2012) is unusual in that it explicitly declares its strengths as well as its weaknesses. In this case, the authors identify the large sample size of 230 in total and use of a questionnaire particularly suited to use with the school children participants.

Worthy of more detailed consideration is Hayhurst et al.'s (2015) measurement of resilience in youth during a sail-training voyage around New Zealand. This article is unusual in that it demonstrates how a research design can be modified in the course of a study to address perceived limitations. Their study took place in two parts, with the first informing the design of the second. First among the authors' perceived limitations in Study 1 were concerns about the sample: The experimental group comprised school pupils with a mean age of 16.6 years who were selected for the sail-training voyage by their schools, whereas the control group was comprised of university students on a convenient psychology course and had an average age 19.4 years. These two groups were neither equivalent in character nor of the same mean age. Second, the pre-test measure was taken on the first day of the voyage when program participants might assess their resilience as rather low in the face of the challenge ahead. A third weakness was the lack of any follow-up measurements in the wake of the possibility that the participants might have experienced post-group euphoria when they were surveyed on the last day of the voyage. A fourth perceived weakness related to understanding the face value of the results rather than the quality of the results, in that the authors did not examine any of the psychosocial processes that may have contributed to the increased resilience.

In their Study 2, Hayhurst et al. (2015) were able to address these weaknesses. The control group comprised high school pupils of mean age 16.4 years, closely related to the 16.5 years for the experimental group. The design of the study was changed to include assessment of the participants on four different occasions: one month prior to the voyage, the first day of the voyage, the last day, and five months after the voyage. The degree of stability of the first two scores and the second two scores would indicate whether intervention-related emotions were influencing any change in resilience as a result of the intervention. A number of psychosocial measures were also made on the first and last days of the voyage. The creative and rigorous methods used in Study 2

permitted the measurement of a small but significant increase in resilience as a result of the voyage, and good stability in the scores before and after the voyage, i.e., little evidence of intervention-related emotional effects.

## Summary, Conclusions, and Implications

As discussed at the beginning of this article, PSD is a difficult concept to define, and presents a challenge for any researcher to measure. The rigor of quantitative research in the OAE sector aimed at measuring PSD has been the subject of much criticism over the last 20 years. The three areas in which we see scope for addressing the issue of rigor are drawing on adequate samples, using appropriate questionnaires, and carefully considering the many variables that may influence outcomes. What follows are recommendations to help guide researchers as they aim to boost the robustness of their investigations.

In authors' discussions of the limitations of their articles, the size and composition of the sample is the most frequently acknowledged problem. Frequently, researchers have little option but to use a readily available sample. Nevertheless, as we, as a field, now have enough quantitative measures of the benefit of OAE for PSD to suspect an effect of about .3 or .4 (see Figure 1), it is suggested that researchers take this effect size together with a statistical power of about .8 and a significance criterion of .05 to determine sample size from tables (Cohen, 1988). In most cases, a sample size of between 80 and 100 in total will be enough for undertaking worthwhile research.

There are two other noteworthy points on sampling. First, bear in mind that analyzing data at the subset level, such as males versus females, reduces the effective sample size. And second, remember that a control group will require a sample of equivalent size to the experimental group. In longitudinal studies there is almost always an attrition of sample size. Initial sample size should be large enough to cope with all these possible variations. Second, lack of time and resources might prevent the acquisition of equivalent experimental and control groups, but good research can be done using quasi-experimental methods with an experimental group only or using a nonequivalent control group with an understanding of its likely influence on the study.

It is tempting to take results from a study with a limited sample size and selective composition and make inferences about the behavior of a much larger and diverse population. However, our investigations confirm that it is far better to be more realistic about what the sample represents by way of a specific population. In some cases this may be a smaller population that is more accurately represented by the sample. Although the conclusions of such a study would be less widely applicable, they would be more rigorous and might constitute valuable source material for a future meta-analysis.

The diversity of findings and large number of variables, in what is a relatively small body of OAE research (Figure 1), mitigates against the use of a few tried and tested questionnaires. It is suggested that, if necessary, researchers develop their own questionnaire in a rigorous manner and then report on this in their article. There is a good deal of advice in textbooks, such as Loewenthal (2001), and in publications, such as Sibthorp (2000), on developing and testing new questionnaires. We urge those who are

developing questionnaires to pay particular attention to the appropriateness of the measurement scale for their participants and the likelihood of ceiling and floor effects, as these aspects are less well covered in the literature.

With regard to the timing of the use of questionnaires during the research, to avoid participants' pre- and post-test emotions influencing their scores, it is advisable to administer questionnaires a few weeks before and after the intervention. To avoid over-familiarity with the questionnaire, the question order might be changed between administrations.

It is ironic that the multiple dimensions of OAE and their integrated effect, which provide such a rich experiential learning environment with many variables, should also create problems for rigor in OAE research: We cannot claim the one without managing the other. Unfortunately, unlike the questionnaire and the sample, where something approaching a formulaic solution is available, making decisions on the plethora of variables requires cautious and meticulous consideration at the project design stage.

The most comprehensive and most commonly cited meta-analysis in OAE is that of Hattie et al. (1997). In common with many other authors, they concluded that "too little is known . . . about why adventure programs work" (p. 43), and urged researchers to continue investigating causes and effects, with a view to them being brought together in future meta-analyses. This kind of research, they suggested, was needed to more deeply understand the interdependency of the many variables inherent in OAE. It is the measurement of the magnitude of effects that has concerned us here, and Hattie et al.'s call to continue investigating these, in association with causes, remains as applicable today as it was then. If we fail to measure effects with good-quality research, then perceived weaknesses in our empirical studies will limit its usefulness in being credibly used to inform practice and influence policy makers, while arguably leaving OAE's corpus of quantitative research vulnerable to criticism from related disciplines.

## Appendix

Papers reviewed from *Journal of Experiential Education* and *Journal of Adventure Education and Outdoor Learning*.

Bell, B. (2006). Wilderness orientation: Exploring the relationship between college preorientation programs and social support. *Journal of Experiential Education*, *29*, 145-167.

Bobilya, A., Akey, L., & Mitchell, D. (2011). Outcomes of a spiritually-focused outdoor education program. *Journal of Experiential Education*, *33*, 301-322.

Clark, J., Marmol, L., Cooley, R., & Gathercoal, K. (2004). The effects of wilderness therapy on the clinical concerns of troubled adolescents. *Journal of Experiential Education*, *27*, 213-232.

Cross, R. (2002). The effects of an adventure education program on perceptions of alienation and personal control among at-risk adolescents. *Journal of Experiential Education*, *25*, 247-253.

Ewert, A., & Yoshino, A. (2011). The influence of short-term adventure-based experiences on levels of resilience. *Journal of Adventure Education and Outdoor Learning*, *11*, 35-50.

Freeman, P. (2002). The role of outdoor recreation in family enrichment. *Journal of Adventure Education and Outdoor Learning*, *2*, 131-145.

Gassner, M., & Russell, K. (2008). Relative impact of course components at outward bound Singapore: A retrospective study of long-term outcomes. *Journal of Adventure Education and Outdoor Learning*, *8*, 133-156.

Gustafsson, P., Szczepanski, A., Nelson, N., & Gustafsson, P. (2012). Effects of an outdoor education intervention on the mental health of schoolchildren. *Journal of Adventure Education and Outdoor Learning*, *12*, 63-80.

Harper, N. (2009). The relationship of therapeutic alliance to outcome in wilderness treatment. *Journal of Adventure Education and Outdoor Learning*, *9*, 45-59.

Hatch, K., & McCarthy, C. (2005). Exploration of challenge courses' long-term effects on members of college student organisations. *Journal of Experiential Education*, *27*(3), 245-265.

Kaly, P., & Heesacker, M. (2003). Effects of a ship-based adventure program on adolescent self-esteem and ego-identity development. *Journal of Experiential Education*, *26*, 97-104.

Larson, B. (2007). Adventure camp programs, self-concept, and their effects on behavioral problem adolescents. *Journal of Experiential Education*, *29*, 313-330.

Martin, A., & Leberman, S. (2005). Personal learning or prescribed educational outcomes: A case study of the outward bound experience. *Journal of Experiential Education*, *28*, 44-59.

Mygind, E. (2009). A comparison of childrens' statements about social relations and teaching in the classroom and outdoor environment. *Journal of Adventure Education and Outdoor Learning*, *9*, 151-169.

O'Connell, T. (2010). The effects of age, gender and level of experience on motivation to sea kayak. *Journal of Adventure Education and Outdoor Learning*, *10*, 51-66.

Orren, P., & Werner, P. (2007). Effects of brief wilderness programs on relations to adolescents' race. *Journal of Experiential Education*, *30*, 117-133.

Passerelli, A., Hall, E., & Anderson, M. (2010). A strengths-based approach to outdoor adventure education: Possibilities for personal growth. *Journal of Experiential Education*, *33*, 120-135.

Sheard, M., & Golby, J. (2006). The efficacy of an outdoor education curriculum on selected aspects of positive psychological development. *Journal of Experiential Education*, *29*, 187-209.

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*, 159-169.

Sugerman, D. (2001). Motivations of older adults to participate in outdoor adventure experiences. *Journal of Adventure Education and Outdoor Learning*, *1*(2), 21-33.

Wang, C. K. J., Ang, R., Teo-Koch, S., & Kahlid, A. (2004). Motivational predictors of young adolescent' participation in an outdoor adventure course: A self-determination theory approach. *Journal of Adventure Education and Outdoor Learning*, *4*, 57-65.

Whittington, A., & Mack, E. (2010). Inspiring courage in girls: An evaluation of practices and outcomes. *Journal of Experiential Education*, *33*, 166-180.

## Authors' Note

The title of this article is borrowed from the book *Assessing Personal and Social Development: Measuring the Unmeasurable?* by Inman, Buck, and Burke (1998)

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Note

1. In North America, and elsewhere, *intra- and interpersonal growth* can be considered broad equivalents of personal (intra—within oneself) and social (inter—with others) development (see, for example, Priest & Gass, 2005).

## References

Allison, P., & Von Wald, K. (2010). Choices, values and untidy processes: Personal, social and health education on educational expeditions. In S. K. Beames (Ed.), *Understanding educational expeditions* (pp. 55-65.). Rotterdam, Netherlands: Sense.

Barrett, J., & Greenaway, R. (1995). *Why adventure? The role and value of outdoor adventure in young people's personal and social development*. Coventry, UK: Foundation for Outdoor Adventure.

Bobilya, A., Akey, L., & Mitchell, D. (2011). Outcomes of a spiritually-focused outdoor education program. *Journal of Experiential Education*, *33*, 301-322.

Borgers, N., Hox, J., & Sikkel, D. (2004). Response effects in surveys on children and adolescents: The effects of number of response options, negative wording and neutral mid-point. *Quality & Quantity*, *38*, 17-33.

Brookes, A. (2003). A critique of neo-Hahnian outdoor education theory. Part one: Challenges to the concept of character building. *Journal of Adventure Education and Outdoor Learning*, *3*, 49-62.

Buck, M., & Inman, S. (1998). Personal and social development at the crossroads. In S. Inman, M. Buck, & H. Burke (Eds.), *Assessing personal and social development: Measuring the unmeasurable?* (pp. 1-16). London, England: Falmer.

Carifio, J., & Perla, R. (2007). Ten common misundertsandings, persistent myths and urban lgends about likert scales and likert response formats and their antidotes. *Journal of Social Sciences*, *3*, 106-116.

Cason, D., & Gillis, H. (1994). A meta-analysis of outdoor adventure programming with adolescents. *Journal of Experiential Education*, *17*(1), 40-47.

Clark, J., Marmol, L., Cooley, R., & Gathercoal, K. (2004). The effects of wilderness therapy on the clinical concerns of troubled adolescents. *Journal of Experiential Education*, *27*, 213-232.

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

English Outdoor Council. (2005). *High quality outdoor education: A guide to recognising and achieving high quality outdoor education in schools, youth services, clubs and centres*. Retrieved from http://www.englishoutdoorcouncil.org/HQOE.pdf

Ewert, A., & Sibthorp, J. (2009). Creating outcome through experiential education: The challenge of confounding variables. *Journal of Experiential Education*, *31*, 376-389.

Gillis, L., & Speelman, E. (2008). Are challenge (Ropes) courses an effective tool? A meta-analysis. *Journal of Experiential Education*, *31*, 111-135.

Gustafsson, P., Szczepanski, A., Nelson, N., & Gustafsson, P. (2012). Effects of an outdoor education intervention on the mental health of schoolchildren. *Journal of Adventure Education and Outdoor Learning*, *12*, 63-80.

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

Hattie, J., Marsh, H., Neill, J., & Richards, G. (1997). Adventure education and outward bound: Out-of-class experiences that have a lasting effect. *Journal of Educational Research*, *67*(1), 43-87.

Hayhurst, J., Hunter, J., Kafka, S., & Boyes, M. (2015). Enhancing resilience in youth through a ten day developmental voyage. *Journal of Adventure Education and Outdoor Learning*, *15*(1), 40-52.

Hopkins, D., & Putnam, R. (1993). *Personal growth through adventure*. London, England: David Fulton.

Inman, S., Buck, M. & Burke, H. (1998). Assessing Personal and Social Development: Measuring the unmeasurable? London, England: Falmer Press.

Lipsey, M., & Wilson, D. (1993). The efficacy of psychological, educational and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*, 1181-1209.

Loewenthal, K. (2001). *An introduction to psychological tests and scales* (2nd ed.), Hove, UK: Psychology Press.

Martin, A., & Leberman, S. (2005). Personal learning or prescribed educational outcomes: A case study of the outward bound experience. *Journal of Experiential Education*, *28*, 44-59.

McBee, M. (2010). Modeling outcomes with floor or ceiling effects: An introduction to the Tobit Model. *Gifted Child Quarterly*, *54*, 314-320.

McKenzie, M. (2003). Beyond "The outward bound process": Rethinking student learning. *Journal of Experiential Education*, *26*, 8-23.

Neill, J. (2007). *Life effectiveness questionnaire: A research tool for measuring personal change*. Retrieved from http://wilderdom.com/leq.html

Neill, J. (2008). *Enhancing life effectiveness: The impacts of outdoor education programs* (Unpublished PhD thesis). Department of Philosophy, University of Western Sydney, Sydney, Australia.

Nichols, G. (2000). A research agenda for adventure education. *Australian Journal of Outdoor Education*, *4*(2), 22-31.

Priest, S., & Gass, M. (2005). *Effective leadership in adventure programming* (2nd ed). Champaign, IL: Human Kinetics.

Pring, R. (1982). Personal and social development: Some principles for curriculum planning. *Cambridge Journal of Education*, *12*, 3-14.

Rickinson, M., Dillon, J., Teamey, K., Morris, M., Choi, M., Sanders, D., & Benefield, P. (2004). *A review of research on outdoor learning*. Shrewsbury, UK: National Foundation for Educational Research and King's College London.

Scrutton, R. (2014). Outdoor adventure education for children in Scotland: Quantifying the benefits. *Journal of Adventure Education and Outdoor Learning*. Advance online publication. doi: 10.1080/14729679.2013.867813

Sibthorp, J. (2000). Measuring weather . . . and adventure education: Exploring the instruments of adventure education research. *Journal of Experiential Education*, *23*, 99-107.

Social Research Methods. (2013). *Statistical power*. Retrieved from www.socialresearchmethods.net/kb/power.php

Williams, R. (2013). Woven into the fabric of experience: Residential adventure education and complexity. *Journal of Adventure Education and Outdoor Learning*, *13*(2), 107-124.

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

Young Adult Learners Partnership. (2003). *Explaining personal and social development*. Retrieved from http://www.education.gov.uk/publications/standard/publicationDetail/Page1/RR480

Zelezny, L. (1999). Educational interventions that improve environmental behaviors: A meta-analysis. *The Journal of Environmental Education*, *31*(1), 5-14.

## Author Biographies

**Roger Scrutton** is Honorary Research Fellow in Outdoor Education at the University of Edinburgh. Previously he spent 34 years in the School of Geosciences at Edinburgh conducting teaching and research in geophysics. He is a Fellow of the Royal Society of Edinburgh, where he helps to deliver the Young People's Programme.

**Simon Beames** is Lecturer in Outdoor Education at the University of Edinburgh. He is the creator of the Outdoor Journeys Programme - a cross-curricular outdoor learning initiative for school children - and a frequent speaker and workshop leader for teachers on how they may "take learning outdoors".