The final publication is available at Springer via http://dx.doi.org/10.1007/s10964-016-0463-9

The Impact of the Project K Youth Development Program on Self-Efficacy: A Randomized Controlled Trial

Kelsey L. Deane[a]*, Niki Harré[b], Julie Moore[c] and Matthew G. R. Courtney[d]

University of Auckland[a,b,d] and Graeme Dingle Foundation[c]

[a] **Corresponding Author:** Kelsey L. Deane, PhD, Faculty of Education & Social Work, The University of Auckland. Private Bag 92601, Auckland, 1150, New Zealand. Phone: +64 9 373 7513. Fax: +64 9 623 8898. Email: k.deane@auckland.ac.nz

[b] School of Psychology, The University of Auckland. Private Bag 92019, Auckland, 1142, New Zealand.

[c] Graeme Dingle Foundation, PO Box 305 474, Triton Plaza, North Shore, 0757, New Zealand.

[d] Faculty of Education & Social Work, The University of Auckland. Private Bag 92601, Auckland, 1150, New Zealand.

The Impact of the Project K Youth Development Program on Self-Efficacy: A Randomized Controlled Trial

Abstract

A key issue for youth development programs is whether the learning they provide is transferred to participants' daily lives. It is also important that they are effective for the diverse range of participants they attract. This study used a randomized controlled trial design to measure the impact of Project K, a New Zealand-based youth development program, on academic and social self-efficacy. Project K combines a three-week wilderness adventure, a 10 day community service component, and one year of mentoring to promote positive growth in 14-15 year olds with low self-efficacy. At baseline, the evaluation included 600 Project K (46% female) and 577 Control participants (48% female) and revealed that Project K was effective in improving both social and academic self-efficacy from pre- to post-program with effects being sustained one year later. Parents' perceptions of changes in the participants' interpersonal skills supported these findings. Differential program effects were found across participant subgroups, particularly one year after program completion. The implications of these differences are discussed.

Keywords:

Positive Youth Development, Self-Efficacy, Program Evaluation, Randomized Controlled Trial, Experimental Design

Introduction

As psychological beings, what people believe they can do matters greatly in terms of the actions they undertake. Such self-beliefs are especially important for young people as they face new challenges and are expected to make major choices based on their perceived capabilities. This article discusses the evaluation of a New Zealand youth development program (Project K) focusing on its effectiveness in raising the participants' self-efficacy, a key self-belief as will be detailed below. At a broad level, the article addresses the critical question of if, and how, youth development programs can support young people to develop positive beliefs about their efficacy. In particular, it considers whether youth development programs, in which young people are removed to a greater or lesser extent from their everyday lives, can have an impact on their sense of efficacy *within* their everyday lives. This study also investigated the extent to which any positive self-efficacy outcomes held across the diverse range of program participants. This addresses another key issue for youth development programs: how to ensure all participants benefit, and how to avoid replicating patterns of social disadvantage.

We start with an overview of the literature on self-efficacy and its relevance to youth development programs. We then discuss social and academic self-efficacy in particular, which are the key focus of Project K, and central to young people's lives. Next, we review literature on the outcomes of youth development programs for diverse participants, before describing the program and how it attempts to improve participants' self-efficacy. Our hypotheses are then outlined.

Self-efficacy and its Relevance to Youth Development Programs

There is considerable evidence that initiative and motivation are, to a large degree, contingent on a person's belief in his or her capabilities, over and above actual skill and ability (Bandura, 1977; 1997; Wigfield, Eccles, Schiefele, Roeser, & Davis-Kean, 2006). This belief in one's ability to carry out the specific tasks needed to obtain a goal is often referred to as "self-efficacy" and was originally proposed by Albert Bandura (1977). As he conceived it, self-efficacy promotes persistence, commitment and the willingness to take on challenging goals within the relevant domain. There is increasing evidence that self-efficacy is related to adaptive health (Hall, Stewart, Arger, Athenour, & Effinger, 2014; Litt & Kadden, 2015) and academic behaviors (Høigaard, Kovač, Øverby, & Haugen,

2014; Jungert, Hesser & Traff, 2014) as well as positive undertaking of social roles (Dicke et al., 2014; Glatz & Buchanan, 2015) and physical activity (Bauman et al., 2012).

Bandura (1997) proposed that self-efficacy can be enhanced by successful mastery performances (see also Rodgers, Murray, Courneya, Bell, & Harber, 2009; Voskuil & Robbins, 2015; Warner et al., 2014), or observing others, especially similar others, execute a relevant task (Bandura, 1997; see also Voskuil & Robbins, 2015). This emphasis on direct experiences or observations also means that self-efficacy was initially assumed to be relatively specific to particular skills, contexts or life domains. For example, a person may learn how to play the guitar through practice and imitating skilled players, which leads to self-efficacy for playing the guitar. It does not necessarily follow that this person feels more confident and is more persistent when it comes to sports, academic endeavors or human relationships.

However, there have also been moves toward measuring "generalized" self-efficacy; that is a sense of confidence and willingness to set and persist with goals that applies across domains (e.g., Craig, McInroy, Austin, Smith, & Engle, 2012; Lennings, 1994; Tipton & Worthington, 1984). There is evidence for both domain specificity (Rodgers et al., 2009), and generalizability across domains (Hazenberg, Seddon & Denny, 2015; Jungert et al., 2014; Widmer, Duerden, & Taniguchi, 2014) or across skills within a single domain (Hayhurst, Hunter, Kafka, & Boyes, 2015). Furthermore, in a recent review of young people's self-efficacy for physical activity, Voskuil and Robbins (2015) emphasized how the self-appraisal, which is at the core of self-efficacy, can be focused on the behavior, the domain, or the situation. It follows from this that, in developing and demonstrating self-efficacy, the individual is asking, "What type of task is this?" and "Do I have the skills required?"

These issues are at the heart of youth development programs and the frameworks on which these are based. Insofar as they are promoting self-efficacy (which we will address next), youth development programs are hoping that participants will make a psychological bridge between tasks they have mastered within the program and everyday tasks. Nevertheless, youth development programs are artificially constructed ecological contexts designed to purposefully promote development. This may create challenges if there is a lack of attention to transferring the learning to other naturally occurring contexts (Deane & Harré, 2014a).

Reviews have shown that efficacy-related beliefs are indeed frequently targeted by youth development programs (Catalano, Berglund, Ryan, Lonczak & Ryan, 2004; Roth & Brooks-Gunn, 2003). Such beliefs also feature prominently in the frameworks undergirding many of these programs. For example, one of the four internal "building blocks" for healthy youth development proposed by the Search Institute®'s Developmental Assets® framework is social competencies (Benson, 2007). Competence and Confidence are also two of the "5 C's" of the Positive Youth Development framework thought to reflect youth thriving (Jeličić, Bobek, Phelps, Lerner, & Lerner, 2007).

So do youth development programs succeed in their efforts to boost efficacy-related beliefs beyond the specific program context? Meta-analyses of adventure-based programs (Hattie, Marsh, Neill & Richards, 1997; Wilson & Lipsey, 2000), youth mentoring (DuBois, Portillo, Rhodes, Silverthorn & Valentine, 2011), and after-school programs that target young people's social and emotional learning (Durlak, Weissberg, & Pachan, 2010) have all measured and found such positive effects on self-efficacy or confidence. However, each of these studies also demonstrates that effects are not consistent across programs. In fact, a meta-analysis of three rigorously evaluated youth empowerment programs provided no evidence to support an impact on self-efficacy (Morton & Montgomery, 2013).

Going back to our earlier argument, this lack of impact may be because there is insufficient attention, in some programs, to encouraging young people to see their program-related mastery experiences as evidence of their competence, or at least potential competence, in an everyday domain. Notably, one recent study attempted, with some success, to raise young people's academic self-efficacy in an adventure program setting by encouraging them to make connections between their adventure-related achievements and school work (Widmer et al., 2014).

Zimmerman and Cleary's (2006) theory of self-regulation for learning offers further insight into how programs may encourage young people to see the mastery they have acquired in the program context as demonstrating that they have the skills to successfully master tasks within everyday domains. Self-regulation is a learning cycle that aids decisions about which goals to pursue and which strategies to employ prior to and during goal pursuit. Initially, individuals reflect on the value of the task at hand, and how well they believe they can perform (their self-efficacy) before deciding to

pursue a goal. Good self-regulators use strategies (e.g., time management, proximal goal setting, planning, etc.) to monitor their progress against goals and to determine if further effort is needed and if they want to persevere.

A review we conducted of youth adventure programs showed clear evidence of a self-regulation learning cycle in action (Deane & Harré, 2014a). Importantly, participants in these programs confront novel and challenging experiences and then, with encouragement, support and observation of successful others (program facilitators and peer participants), learn how to successfully navigate these experiences. This is followed with reflective debriefs that help participants to evaluate their performance and make a bridge between these experiences and situations that they more commonly face. In theory at least, an experiential learning cycle as described above should aid transfer of self-efficacy to other desirable goals. This is particularly likely if participants are exposed to multiple iterations of the cycle, as is the case with Project K, the program at the centre of this article (this will be discussed in detail later).

The Importance of Social and Academic Self-Efficacy for Adolescent Development

Self-efficacy is relevant to most of the settings that young people encounter. Academic and social domains are, however, particularly relevant, as these are key contexts in which young people are developing and making choices. Furthermore they are shared contexts, common to almost all young people and therefore appropriate targets for youth development programs aimed at diverse participants.

Social self-efficacy reflects an individual's belief in his or her capability to communicate with others, build relationships, manage interpersonal conflict, and assert personal viewpoints. It has consistently been linked to better psychological and behavioral adjustment, for example lower levels of depression (Bandura, Barbaranelli, Caprara, & Pastorelli, 1996; Caprara, Gerbino, Paciello, DiGunta, & Pastorelli, 2010; Hermann & Betz, 2006), loneliness (Hermann & Betz, 2006), withdrawal (Connolly, 1989; Wichmann, Coplan, & Daniels, 2004), other internalizing symptoms (Connolly, 1989; Kim & Chichetti, 2003) and delinquent behavior (Allen, Leadbeater & Aber, 1990; Caprara et al., 2010). It can also engender a more collaborative and productive interpersonal style

(Desivilya & Eizen, 2005). Given the ubiquity of social relationships to almost every aspect of human life, this domain should be considered crucial for youth development programs.

Similarly, research in the academic domain has highlighted the importance of skills such as the ability to focus on schoolwork, avoid distractions, and complete homework on time for succeeding at school (a concept also called self-regulation for learning; Caprara et al., 2008; Pajares, 2002). Higher academic self-efficacy has also been associated with scholastic achievement (Bandura et al., 1996; Bandura, Pastorelli, Barbaranelli & Caprara, 1999; Carroll et al., 2009), greater educational aspirations (Bandura et al, 1996; Carroll et al., 2009) and lower school dropout (Caprara et al., 2008). Academic self-efficacy should thus also be considered an essential outcome of youth development programs, especially given the role of school success in the transition to adulthood.

Youth Development Programs for Diverse Participants

As well as considering how to boost self-efficacy and encourage participants to generalize their mastery experiences to key everyday domains, youth development programs should, ideally, serve all participants equally. Indeed the philosophy of Positive Youth Development, which underpins many programs, assumes that all young people require access to asset-rich settings in order to thrive, regardless of risk status or other demographic characteristics (Damon, 2004). Individual and cultural responsiveness are also acknowledged to be features of high quality youth development programs (Durlak et al., 2007; Roth & Brooks-Gunn, 2003). Thus, one should expect such programs to be effective across key diversity dimensions.

Surprisingly few evaluations of youth development programs have explored if effectiveness differs across demographic groups (Deane & Harré, 2014a; DuBois et al., 2011; Durlak et al., 2010). Those that have done so have found mixed results. Better outcomes have been reported for females involved in some adventure programs (McKenzie, 2003; Caines & Centre, 2010); however, Glass and Benshoff (2002) found no gender or ethnic differences in their evaluation of a challenge course for young people. Similarly, Scales, Benson, Leffert and Blyth (2000) demonstrated that spending time in a youth program was predictive of thriving across all racial-ethnic groups involved in their study of 6000 adolescents in the U.S. In contrast, Orren and Werner (2007) found that the self-esteem of African American participants involved in a brief wilderness program targeting diverse youth

deteriorated from pre- to post-assessment. Further, DuBois et al.'s (2011) meta-analysis of mentoring programs revealed stronger effects for young people with either high levels of individual or environmental risk but not for those who had a combination of both. Because program effects may be varied for different groups, investigating the impact of programs on a diverse range of participants should be of concern to those interested in promoting growth for everyone who takes part.

Current Study

This article outlines the findings from an evaluation of Project K, a comprehensive, nationwide New Zealand youth development program that targets young people with low self-efficacy but who do not exhibit high risk behavior. In particular, Project K aims to positively influence the social and academic self-efficacy of a wide range of demographically diverse participants. Project K's program theory is based on an experiential learning cycle that encourages participants to transfer the efficacy they have developed within the program to their day to day lives (Deane & Harré, 2014b). Furthermore, it involves three stages, as will be detailed below, giving participants multiple opportunities for efficacy growth and transfer. As such, Project K provided an opportunity to investigate our expectations that programs with such a focus should increase participants' self-efficacy outside of the program context, and that the effects would not differ across diverse participants.

The evaluation used a randomized controlled trial design. Despite the strength of this design, few evaluations of youth development programs have employed experimental or quasi-experimental designs or investigated long-term effects (Deane & Harré, 2014a; DuBois et al., 2011; Durlak et al., 2010; Brooks-Gunn & Roth, 2014). This is certainly the case in New Zealand, where, until now, no published examples of randomized controlled trial evaluations of youth development programs existed. Given the resources and time needed to implement high-quality experimental research in an applied community-based setting and the inherent challenges acknowledged to exist in such endeavours (Deane & Harré, in press), this is not surprising. Nevertheless, there is a global trend toward "evidence-based" practice (Nunns, Peace, & Witten, 2015) and randomized controlled trial

evaluations can provide useful evidence as to the overall value of a program, as well as its specific strengths and the participants who may gain the most benefit.

Project K is owned by the New Zealand-based Foundation for Youth Development (known as FYD) and operates in 11 regions throughout the country. It has three components, an outdoor education experience called the "Wilderness Adventure", a "Community Challenge" and individual mentoring. For the three week "Wilderness Adventure" participants are transported to an outdoor camp. There they prepare for a seven to 10 day expedition by developing a number of requisite skills (e.g., goal-setting, problem-solving, communication, and leadership). The emphasis is on participation in team-building and challenge-based activities. The "Community Challenge" follows. This involves participants working together in their home community to explore available facilities and resources and network with key community figures. Workshop sessions address issues pertinent to adolescent health and well-being (e.g., relationships, sex, drugs and alcohol, nutrition, and community opportunities) and participants develop a collaborative project designed to address a community need. For the third component, participants are paired with an adult mentor for a year. Mentors are trained to develop non-judgmental relationships with the young people, and help them set four goals to achieve for the year. One of these must be an academic goal. The end of the relationship (and Project K) is marked by a graduation celebration.

As touched on earlier, Project K's program theory assumes that by carefully crafting opportunities for challenging mastery experiences and coupling these with facilitated self-reflection during which transference of learning is emphasized, self-efficacy (particularly in the interpersonal domain due to the strong social nature of the activities) will be developed and this will flow over into other important domains of their lives (Deane & Harré, 2014b). In addition, program facilitators are trained to have high expectations of participants as these expectations have been shown to have an important influence on efficacy beliefs in a variety of contexts (Bandura, Barbaranelli, Caprara, & Pastorelli., 2001; Schunk & Meece, 2006). Crucially, program facilitators are encouraged to have high expectations of all participants, and there is evidence that this is indeed the case (Hollis, Deane, Moore, & Harré, 2011). Finally, Project K's three stages allows participants numerous opportunities

to test themselves and to obtain ongoing support which reinforces their self-beliefs in different contexts (Deane & Harré, 2014b).

In line with the literature reviewed and the description of how Project K operates, we expected participants to show meaningful self-efficacy gains applicable to their daily lives. More specifically, the interpersonal skills needed for the team-based activities and the social nature of the first two program stages led us to expect an impact on social self-efficacy. The need for participants to set at least one academic goal and have this progress monitored and supported by a mentor who reinforces the learning transfer from the previous program stages also led us to expect an increase in academic self-efficacy. Nevertheless, because Project K operates within a social, but not an academic domain and so has more opportunity to directly influence the former beliefs, we expected the effect on this latter domain to be weaker. The efficacy outcomes reported in this article arose from Project K data collected at baseline, the immediate end of the program and one year later from randomly assigned program and control participants. Consequently, we hypothesized that Project K participants would show an increase in social and academic self-efficacy from baseline to one year post-program and the rate of increase would be greater than that of the control group (Hypothesis 1). Further, we hypothesized that there would be a stronger effect on social self-efficacy compared to academic self-efficacy (Hypothesis 2). Recognizing the potential limitations of self-report (Crockett, Schulenberg & Petersen, 1987) and the value of triangulating such data, we sought additional ratings of the participants' efficacy. As parents and teachers are the adults most likely to observe changes in their children's or students' competency, their perceptions of their child's or student's interpersonal skills (a proxy for social self-efficacy) were sought at baseline and at the end of the program. Unfortunately, there were difficulties gaining sufficient reliable data from teachers. Consequently, we only include a hypothesis pertaining to parents. We hypothesized that the parents of Project K participants would rate their children's interpersonal skills more highly upon program completion than the parents of control participants (Hypothesis 3). Finally, Project K targets young people with low self-efficacy but works with schools along the full socio-economic spectrum as well as co-educational and single-sex schools, and schools that have high proportions of ethnic minorities. Given the emphasis of the program theory on serving a wide range of diverse participants (Deane & Harré, 2014b) and evidence

suggesting that high but equal expectations are communicated by facilitators to all participants (Hollis et al., 2011), we hypothesized that the program effects on the above outcomes would be consistent across participant subgroups (Hypothesis 4).

Methods

Program Selection and Characteristics

Eighty-four Project K programs (a program being one delivery to a group of 12 participants who are from the same school) from 42 schools across all 11 regions in which Project K operates were delivered during the timeframe for the randomized controlled trial. Of the 84 programs, 50 were included in the randomized controlled trial evaluation. Regions delivering a small number of programs were required to include all of their programs in the evaluation. New programs (i.e., sourced from a school for the first time) were also required to take part. Regions with large numbers of programs were usually not required to evaluate subsequent programs once they had one or two evaluated programs underway. For various reasons (e.g., not following selection guidelines), a few programs were dropped after being selected as an evaluated program.

The 50 evaluated programs were sourced from 38 schools. New Zealand schools are assigned a *decile rating* which reflects the socio-economic status (SES) of their students. Decile 1 represents the 10% of schools with the greatest proportion of students coming from low SES areas and decile 10 the 10% of schools with the lowest proportion of such students. The 50 programs drew participants from schools ranging across the full continuum (1 to 10), the mean decile rating was 5.98 (SD = 2.88).

Participant Selection and Procedure

Project K's program participant selection process occurred in three stages. Moore's (2005) Self-Efficacy Questionnaire (SEQ) was first administered by a trained Project K staff member to all students in Year 10 (13-15 years old) during class time (this was the baseline measure or Time 0 for those who were eventually selected into the Project K or Control group). The questionnaire contained three self-efficacy subscales: academic, social and help-seeking (these are explained in more detail below). Each subscale was summed. Additional perspectives of the students' efficacy in these three domains were also sought. Two teachers' ratings for each student were obtained, given the likelihood

that students behave differently in different classes. Since teachers were required to comment on each Year 10 student in the school, to reduce the survey administration time, a very brief rating scale containing three statements, with each thought to reflect one of the subscales, was used. The items were as follows "Is self-motivated and consistently works close to ability levels" (academic); "Interacts positively with peers" (social) and, "Is able to communicate positively with teachers" (help-seeking). Responses were indicated on a balanced six-point Likert scale ranging from "Not well at all" to "Very well" for each question. The median was calculated for each of the nine scores (three scores for each of the summed self-reported subscales and three for each teacher). Individuals with scores below the median were coded as "1" and those scoring equal or above the median were coded as "0". These scores were then summed and students were ranked.

In the second stage, the list of the highest ranking individuals, which were those with the lowest measured self-efficacy relative to their peers, was reviewed by a program liaison team (consisting of the program director, the school guidance counsellor and other school personnel familiar with the students) to identify any students whose behavioral or mental health history would likely place themselves or others at too great a risk by participating in the program. These individuals were excluded from the selection pool and provided with information about other support services if there was any indication that their needs were not currently being met.

In the case of the evaluated programs included in this study, at least the top ranking 20 female and 20 male students from the remaining list and their families were then invited to attend an information evening (fewer were invited in the case of non-evaluated programs as there were only 12 places in the program). Twenty-four young people who assented and provided parental consent to participate in the program and the evaluation were then randomly and evenly allocated to the Project K or "program" group (n = 12), or the control group (n = 12). Any remaining people on the list who did not have their name drawn were retained as reserves for the program in case someone else had withdrawn. Within two months of completing the baseline SEQ (Time 0), Project K students began the Wilderness Adventure. To reduce demoralization for the control participants, an active control group process was developed. Control participants were invited to an adventure day at the beginning of the program and at the end of the year. The program staff also met the control participants once

each school term, lunch was provided and they were connected to other support services if a need was identified.

Both groups completed the self-efficacy evaluation measures of interest here within the same time frame. Control participants completed the end of program (Time 1) measure at a meeting with the program staff. Project K participants also completed these measures in the presence of staff members within two weeks of the graduation ceremony. Data collection for the one year post-program follow up occurred between twelve and fourteen months after program completion (Time 2). Because none of the participants were involved in the program at this time, most surveys were sent and returned via post; however, a few were administered at their homes by Program Managers or FYD staff (unfortunately, exact numbers of those administered at home were not recorded). The data were obtained by the first two authors from FYD after receiving approval from their institution's Human Participants' Ethics Research Committee.

Participant Sample Characteristics

The participants were in Year 10 of secondary school (aged 13 to 15 years) at program start. The number of individuals who participated in any part of the full randomized control trial study (which included a host of measures other than self-efficacy that are not the focus here) was 1177 (Project K program participants = 600 and Controls = 577). Theoretically, 1200 participants should have been included in the randomized controlled trial across the 50 programs (12 per group, per program) but some control participants dropped out of the evaluation at the last minute. There were 278 and 277 females in the Project K and Control groups, respectively. Two Project K participants and two Control participants did not specify their gender. Ethnic identity was coded using the Total Response method whereby a participant was counted in each group he or she selected. As a result the total ethnicity responses were greater than the number of participants in the sample. The ethnic breakdown of the baseline sample was as follows for the Project K Condition: 433 (75.0%) European (ancestry or 1st generation European); 139 (24.1%) Māori (indigenous New Zealanders); 84 (14.6%) Pacific (i.e., Tongan, Samoan, Fijian, Tokelauan, Niuean, or Cook Island Māori or other Pacific Island); 25 (4.3%) Asian (Chinese, Indian, Southeast Asian, or Other Asian); 14 (2.4%) Other; and, for the Control Condition: 392 (69.1%) European; 143 (25.2%) Māori; 96 (16.9%) Pacific; 38 (6.7%) Asian; and 20

(3.5%) Other. Note that approximately 16% of the Project K and 16.5% of the Control sample identified with more than one of the above ethnic categories. Twenty-three participants in the Project K group and ten in the Control group did not specify their ethnicities.

Of the 600 Project K participants who were initially included in the evaluation, 76 (12.67%) did not "graduate" from the program. To graduate, Project K participants were required to complete at least 75% of the program and accomplish three goals. Reasons for program non-completion varied but included students withdrawing themselves from the program due to other commitments, needing more time for school work or because they moved from the region. Participants could also have been withdrawn from Project K by program staff had they not committed to the program. For the purposes of this study, these participants were considered the program attriters but are included in the analyses of program effects in line with guidelines that Intention to Treat analyses should be used to ascertain effects for any participants who were initially intending to receive the treatment or intervention (Have et al., 2008; Parker, Bush & Harris, 2014).

Baseline (Time 0) self-efficacy data could only be located for 1149 of these participants (Project K = 573 and Control = 576). It was unclear if the efficacy data were lost or if the participants joined the program later as reserves. Evaluation attrition also occurred. At the end of the program (Time 1), 899 participants completed the self-efficacy survey: 482 Project K participants and 417 Controls, reflecting a 19.67% attrition rate for the former and 27.73% attrition rate for the latter. At one year post-program (Time 2), 601 participants responded: 317 Project K participants and 284 Controls (47.25% attrition from baseline for the former and 50.78% attrition for the latter). Evaluation attrition was influenced to a degree by program attrition as contact details were often lost when program attriters moved regions. At Time 2, program staff had a much more difficult time maintaining contact with students who were not involved in the program and/or had left school because many contact details had changed. In other cases, participants did not return surveys or declined to participate. Unfortunately, no systematic records of the reasons for evaluation attrition were kept. Measures to deal with missing data and evaluation attrition are discussed in the Analyses section below.

Measures

Academic and Social Self-efficacy.

As previously discussed three self-efficacy scales were initially developed for this study, and all three were used for the selection process. However a verification of the factor structure of the "help-seeking" scale with the current sample at Time 0 failed to provide support for its discriminant validity (neither exploratory nor confirmatory procedures provided an interpretable solution). The help-seeking scale was therefore not included in the analyses and will not be discussed further here. The academic and social subscales of Moore's (2005) SEQ align closely with the academic and social self-efficacy subscales of Muris' (2001) SEQ-C and Bandura's children's self-efficacy scale (Bandura et al., 1999) with some items adapted to suit the New Zealand youth context. All items in the SEQ were pre-piloted with two classes of Year 10 students from one New Zealand secondary school. The students completed the SEQ anonymously and were then placed into small groups. To elicit information on their interpretation of the items, each group was allocated one of the self-efficacy domains and asked to paraphrase the questions in their own words. This was followed by a discussion to gather general feedback. From the feedback it was clear that the students did not interpret two of the questions as intended, and these were reworded.

Academic items primarily assessed the degree to which individuals felt that they could successfully self-regulate their learning (e.g., "How well can you study when there are other interesting things to do?") but an item about meeting teachers' expectations (i.e., "How well can you succeed in satisfying your teachers with your schoolwork?") was also included. The social items pertained to relationship building, communication and assertiveness in classroom or group situations (e.g., "How well can you have a chat with an unfamiliar person your age?"). Both subscales contained eight items measured on a balanced six point Likert scale (1 = Not well at all to 6 = Quite well).

Because the school-specific survey items were no longer applicable to those who had left school at the final time point (n = 111), the academic subscale of the SEQ was no longer relevant, thus it was removed from their follow up (Time 2) surveys. In addition, because four of the social self-efficacy items were specific to an academic environment, these were slightly modified at Time 2 in an attempt to maintain relevance to school leaver participants. For example, the word "classmates" was exchanged with "people your own age". School students (n = 490) were administered the academic

self-efficacy scales as well as the eight original and four modified social self-efficacy items. The school leavers, on the other hand, only received eight social self-efficacy items (four original and four modified).

Parents' perspectives of the participants' interpersonal skills.

The six-item prosocial and communication skills subscale of the 12-item Social Competence Scale—Parent Version (Corrigan, 2002) was used in this study (considered a measure of parents' perceptions of their children's interpersonal skills) as a means to triangulate the self-reported social self-efficacy ratings. An example item was "Your child can give suggestions and opinions without being bossy". Responses were coded on a balanced five-point Likert scale (1 = Not at all, to 5 = Very Well). The scale was originally developed as part of the Fast Track Project, a project designed to assess an academic, social skills and behavioral regulation intervention involving children and adolescents in the U.S. (Corrigan, 2002).

Parents of Project K and Control participants who could be matched to their children's pre to end of program data (n = 481) completed the full scale at baseline and at the end of the program; however, exploratory procedures involving the six-item prosocial behaviors and communication skills and the six-item emotional self-regulation did not result in an interpretable solution for the current sample. Because the prosocial and communication (interpersonal) skills subscale aligned more strongly with the participant social self-efficacy scale compared to the emotional self-regulation subscale it was thought to be a better measure for triangulation. This scale was therefore retained and the emotional self-regulation scale was dropped. Details regarding convergent validity and invariance across time points for the prosocial and communication subscale are described below.

Analyses

Measurement models and invariance tests.

To test the stability of the social and academic efficacy factors at each of the three time points, (a) model fit, (b) convergent and discriminant validity, and (c) invariance tests were undertaken. All such analyses were carried out with the assistance of MPlus 7.35 (Muthén & Muthén, 2015) using MLR estimation (suitable for the moderately skewed 6-point ordinal variables in the dataset; Muthén & Muthén, 2015). Model fit for the measurement models was assessed across four indices in accordance

with Fan and Sivo (2007): acceptable fit = RMSEA ≤ .08; CFI ≥ .90; SRMR ≤ .08; $\hat{g} \geq .90$; good fit = RMSEA ≤ .05; CFI ≥ .95; SRMR ≤ .05; $\hat{g}$ (gamma hat) ≥ .95. In accordance with Hair et al. (2010), convergent validity was met where item factor loadings met a minimum of |.5| (|.45| deemed acceptable due to rounding). In accordance with Voorhees, Brady, Calantone, & Ramirez (2016), discriminant validity was met where the requirements for the HT-$MT_{.85}$ (heterotrait-monotrait ratio minimum of .85) and AVE-SV (in which the average variance extracted in a factor exceeds shared variance between associated factors) tests were met.

To ensure that the social and academic efficacy factors constituted meaningful factors at each of the three time points, seven measurement models were tested (see online supplement, Figure S1). Each of these measurement models, all involving the social and academic efficacy factors, demonstrated acceptable to good model fit (see online supplement, Table S1). Across all models, item-factor loadings ranged between .46 and .92 suggestive of convergent validity. In addition, each of the seven models met the requirements of discriminant validity, i.e., HT-$MT_{.85}$ and AVE-SV tests (see online supplement, Table S2). This suggested that the meaning of each factor was sufficiently distinct at each time point.

Multiple longitudinal and multi-group invariance tests were also carried out to test the stability of the factor structures over time because the purpose of this study was to track shifts in social and academic self-efficacy over time (see online supplement, Figure S2). For longitudinal invariance tests, parameters were specified as parallel specific factors over time (McArdle, 2007). In accordance with McArdle (2007), where metric invariance was met, assessments of shifts in factors over time could be defended. In accordance with Cheung and Rensvold (2002), configural invariance is met when RMSEA ≤ .05, and metric invariance is met when |Δ CFI| (compared to configural CFI) ≤ .01. Results suggested that, when both social and academic efficacy factors were measured among students remaining in the evaluation from Time 0 to Time 1 (test a; $n = 899$), the requirements for metric invariance were met.

In addition, results suggested that when both social and academic efficacy factors were measured among the school student participants (who received the original academic and social self-efficacy items at Time 2) across all three time points (test b; $n = 490$), the requirements for metric

invariance were met. However, because the school leaver participants (n =111) did not receive the academic self-efficacy scale and responded to four modified social self-efficacy items that no longer pertained to the academic context, a multi-group invariance test assessing equivalence of the modified social self-efficacy factor for student participants (n = 490) versus school leavers (n = 111) was conducted (Test d). In this case the configural test failed. Similarly, test c, the longitudinal invariance test assessing equivalence of the single social efficacy factor over the three time points for the 601 participants who responded at Time 2 (including school leavers), failed to meet the requirements of configural invariance (see online supplement, Table 3; full output for invariance testing available from the corresponding author). This suggested that the meaning of social efficacy was different for adolescents outside of the secondary school context. This resulted in the need to remove the school leavers (n = 111) from the analyses at Time 2, reducing the sample at this time point to 490. The limitations with regards to generalizability are considered in the discussion.

Finally, when testing was carried out on the two-factor model distinguishing between parental perceptions of their child's interpersonal skills and self-reported social efficacy from Time 0 to Time 1 (test e, n = 481), the requirements for metric invariance were also met. Further the two factors were correlated at each time point (r = .21, p. < .001 at Time 0 and r = .24, p. < .001 at Time 2) indicating congruence between youth and parents' perspectives of the young person's social competency.

Cronbach's alphas for the academic self-efficacy scale were .87, .90 and .90 cross-sectionally for Time 0, 1 and 2, respectively. For social self-efficacy these were .85, .87, and .88, respectively. For parents' perceptions of prosocial and communication skills Cronbach's alpha was .86 at both Time 0 and Time 1.

Missing values and evaluation attrition.

There were no missing values for the academic and social self-efficacy items at Time 0 or Time 1 when each time point was analysed cross-sectionally and only one participant in the Time 2 dataset including the 490 school students who responded to both the academic and social self-efficacy questions did not respond to all items. However, four participants had no baseline self-efficacy data and 48 (9.8%) had no Time 1 data for the longitudinal dataset created to include Time 0, 1 and 2 data based on the 490 participants in the Time 2 sample. Therefore a missing value analysis was conducted

and because Little's MCAR statistic indicated the data could be considered missing at random (χ^2 [310] = 309.95, p = .49), missing data were imputed using the Expectation Maximization algorithm in SPSS 23.

A longitudinal dataset containing parents who responded to the Social Competence Scale—Parent Version at Time 0 and Time 1 was created. Fewer than 5% data were missing at an item-level for both time points and a non-significant Little's MCAR statistic obtained from the missing values analysis indicated that the missing values could be considered missing at random (χ^2 [389] = 391.563, p = .45). Missing values were again imputed using the Expectation Maximization algorithm in SPSS 23.

Data missing due to evaluation attrition (non-respondents at Time 1 and 2), however, could not be considered missing at random. It was therefore not defensible to impute data longitudinally and necessary to assess the effects of attrition on the (a) baseline equivalence of the 899 Project K and Control participants who responded at Time 1, and (b) the 490 school students who provided analyzable data at Time 2. To do this, propensity score matching analyses (Lanza, Moore & Butera, 2013) were conducted using the FUZZY extension add-on in SPSS 23 first for the Time-0-to-Time-1 sample (n = 899), and then for the Time-0-to-Time-1-to-Time-2 sample (n = 490). The first step involved binary logistic regressions to ascertain if the participants' demographic characteristics (contrast-coded variables for Gender, Māori vs. non-Māori, Pacific vs. Non-Pacific, and Asian vs. non-Asian ethnicity) and their baseline academic and social self-efficacy scores were significant predictors of membership in either the Project K or the Control group (Control = 0, and Project K = 1). None of the variables included in the model were significant predictors of group membership for the Time-0-to-Time-1 sample with 899 participants. Further, no significant predictors (demographic characteristics and baseline scores including parents' baseline ratings of participants' interpersonal skills) were obtained in the logistic regression predicting Project K or Control group membership when comparing 481 of the parents' ratings of their children's interpersonal skills for the youth participants in the Time-0-to-Time-1-sample. This suggests that the groups were not distinguishable at baseline and the baseline equivalence established by the randomization process still held for all end of program (Time 1) assessments. Consequently, matching based on propensity scores was not needed.

For the sample with data for all three time points (n = 490), Māori (vs. non-Māori) and Pacific (vs. non-Pacific) ethnicity were significant predictors of group membership (*Wald* = 4.86, b = .55, p <.05, *odds ratio* = 1.74 for Māori and *Wald* = 5.06, b = -.67, p <.05, *odds ratio* = .51 for Pacific) indicating that the Project K and Control groups were not balanced in terms of ethnic group proportions. Consequently, propensity scores were used to match Control respondents with Project K respondents without replacement using optimal matching with a match tolerance of .02 (using the FUZZY extension add-on in SPSS 23). This resulted in a sample of 420 matched participants (210 in each group).

Binary logistic regressions were also conducted to assess if respondents could be differentiated from non-respondents at Time 1 (respondents = 0, and non-respondents = 1). Māori and Pacific ethnicity were significant predictors of whether or not someone responded at Time 1 (*Wald* = 5.16, b = -.38, p <.05, *odds ratio* = .69 for Māori, and *Wald* = 4.30, b = -.42, p <.05, *odds ratio* = .66 for Pacific). Māori and Pacific participants were less likely to respond. Baseline academic self-efficacy was also a significant predictor (*Wald* = 6.49, b = -.24, p <.05, *odds ratio* = .79). Those with lower academic self-efficacy at baseline were less likely to respond.

At Time 2, because the school leavers had to be removed from the program effect analyses, they were included with the non-respondents in a binary logistic regression predicting membership in the respondent vs. non-respondent groups. Māori, Pacific and Asian ethnicity and baseline academic and social self-efficacy were all significant predictors. Māori participants were less likely to be included in our Time 2 analyses (*Wald* = 16.14, b = -.60, p <.001, *odds ratio* = .55) as were Pacific students (*Wald* = 8.83, b = -.54, p <.01, *odds ratio* = .58) and those with lower academic self-efficacy at baseline (*Wald* = 9.50, b = -.25, p <.01, *odds ratio* = .78). In contrast, Asian participants (*Wald* = 15.02, b = 1.16, $p.$ <.001, *odds ratio* = 3.19) and those with higher social self-efficacy at baseline (*Wald* = 5.16, b = .18, $p.$ <.05, *odds ratio* = 1.20) were more likely to be included in the Time 2 analyses. The limitations associated with evaluation attrition are considered in the discussion.

Multi-level analyses of program effects.

To account for the dependency inherent in the nested structure of the Project K randomized controlled trial data (i.e., observations nested in individuals and individuals nested within programs),

Hierarchical Linear Modelling was employed, using HLM7 software (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011). This also allowed us to appropriately model cross-level interactions (Hox, 2002). To address the first hypotheses regarding Project K's effectiveness in raising the academic and social self-efficacy of participants, we first report the linear growth trajectories of participants' scores from Time 0 to Time 2 using three-level models. The lowest level of analysis was the repeated measures (i.e., Time; baseline = 0, post-program = 1, and one year post-program = 2), individuals were at the second-level, and programs at the third. Program effects for the trajectories were estimated by including a grand mean-centred "Condition" variable (0.49 = Project K, -0.51 = Control) while statistically adjusting for the main effects of the following grand mean-centred demographic variables: gender (Male = -0.47, Female = 0.53) and Māori (Māori = 0.75, Non-Māori = -0.25), Pacific (Pacific = 0.84, Non-Pacific = -0.16), and Asian (Asian = 0.94, Non-Asian = -0.06) ethnicity at Level-2 and school decile rating (Decile) at Level-3.[1] Outcomes were analyzed separately (see Eqs. 1.0-2.3 in Appendix S2 of the online supplement for the equation specifications of the three-level models). HLM estimates for the growth models are calculated with Empirical Bayes estimation which account for the lower reliability of estimates with missing data (Hox, 2002), making the estimates for the person-level trajectories robust despite attrition at Time 1 and 2. Given the much larger sample of the original Intention to Treat sample (n = 1140 for the efficacy outcomes at baseline after list-wise deletion of those missing values for any demographic variables) this article focuses on findings from this sample (note that the demographic variables were included as covariates so that they were statistically adjusted for); however, we also conducted an analysis of the linear trajectory for the propensity matched sample of 420 participants across the three time points to compare program effects across the two samples. Because there is equivalence between the Project K and

[1] The "European" and "Other" variables were not included as covariates in any analyses that incorporated ethnicity to avoid model multicollinearity.

Control groups in this sample at baseline, no demographic covariates are included in the models (See Eqs. 2.4-3.0 in Appendix S2 of the online supplement).

Two-level models with individuals at Level-1 and programs at Level-2 were also generated to determine effects on academic (ASE1) and social self-efficacy (SSE1) at program completion and one year post-program independently. Differences between Project K and Control participants at Time 1 were investigated, adjusting for the baseline measure of the relevant outcome (i.e., either ASE0 or SSE0, grand mean-centred). The main effects for gender, ethnicity, and school decile rating were also adjusted for in each of the two-level models (coded as above). Further, because Project K and Control groups were not equivalent at baseline for the 490 included in the Time 2 analyses, these analyses were repeated with the propensity matched sample (n =420) and the effects were compared. The demographic variables are included in the comparative analyses with the propensity matched sample because interactions between the demographic variables and program condition are included (explained below). An assessment of the effects sizes for program effects on academic compared to social self-efficacy enabled us to address the second hypothesis (also explained below).

A similar two-level model with individuals at Level-1 and programs at Level-2 (exchanging only the outcome and baseline variables) was conducted to address the third hypothesis regarding whether parents' perceptions of the participants' prosocial and communications skills mirrored the participants' self-reported social self-efficacy outcomes over the duration of the program (See Eqs. 3.1-4.3 in Appendix S2 of the online supplement).

Finally, to address the fourth hypothesis that the program would not be differentially effective across demographic subgroups—for males vs. females, across different ethnicities (i.e., Māori, Pacific and Asian) and school decile contexts—interactions between the baseline levels of each of the aforementioned variables with Condition (i.e., Project K or Control) were assessed in the two-level models (at Level-1 interaction terms were calculated prior to model specification by multiplying the two grand mean-centred variables (e.g., Condition x Gender).

An additional exploratory test of differential program effects was conducted to ascertain if outcomes differed depending on the starting position of the participants on all outcomes of interest. Level-1 interaction terms were therefore also calculated by multiplying the grand mean-centred

Condition variable with each grand mean-centred baseline measure (e.g., Condition x ASE0). Details of the model specifications for the interaction models are provided in the online supplement (See Eqs. 3.0-4.2 in Appendix S2 of the online supplement). Restricted maximum likelihood estimation (REML) was used for all analyses and the estimates calculated with robust standard errors are reported as some of the variables deviated slightly from a normal distribution.

Significant interaction effects were further examined using simple slope analyses calculated with utilities developed by Sibley (2009) in Microsoft Excel. The test of the effect for each simple slope may be thought of as roughly equivalent to an independent groups *t*-test. Interaction effects between a binomial predictor and a binomial moderator (e.g., those between condition and gender) can be interpreted much like those found in an Analysis of Variance (ANOVA) and the associated simple slopes for these types of interactions represent differences between group means. To reduce the risk of Type I error associated with the post-hoc simple slope tests, the criterion for statistical significance ($p < .05$) was adjusted with a Bonferroni correction by dividing .05 by 2 (the number of slopes tested) resulting in an $\alpha = .03$.

Effect sizes.

The program effects sizes are reported as unstandardized regression weights because these represent regression-adjusted differences between the Project K and Control group on the original metric used (i.e., the 6 point Likert scale) but because efficacy is an abstract construct it is still difficult to grasp the practical significance of these results. When using multiple regression techniques, the convention is to report the R^2 for the model. This provides an estimation of the percentage of variance in the outcome that can be explained by the variables in the model. This can be further broken down into part correlations (η^2) which signify the amount of variance explained by one variable independently of all others (Field, 2009). This is problematic when using multilevel modelling because there is variance to be explained for multiple variables and at multiple levels. At times, adding a predictor may even generate negative values for the additional variance explained (Snidjers & Bosker, 1999), hence why we chose to focus on differences on the original metric.

Nevertheless it is useful to understand the program effect size with respect to other similar types of programs, thus standardized mean differences (Cohen's *d*) were calculated between groups at

each time point for all outcomes of interest. In addition, Cohen's *d* effect sizes were also calculated independently for each group for Time 0 to Time 1 means for academic and social self-efficacy and parent perceptions of the participants' interpersonal skills and for Time 0 to Time 2 means for academic and social self-efficacy for the original and the propensity matched samples. The absolute differences between the Project K and Control group effect sizes over time enable a rough comparison with effect sizes reported for similar types of youth programs. They also enable a comparison of the effect sizes for academic compared to social self-efficacy to address our second hypothesis.

Results

Descriptive Statistics, Bivariate Correlations, and Intra-Class Correlations (ICC)

Table 1 presents the means and standard deviations for each of the person-level variables (and effect sizes between groups at each time point) and Table 2 presents the bivariate correlations for those specified in the hierarchical linear models. The ICCs for intercept-only HLMs signify the degree of dependence in the observations at each level of analysis. For the three-level models, The $ICC_{Level\ 2}$ was 0.35 and 0.30 for academic and social self-efficacy (.43 and .39 for the propensity matched sample), respectively. The $ICC_{Level\ 3}$ was 0.04 for both outcomes for the original sample (.02 and .04 for the propensity matched sample). All values are based on intercepts that varied significantly ($p <$ .05) across individuals (r_0) and across programs (u_0), with the exception of the $ICC_{Level\ 3}$ for academic self-efficacy in the propensity matched sample. For the two-level models, the ICC for ASE1 was .02 and for ASE 2, the ICC = .03 for the original and .02 for the propensity matched sample. For SSE1, the ICC = .05 and for SSE2, the ICC was .06 (.05 for the propensity matched sample). The intercepts for social self-efficacy across all samples varied significantly ($p <$.05), but those for academic self-efficacy did not. This suggests that even when the ICC was significant at the program-level, only up to 6% of the variance could be attributed to program-level factors.

Hypothesis 1: Did Program Participants Increase in Self-Efficacy Relative to Control Participants?

The non-significant main effect of Condition on the intercepts in the academic and social self-efficacy growth models indicates that there was no difference between Project K and Control in the baseline levels of these measures. However, significant cross-level interactions between Condition and Time

for both academic (*b* [unstandardized beta] = 0.22, $p < .001$) and social self-efficacy ($b = 0.20$, $p < .001$) in the linear growth models indicate that the linear trajectory for these scores (from baseline to one year post-program) differ depending on program condition (see Table 3). The same pattern of effects was obtained when these analyses were conducted on the propensity matched sample ($b = 0.15$, $p < .01$ for academic and $b = 0.16$, $p < .01$), although the regression coefficients were slightly smaller.

The simple slopes analyses (for the full Intention to Treat sample) for academic self-efficacy revealed that the linear increase from Time 0 to Time 2 was significant for the Project K group (simple slope = 0.23, $t = 5.816$, $p < .001$) but not the Control group (simple slope = 0.008, $t = 0.24$, $p = .81$). The regression-adjusted mean for baseline scores for Project K was 3.32 and it was 3.28 for the Control group and 3.78 compared to 3.30 (respectively) at one-year post program. The simple slopes analyses for social self-efficacy demonstrated that the linear trajectory for both Project K and Control was significant over the three time points (simple slope = 0.17, $t = 7.13$, $p < .001$ for Control; simple slope = 0.37, $t = 11.44$, $p < .001$ for Project K) but the effect was stronger for the Project K group (Time 0 regression-adjusted means: $M_{PK} = 4.12$ and $M_{Control} = 4.11$; and Time 2 regression adjusted means: $M_{PK} = 4.76$; $M_{Control} = 4.34$).

The results for the growth models provide an overview of the overall effects of Project K on the two outcomes of interest. However, the two-level models allowed for a more detailed investigation of the effects at Time 1 and Time 2. Table 4 presents the immediate post-program effects (Time 1) for academic and social self-efficacy and parents' perceptions of the participants' interpersonal skills for the two-level HLM models. At the end of the program, Project K participants on average had significantly higher academic ($b = 0.44$, $p < .001$) and social self-efficacy scores ($b = 0.36$, $p < .001$) than Control participants, adjusting for the baseline level of each outcome and the main effects of gender, the ethnicity variables, and school decile rating and interactions between Condition and each of the covariates.

Table 6 presents the one year post-program effects (Time 2) for academic and social self-efficacy for the original sample with 490 participants and the comparative propensity matched sample with 420 participants. The findings demonstrate that Project K participants still had higher academic

(b = 0.25, p < .001 for the original sample and b = 0.27, p < .001 for the propensity matched sample) and social self-efficacy (b = 0.25, p < .001 for the original sample and b = 0.29, p < .001 for the propensity matched sample) one year after program completion.

Hypothesis 2: Was the Program Effect Stronger for Social Self-Efficacy?

The effect sizes (Cohen's *d*) representing differences between Control (baseline) and Project K (reference) at Time 0 were -.04 for academic and -.06 for social self-efficacy indicating virtually no difference between the groups at baseline as expected. At Time 1, in support of the Project K intervention, the between group differences in means represented an effect size of .43 for academic and .44 for social. Similarly, for Time 2, the between group effect sizes (for the propensity matched sample) were .35 for academic and .44 for social (see Table 1).

A comparison of standardized mean shifts across Time 0 and Time 1 data points provide more evidence in support of the program: the Time 0 to Time 1 shift for the Project K group was .50 for academic and .64 for social. For the Control group the effect sizes were .00 and .03 respectively. For Time 0 to Time 2, the effect size for the Project K group (propensity matched sample) was .39 for academic and .87 for social, and, for the Control, the effect sizes were .06 and .33 respectively (see Table 5).

Because these latter effect size calculations represent changes over time and account for baseline scores (which were not different between Project K and Control on any measure), the absolute differences in the Project K and Control effect sizes for the latter calculations provide more accurate estimates of the magnitude of Project K's effects. According to traditional effect size guidelines (see DuBois et al., 2011), the findings indicate that for academic self-efficacy the program effect from baseline to the end of the program is considered moderate (0.50 effect size difference between the groups) whereas the effect size for social self-efficacy is considered moderate to large (.61 effect size difference). The effect size representing differences from baseline to one year post-program is considered small to moderate for academic self-efficacy (.33 effect size difference between Project K and Control) and moderate for social self-efficacy (.54 effect size difference). In accordance with our second hypothesis, this suggests that the effect on social self-efficacy was stronger than that for academic self-efficacy.

Hypothesis 3: Did Parents of Project K participants rate their child's interpersonal skills more highly after the program than Parents of Control participants?

A significant Condition effect ($b = 0.24$, $p < .001$) for the two-level model assessing parents' perceptions of participants' interpersonal skills at Time 1 indicated that the parents of Project K participants did report higher interpersonal skills for their children compared to the parents of Control participants, adjusting for the baseline level of each outcome and the main effects of gender, the ethnicity variables, and school decile rating and interactions between Condition and each of the covariates (see Table 4). The effect size representing the difference between the Control group (baseline) and Project K (reference) parents' perceptions was -.04 at Time 0 and .33 at the end of the program (Time 1). For Project K parents' perceptions, the effect size from Time 0 to Time 1 was .50, and for the Control it was .12, indicating a small to moderate program effect.

Hypothesis 4: Were Program Effects Consistent across Participant Subgroups?

The two-level HLM models also facilitated analysis and interpretation of interaction effects based on Condition by participant characteristics thus providing evidence of whether our fourth hypothesis was supported. Table 4 also illustrates the main effects for baseline self-efficacy and parental perception scores, gender, ethnicity and school decile and the interaction effects for the above mentioned variables with Condition at the end of the program (Time 1) and Table 6 presents these for one year post-program (Time 2). Since our interest was in determining whether or not the program is differentially effective across participant subgroups we focus on the interaction effects and refer readers who are interested in the overall demographic effects to Tables 4 and 6.

A significant interaction effect for social self-efficacy at Time 1 indicated that over the duration of the program, Project K was differentially effective in boosting self-efficacy for young people who identified as Pacific (also known as Pasifika compared to non-Pacific youth ($b = -0.45$, $p < .05$). Analysis of the simple slopes for this effect revealed that the slope for non-Pacific participants (representing the difference between Project K and Control students) was significant ($b = .51$, $t = 5.79$, $p < .001$; $M_{PK} = 4.57$ and $M_{Control} = 4.05$) as was the slope for Pacific participants ($b = .20$, $t = 2.49$, $p < .05$; $M_{PK} = 4.43$ and $M_{Control} = 4.43$); however the latter effect was weaker. See Panel A in Figure 1. No other interaction effects were obtained for Time 1 outcomes.

Interestingly, at Time 2, a significant Condition x Pacific interaction effect was also obtained but for academic and not social self-efficacy (b = -0.45, p < .01). This effect was replicated with the propensity matched sample (b = -0.75, p < .01). The simple slope analysis revealed a significant slope for non-Pacific (b =.50, t = 4.25, p < .001; M_{PK} = 3.81 and $M_{Control}$ = 3.31) but not for Pacific (b = .05, t = .44, p =.67; M_{PK} = 3.49 and $M_{Control}$ = 3.44). See Panel B in Figure 1.[2]

A consistent Condition x School Decile interaction effect was obtained between the original and propensity matched samples for Time 2 academic self-efficacy (b = -0.05, p < .05 for the original sample and b = -0.07, p < .05 for the propensity matched sample). The simple slope for participants from schools with a decile rating one standard deviation lower than the mean was significant b =.47, t = 3.67, p < .001; M_{PK} = 3.69 and $M_{Control}$ = 3.22). The slope for participants from schools with a decile rating one standard deviation higher than the mean was not significant (b =.10, t = 0.66, p = .51; M_{PK} = 3.60 and $M_{Control}$ = 3.54. See Panel A in Figure 2.

Similarly, a significant Condition x Māori ethnicity interaction effect was obtained for social self-efficacy at Time 2 (b = -0.36, p < .05). This effect was again replicated with the propensity matched sample (b = -0.31, p = .05). The simple slope analysis for this effect revealed a significant slope for both non-Māori (b =.40, t = 4.74, p < .001; M_{PK} = 4.59 and $M_{Control}$ = 4.91) and Māori (b = .18, t = 1.96, p =.05; M_{PK} = 4.48 and $M_{Control}$ = 4.30) but the effect was stronger for the former. See Panel B in Figure 2.

A significant Condition x Gender interaction was also obtained for Time 2 academic self-efficacy with the original sample (b = -0.34, p < .05) but this was not replicated with the propensity matched sample thus the simple slope effects were not explored. Conversely, a significant Condition x Asian ethnicity interaction was obtained for Time 2 social self-efficacy but only for the propensity matched sample (b = -0.34, p < .05). The simple slope for the Asian participants was significant (b =

[2] The estimates for all of the simple slope analyses at Time 2 were based on the propensity matched sample because the sample size was similar to the original sample and because it provides a stronger counterfactual analysis of program effects.

.49, $t = 6.32$, $p < .001$; $M_{PK} = 4.61$ and $M_{Control} = 4.12$) but the effect for non-Asian participants was not ($b = .09$, $t = 0.91$, $p = .37$; $M_{PK} = 4.46$ and $M_{Control} = 4.37$). See Panel C in Figure 2.

The final interaction effect pertains to our exploration of differential program effects according to the participants' starting point. A significant Condition x baseline Academic Self-Efficacy interaction was obtained for Time 2 academic self-efficacy with the original sample ($b = -0.24$, $p < .01$) and the propensity matched sample ($b = -0.27$, $p < .01$). The slope for participants starting with lower academic self-efficacy at baseline (one standard deviation lower than the mean) was significant ($b = .51$, $t = 4.36$, $p < .001$; $M_{PK} = 3.44$ and $M_{Control} = 2.93$) and the slope for those starting with higher academic self-efficacy (one standard deviation higher than the mean) was not ($b = .03$, $t = .30$, $p < .001$; $M_{PK} = 3.85$ and $M_{Control} = 3.82$). See Panel D in Figure 2.

There were no consistent differential effects for any particular participant subgroup. The program was effective in impacting either academic or social self-efficacy for all participant subgroups during the program and one year post-program. Nevertheless, depending on the outcome and participant group of interest, program effects did differ - particularly over the follow up period. Thus our fourth hypothesis was not confirmed. Interestingly, where stronger effects did occur they tended to be for those who reported lower levels of the outcome at baseline. To clarify, the main effects for school decile, gender, ethnicity found in Table 3 represent baseline efficacy levels because the intercept for Time is 0 (the baseline measurement). These estimates demonstrate that several subgroups for which stronger effects were obtained reported significantly lower efficacy levels at baseline compared to their reference group (e.g. Asian participants reported lower social self-efficacy than non-Asian students and participants from low SES schools reported lower academic self-efficacy than those from high SES schools). This finding is considered in the discussion below.

Discussion

There is considerable evidence that people's beliefs about their capabilities (i.e., self-efficacy beliefs) are crucial to their willingness to take on challenges and persist in the face of setbacks (Bandura, 1977; 1997; Wigfield et al., 2006). Such beliefs have been associated with adaptive behaviors in a wide variety of domains (e.g., Bauman et al., 2012; Hall et al., 2014; Dicke et al., 2014; Glatz &

Buchanan, 2015; Høigaard et al., 2014; Jungert et al., 2014; Litt & Kadden, 2015). Youth development programs commonly target self-efficacy related beliefs (Catalano et al., 2004; Roth & Brooks-Gunn, 2003), given their particular importance to young people who are in the process of making major life choices. The question remains, however, as to if, and how, targeting self-efficacy within the context of a youth development program can positively influence young people's beliefs within the everyday domains in which they live. An accompanying question is whether any positive influence applies to the diverse range of participants that are targeted by youth development programs.

The current study explored these questions through a randomized controlled trial evaluation of Project K, a 14-month long New Zealand youth development program with adventure, community and mentoring components. It focused on academic and social self-efficacy and parents' perceptions of interpersonal skills. We found significant increases in Project K participants' academic and social self-efficacy relative to Control participants both at the end of the program and one year later. Notably, there was evidence that the size of the program effect on social self-efficacy was slightly larger than that for academic self-efficacy, as predicted. There was also congruence between parent and youth perceptions of social efficacy as parents of Project K participants also rated their children's social competence more highly after the program than did Control parents. While most of the findings held across participants with different demographic characteristics over the duration of the program, notable variation was found over the follow up period as will be discussed further.

The estimated effect sizes of Project K on self-efficacy outcomes over the duration of the program were moderate to large and from baseline through to follow up, the effects were small to moderate. These effects compare favourably with those found in meta-analyses of other youth development programs (DuBois et al., 2011; Durlak & Weissberg, 2007; Hattie et al., 1997). For example, DuBois et al.'s (2011) youth mentoring meta-analysis yielded a pre-test to follow-up effect of .17 over seven studies. Hattie et al.'s (1997) meta-analysis of adventure-based programs (for adults and youth) found a .40 pre-test to follow up effect size for the youth sample which is comparable to the cumulative pre-test to follow up Project K effects on academic (.33) and social self-efficacy (.54).

So why does Project K have this impact? We suggest this is likely due to the program's emphasis on an experiential learning cycle (Deane & Harré, 2014a) that facilitates the self-regulation process thought to be at the core of efficacy-building (Zimmerman & Clearly, 2006). Project K participants engage in direct and vicarious experiences of mastery (both important sources of self-efficacy; Bandura, 1997) and reflect on these with program facilitators and peers during post-experience debriefs. Particular attention is paid to generalizing the learning to other more normative contexts and new goals are set (Deane & Harré, 2014b). Importantly, the three components and the 14 month time-frame of Project K allow for multiple iterations of this learning cycle, both clearly outside (the Wilderness Adventure) and more closely associated with (the Community Challenge and Mentoring) participants' daily lives.

We expected, and found, stronger program effects for social than academic self-efficacy. Project K offers numerous social tasks that are likely to directly build confidence in this domain. First, participants are isolated in a wilderness context with unfamiliar peers and must work together to achieve goals that have authentic consequences. Social challenges are also intrinsic to the development of a community project and networking with community figures which occurs in the second program component. Finally, the mentoring component encourages one-to-one relationship building with an adult. In contrast, Project K is not strongly academically focused (the participants must select one academic goal and mentors support them in this regard but the program activities do not focus on academic skills directly). Much of its impact on academic efficacy was thought to arise from the strong program focus on learning transfer. This is always going to be a less reliable process, given that there is clearly a large context or task specific component to self-efficacy (see Voskuil & Robbins, 2015).

While Project K showed a positive impact on academic and social self-efficacy overall, we also found some variation across participant subgroups. It is interesting that, in almost every case, this pointed to an additional boost for those who needed it most. In the social domain, Project K had stronger efficacy-enhancing effects for Asian participants, a group that reported significantly lower social self-efficacy at baseline and whose parents also indicated had they lower interpersonal skills at the beginning of the program. On the other hand, while there was evidence that Pacific and Māori

youth in the Project K group had improved social self-efficacy when compared with the Pacific or Māori Control participants, stronger effects were obtained for non-Māori and non-Pacific participants. Both Māori and Pacific participants, however, had higher average rates of social self-efficacy at baseline, regardless of which program condition they were in.

The possibility that Project K was more effective at boosting efficacy for those who most needed it was partially supported in the academic domain. One year after program completion there was evidence that Project K was more effective in boosting academic self-efficacy for students who started the program with very low academic self-efficacy and for students from low SES schools. Nevertheless, Māori Project K participants, who also reported significantly lower academic self-efficacy at baseline, showed only average gains in this domain.

So while we hypothesized the program would be effective across participant subgroups, the finding that it generally had more impact on those with further to go is not necessarily problematic. Notably, other studies have also found that youth development programs are most effective for those who enter with greater disadvantage or who have less developed incoming skills. For example, Scales, Benson, Roehlkepartain, Sesma, & van Dulmen's (2006) found that service-learning programs in the U.S. appear to close the achievement gap between students from high and low socioeconomic schools. DuBois et al. (2011) also found that youth mentoring programs are more effective for individuals with some degree of individual or environmental risk compared to those at low or extreme levels of risk.

One finding from Hollis et al.'s (2011) study of Māori participants' experiences within Project K may provide some insight as to why this pattern occurred. Five of the six Project K graduates interviewed indicated that the program atmosphere was one of equality in that the expectations were the same for all participants. By conveying equal expectations to all participants, Project K may shift the negative self-perceptions of some participants towards the higher levels of their peers. This also suggests it is useful for programs to have participants operating at a range of levels in regard to the target skills, because peers with greater skills allow opportunities for vicarious learning. Furthermore, if all participants are treated with sensitivity, the more skilled peers may benefit from being able to demonstrate leadership within the relevant domain.

At the same time, while Project K did not fail Māori students in the academic domain, it could potentially work harder to improve their academic self-efficacy, given their relatively low starting point. It is notable that our study assessing Project K's theory of change found the cultural component was stressed strongly by the program developers, but not so strongly by the program deliverers (Deane & Harré, 2014b). This is something that could fruitfully be addressed.

This was a real world program and absolute control of the experience of participants was not possible (or desirable). While there is extensive training given to Project K program deliverers, a detailed implementation manual, and other checks of fidelity, participants will not have received a completely consistent experience. Some will also be more motivated to change than others and there are numerous other factors that are likely to influence how participants respond to the program. All of these factors can bias true estimates of the program's influence. Additional moderating variables recognized to have an important impact on program effectiveness (e.g., participant engagement and facilitator support; Deane & Harré, 2014a) have been collected with more recent Project K programs for future investigations. In a similar vein, we are developing observational studies in an attempt to capture the link between the iterative learning cycle at the core of Project K and participant engagement and outcomes.

While this was a randomized controlled trial study, it is possible that control group participants received different youth development services while their peers were experiencing Project K. Thus, the intended contrast between the control and program participants may be muted. Furthermore, random assignment may serve to create equivalent groups at baseline but it is difficult to control participant retention over the course of an evaluation. In the current study, there was differential drop-out across the program conditions and across demographic subgroups. While we attempted to mitigate this bias by statistically adjusting for ethnic group covariates and baseline efficacy levels as well as by conducting comparative analyses on a propensity matched sample, the participants in each condition may have varied on other characteristics. The non-respondents and school leavers not included in the analyses did differ from those whose data we could use. The generalizability of the results are, therefore, restricted.

The ethnic group effects in particular should also be interpreted with caution because of the differential attrition across the intervention and control groups at both time points and small subgroup sample sizes for the one year post-program follow up. The participant numbers at baseline were reasonable for Māori and Pacific participants but these diminished substantially one year post-program. There were only 34 Māori and 22 Pacific youth in the PK group, and 29 Māori and 21 Pacific in the Control group at this time point. The Asian subgroup was initially small and further diminished to 19 in the Project K group and 15 in the Control. We suggest that others consider over-sampling indigenous and ethnic minority participants in future large scale evaluations of youth development programs to enable more meaningful analysis of ethnic differences.

The findings should not be extrapolated beyond young people exhibiting a "mid-level" of risk. As indicated earlier, young people expressing more severe mental health concerns or who are exhibiting other high risk behaviors are excluded from the program for safety reasons. We, therefore, cannot assume that Project K would be effective in improving self-belief outcomes for youth at a high level of risk.

Further, there are well-recognized potential biases associated with self-report due to social desirability, memory effects and/or comprehension difficulties (Crockett et al., 1987). We included parents' perspectives of the participants' interpersonal skills to enable triangulation with the social self-efficacy self-reports. Unfortunately, these could only be collected immediately after the program due to logistical constraints. Similarly, the initial evaluation design also included teachers' ratings of both academic and social competencies. This third perspective would have further enhanced the credibility of the findings, but we were unable to gain sufficient participation rates post-program to include teachers' ratings.

With regards to a final limitation, we have argued that fostering positive self-efficacy is of particular importance in adolescence because efficacy beliefs influence the life-decisions young people must make. The ultimate end is thus to see an impact on goal-directed behaviors, not just beliefs. In this article, we have stopped short of this because it was important to first establish whether or not Project K is effective in addressing its immediate objective of boosting self-efficacy. We are

currently investigating the impact of Project K on behavioral outcomes and the role of self-efficacy in mediating these effects.

Conclusion

For youth development programs to be effective, they must assist young people in navigating the everyday domains that make up their lives. This is a difficult task, and previous meta-analyses of youth development programs have found mixed (DuBois, Portillo, Rhodes, Silverthorn & Valentine, 2011; Durlak, Weissberg, & Pachan, 2010; Hattie, Marsh, Neill & Richards, 1997; Perry et al., 2012; Wilson & Lipsey, 2000) or no (Morton & Montgomery, 2013) positive impact on self-efficacy. Our study found that Project K, which provides participants with an adventure element, community exploration and individual mentoring, boosted participants' self-efficacy in two key everyday domains: academic and social. We suggest this is likely due to Project K's focus on multiple iterations of an experiential, self-regulation learning cycle and explicit attention to making connections between achievements in the program context and outside domains. Notably, stronger effects were obtained for several participant subgroups that initially had lower efficacy. This may be in part because they were positively influenced by their more confident peers. This suggests programs should not necessarily only target those most in need. Clearly, further research is needed, and is being undertaken, to uncover the most important factors in Project K's success. We strongly recommend that youth program developers and evaluators place close attention to processes that encourage participants to take their learning within the program to the world outside.

Compliance with Ethical Standards

Funding: The New Zealand Ministry of Social Development (MSD) funded FYD for the evaluation for eight program deliveries included in the evaluation presented here. The study was also funded in part from two scholarships to the first author, a New Zealand International Doctoral Research Scholarship from the New Zealand government and a University of Auckland Doctoral Research Scholarship.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent: Informed consent was obtained from all individual participants included in the study.

References

Allen, J. P., Leadbeater, B. J., & Aber, J. L. (1990). The relationship of adolescents' expectations and values to delinquency, hard drug use, and unprotected sexual intercourse. *Development and Psychopathology*, *2*(01), 85-98. doi: doi:10.1017/S0954579400000614

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191-215.

Bandura, A. (1997). *Self-efficacy. The exercise of control*. New York, NY: W.H. Freeman and Company.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Multifaceted impact of self-efficacy beliefs on academic functioning. *Child Development, 67*(3), 1206-1222.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (2001). Self-efficacy beliefs as shapers of children's aspirations and career trajectories. *Child Development*, *72*(1), 187-206.

Bandura, A., Pastorelli, C., Barbaranelli, C., & Caprara, G. V. (1999). Self-efficacy pathways to childhood depression. *Journal of Personality and Social Psychology, 76*(2), 258-269.

Bauman, A. E., Reis, R. S., Sallis, J. F., Wells, J. C., Loos, R. J. F., & Martin, B. W. (2012). Correlates of physical activity: why are some people physically active and others not? *Lancet 380*, 258-271.

Benson, P. L. (2007). Developmental assets: An overview of theory, research, and practice. In R. K. Silbereisen & R. M. Lerner (Eds.), *Approaches to positive youth development* (pp. 33-58). London, England: Sage Publications.

Brooks-Gunn, J. & Roth, J. (2014). Invited commentary: Promotion and prevention in youth development: Two sides of the same coin? *Journal of Youth and Adolescence, 43*, 1004-1007. doi: 10.1007/s10964-014-0122-y

Caines, K. & Center, B. (2010). Differential effects of an outdoor community-based youth development program on adolescent boys and girls. *Journal of Adolescent Health, 46*(2), S73-S73. doi: 10.1016/j.jadohealth.2009.11.176

Caprara, G. V., Fida, R., Vecchione, M., Del Bove, G., Vecchio, G. M., Barbaranelli, C., & Bandura, A. (2008). Longitudinal analysis of the role of perceived self-efficacy for self-regulated learning in academic continuance and achievement. *Journal of Educational Psychology, 100*(3), 525-534. doi: 10.1037/0022-0663.100.3.525

Caprara, G. V., Gerbino, M., Paciello, M., Di Giunta, L., & Pastorelli, C. (2010). Counteracting depression and delinquency in late adolescence: The role of regulatory emotional and interpersonal self-efficacy beliefs. *European Psychologist, 15*(1), 34-48.

Carroll, A., Houghton, S., Wood, R., Unsworth, K., Hattie, J., Gordon, L., & Bower, J. (2009). Self-efficacy and academic achievement in Australian high school students: The mediating effects of academic aspirations and delinquency. *Journal of Adolescence, 32*(4), 797-817. doi: 10.1016/j.adolescence.2008.10.009

Catalano, R. F., Berglund, M. L., Ryan, J. A. M., Lonczak, H. S., & Hawkins, J. D. (2004). Positive youth development in the United States: Research findings on evaluations of positive youth development programs. *The Annals of the American Academy of Political and Social Science, 591*(1), 98-124. doi: 10.1177/0002716203260102

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indices for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255. doi:10.1207/S15328007SEM0902_5

Connolly, J. (1989). Social self-efficacy in adolescence: Relations with self-concept, social adjustment, and mental health. *Canadian Journal of Behavioural Science, 21*(3), 258-269.

Corrigan, A. (2002). Social Competence Scale – Parent Version. Fast Track project technical report. Retrieved from http://www.fasttrackproject.org/techrept/s/scp/scp2tech.pdf

Craig, S. L., McInroy, L., Austin, A., Smith, M., & Engle, B. (2012). Promoting Self-Efficacy and Self-Esteem for Multiethnic Sexual Minority Youth: An Evidence-Informed Intervention. *Journal of Social Service Research, 38*(5), 688-698.

Crockett, L. J., Schulenberg, J. E., & Petersen, A. (1987). Congruence between objective and self-report data in a sample of young adolescents. *Journal of Adolescent Research, 2*(4), 383-392.

Damon, W. (2004). What is positive youth development? *The Annals of the American Academy of Political and Social Science, 591,* 13-24. doi: 10.1177/0002716203260092

Deane, K. L., & Harré, N. (in press). Developing a thoughtful approach to evaluation: Values-driven guidelines for novice evaluators. *Evaluation Matters – He Take Tō Te Aromatawai.*

Deane, K. L., & Harré, N. (2014a). The youth adventure programming model. *Journal of Research on Adolescence, 24*(2), 293-408. doi: 10.1111/jora.12069

Deane, K. L. & Harré, N. (2014b). Program theory-driven evaluation science in a youth development context. *Evaluation & Program Planning, 45,* 61-70. doi: 10.1016/j.evalproplan.2014.03.009

Desivilya, H. S., & Eizen, D. (2005). Conflict management in work teams: The role of social self-efficacy and group identification. *The International Journal of Conflict Management, 16*(2), 183-208.

Dicke, T., Marsh, H. W., Parker, P. D., Kunter, M., Schmeck, A., & Leutner, D. (2014). Self-Efficacy in Classroom Management, Classroom Disturbances, and Emotional Exhaustion: A Moderated Mediation Analysis of Teacher Candidates. *Journal of Educational Psychology, 106*(2), 569-583.

DuBois, D. L., Portillo, N., Rhodes, J. E., Silverthorn, N., & Valentine, J. C. (2011). How effective are mentoring programs for youth? A systematic assessment of the evidence. *Psychological Science in the Public Interest, 12*(2), 57-91. doi: 10.1177/1529100611414806

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*, 294-309. doi: 10.1007/s10464-010-9300-6

Durlak, J. A., & Weissberg, R. P. (2007). *The Impact of After-School Programs that Promote Personal and Social Skills*. Collaborative for Academic, Social, and Emotional Learning. Chicago, IL. Retrieved from www.casel.org

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42*(3), 509-529. doi:10.1080/00273170701382864

Field, A. (2009). *Discovering Statistics Using SPSS*. London: Sage Publications.

Glass, J. S. & Benshoff, J. M. (2002). Facilitating group cohesion among adolescents through challenge course experiences. *The Journal of Experiential Education, 25*(2), 268-276.

Glatz, T., & Buchanan, C. M. (2015). Change and Predictors of Change in Parental Self-Efficacy From Early to Middle Adolescence. *Developmental Psychology, 51*(10), 1367-1379.

Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2010). *Multivariate Data Analysis. Seventh Edition*. Upper Saddle River, New Jersey: Prentice Hall.

Hall, B. C., Stewart, D. G., Arger, C., Athenour, D. R., & Effinger, J. (2014). Modeling Motivation Three Ways: Effects of MI Metrics on Treatment Outcomes Among Adolescents. *Psychology of Addictive Behaviors, 28*(1), 307-312.

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*(1), 43-87. doi: 10.3102/00346543067001043

Have, T., Normand, S. T.., Marcus, S., Hendricks, B. C., & Lavori, P. (2008). Intent-to-treat vs. non-intent-to-treat analyses under treatment adherence in mental health randomized trials. *Psychiatric Annals, 38*(12), 772-783.

Hayhurst, J., Hunter, J. A., Kafka, S., & Boyes, M. (2015). Enhancing resilience in youth through a 10-day developmental voyage. *Journal of Adventure Education and Outdoor Learning, 15*(1), 40-52.

Hazenberg, R., Seddon, F., & Denny, S. (2015). Programme recruitment and evaluation: the effect of an employability enhancement programme on the general self-efficacy levels of unemployed graduates. *Journal of Education and Work, 28*(3), 273-300.

Hermann, K. S. & Betz, N. E. (2006). Path models of the relationships of instrumentality and expressiveness, social self-efficacy, and self-esteem to depressive symptoms in college students. *Journal of Social & Clinical Psychology, 25*(10), 1086-1106.

Høigaard, R., Kovač, V. B., Øverby, N. C., & Haugen, T. (2014). Self-Efficacy mediates the effects of school psychological climate on academic achievement. *School Psychology Quarterly, Advance online publication*. doi: 10.1037/spq0000056

Hollis, H., Deane, K., Moore, J., & Harré, N. (2011). Young Maori perceptions of a youth development program. *Kōtuitui: New Zealand Journal of Social Sciences Online, 6*(1-2), 50-61. doi: 10.1080/1177083X.2011.615847

Hox, J. (2002). *Multilevel Analysis. Techniques and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Jeličić, H., Bobek, D., Phelps, E., Bobek, D., Lerner, J. V., & Lerner, R. M. (2007). Using positive youth development to predict contribution and risk behaviors in early adolescence: Findings from the first two waves of the 4-H study of Positive Youth Development. *International Journal of Behavioral Development, 31*(3), 263−273. doi: 10.1177/0165025407076439

Jungert, T., Hesser, H., & Traff, U. (2014). Contrasting two models of academic self-efficacy – domain-specific versus cross-domain – in children receiving and not receiving special instruction in mathematics. *Scandinavian Journal of Psychology, 55*, 440-447.

Kim, J., & Cicchetti, D. (2003). Social self-efficacy and behavior problems in maltreated and nonmaltreated children. *Journal of Clinical Child and Adolescent Psychology, 32*(1), 106-117.

Lanza, S. T., Moore, J. E., Butera, N. M. (2013). Drawing causal inferences using propensity scores: A practical guide for community psychologists. *American Journal of Community Psychology, 52*, 380-392. doi: 10.1007/s10464-013-9604-4

Lennings, C. J. (1994). An evaluation of a generalized self-efficacy scale. *Personality and Individual Differences, 16*(5), 745-750.

Litt, M. D., & Kadden, R. M. (2015). Willpower versus "Skillpower": Examining how self-efficacy works in treatment for marijuana dependence. *Psychology of Addictive Behaviors, 29*(3), 532-540.

McArdle, J. J. (2007). Factor analysis of longitudinal data. In R. Cudeck & R. MacCallum (Eds.), *Factor Analysis at 100: Historical developments and future directions* (pp. 99-130). Mahwah, NJ: Erlbaum.

McKenzie, M. (2003). Beyond "The Outward Bound Process:" Rethinking student learning. *The Journal of Experiential Education, 26*(1), 8-23.

Moore, J. (2005). Self-efficacy and health behaviours: A test of measures to assess the effectiveness of a positive youth development program. (Unpublished master's thesis). University of Auckland, Auckland.

Morton, M. H. & Montgomery, P. (2013). Youth empowerment programs for improving adolescents' self-efficacy and self-esteem. A systematic review. *Research on Social Work Practice, 23*(1), 22-33. doi: 10.1177/1049731512459967

Muris, P. (2001). A brief questionnaire for measuring self-efficacy in youths. *Journal of Psychopathology and Behavioral Assessment, 23*(3), 145-149. doi: 10.1023/A:1010961119608

Muthén, L. K., & Muthén, B. O. (2015). *Mplus User's Guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nunns, Peace & Witten (2015). Evaluative reasoning in public-sector evaluation in Aotearoa New Zealand: How are we doing? *Evaluation Matters – He Take To Te Aromatawai, 1*, 137- 163. doi: 10.18296/em.0007

Orren, P. & Werner, P. (2007). Effects of brief wilderness programs in relation to adolescents' race. *The Journal of Experiential Education, 30*(2), 117-133.

Pajares, F. (2002). Gender and perceived self-efficacy in self-regulated learning. *Theory Into Practice, 41*(2), 116.

Parker, R., Bush, J. & Harris, D. (2014). Important methodological issues in evaluating community-based interventions. *Evaluation Review, 38*(4), 295-308. doi: 10.1177/0193841X14542610

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2011). HLM 7 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.

Rodgers, W. M., Murray, T. C., Courneya, K. S., Bell, G. J., & Harber, V. J. (2009). The specificity of self-efficacy over the course of a progressive exercise programme. *Applied Psychology: Health and Well-Being, 1*(2), 211-232.

Roth, J. L., & Brooks-Gunn, J. (2003). Youth development programs: Risk, prevention and policy. *Journal of Adolescent Health, 32*(3), 170-182. doi: 10.1016/s1054-139x(02)00421-4

Scales, P. C., Benson, P. L., Leffert, N., & Blyth, D. A. (2000). Contribution of developmental assets to the prediction of thriving among adolescents. *Applied Developmental Science, 4*(1), 27-46. doi:10.1207/S1532480XADS0401_3

Scales, P. C., Benson, P. L., Roehlkepartain, E. C., Sesma, A., Jr., & van Dulmen, M. (2006). The role of developmental assets in predicting academic achievement: A longitudinal study. *Journal of Adolescence, 29*(5), 691-708. doi:10.1016/j.adolescence.2005.09.001

Schunk, D. H., & Meece, J. L. (2006). Self-Efficacy Development in Adolescence. In F. Pajares & T. Urdan (Eds.), *Self-Efficacy Beliefs of Adolescents* (pp. 71-96). Greenwich, CT: Information Age Publishing.

Sibley, C. (2009). Utilities for examining interactions in multiple regression [Computer software]. Auckland: University of Auckland.

Snijders, T., & Bosker, R. (1999). Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling. London: Sage Publications.

Tipton, R. M., & Worthington, E. L. (1984). The measurement of generalized self-efficacy: A study of construct validity. *Journal of Personality Assessment, 48*, 545-548.

Voskuil, V. R., & Robbins, L. B. (2015). Youth physical activity self-efficacy: a concept analysis. *Journal of Advanced Nursing 71*(9), 2002-2019.

Voorhees, C. M., Brady, M. K., Calantone, R., & Ramirez, E. (2016). Discriminant validity testing in marketing: an analysis, causes for concern, and proposed remedies. *Journal of the Academy of Marketing Sciences, 44*(1), 119-134. doi: 10.1007/s11747-015-0455-4

Warner, L. M., Wolff, J. K., Wurm, S., Schuz, B., Parschau, L., & Schwarzer, R. (2014). Sources of self-efficacy for physical activity. *Health Psychology, 33*(11), 1298-1308.

Wichmann, C., Coplan, R. J., & Daniels, T. (2004). The Social Cognitions of Socially Withdrawn Children [Article]. *Social Development, 13*(3), 377-392. doi: 10.1111/j.1467-9507.2004.00273.x

Widmer, M. A., Duerden, M. D., & Taniguchi, S. T. (2014). Increasing and generalizing self-efficacy. *Journal of Leisure Research, 46*(2), 165-183.

Wigfield, A., Eccles, J. S., Schiefele, U., Roeser, R. W., & Davis-Kean, P. (2006). Development of Achievement Motivation. *Handbook of child psychology: Vol 3, Social, emotional, and personality development (6th ed)* (pp. 933-1002). Hoboken, NJ: John Wiley & Sons Inc; US.

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), 1-12. doi: 10.1016/s0149-7189(99)00040-3

Zimmerman, B. J., & Cleary, T. J. (2006). Adolescents' development of personal agency. The role of self-efficacy beliefs and self-regulatory skill. In F. Pajares & D. Miller (Eds.). *Self-efficacy beliefs of adolescents* (pp. 45-69). Greenwich, CT: Information Age Publishing

Table 1

Descriptive Statistics for the Outcomes of Interest and Between Group Effect Sizes

Variable	Project K		Control		Effect Size
	n	*Means (SD)*	*n*	*Means (SD)*	*d [C.I.]*
Time 0 Academic Self-Efficacy	573	3.24 (.92)	576	3.28 (.91)	-.04 [-.10 – .01]
Time 1 Academic Self-Efficacy	482	3.73 (.95)	417	3.33 (.92)	.43 [.37 - .49]
Time 2 Academic Self-Efficacy	255	3.64 (.92)	235	3.39 (.92)	.27 [.19 - .36]
	210	**3.70 (.89)**	**210**	**3.38 (.92)**	**.35 [.27-.44]**
Time 0 Social Self-Efficacy	573	3.96 (.88)	576	4.01 (.89)	-.06 [-.11 – -.01]
Time 1 Social Self-Efficacy	482	4.50 (.79)	417	4.13 (.89)	.44 [.39-.50]
Time 2 Social Self-Efficacy	255	4.54 (.75)	235	4.25 (.86)	.36 [.29 - .43]
	210	**4.56 (.71)**	**210**	**4.21 (.87)**	**.44 [0.37-.52]**
Time 0 Parents' Perceptions of Interpersonal Skills	318	3.40 (.72)	163	3.43 (.76)	-.04 [-.11 - .02]
Time 1 Parents' Perceptions of Interpersonal Skills	318	3.74 (.64)	163	3.52 (.70)	.33 [.28 - .39]

Note. Values in bold are derived from the propensity matched sample.

Table 2

Bivariate Correlations between Variables of Interest

Variable	*1.*	*2.*	*3.*	*4.*	*5.*	*6.*	*7.*	*8.*	*9.*	*13.*
1. Academic S.E. (Time 0)	-									
2. Academic S.E. (Time 1)	.35**	-								
3. Academic S.E. (Time 2)	.36**	.63**	-							
4. Social S.E. (Time 0)	.44**	.03	.09	-						
5. Social S.E. (Time 1)	.09*	.43**	.25**	.37**	-					
6. Social S.E. (Time 2)	.09*	.29**	.45**	.39**	.59**	-				
7. Parents' Perceptions (Time 0)	.22**	.17**	.19**	.18**	.13**	.15*	-			
8. Parents' Perceptions (Time 1)	.09	.30**	.19**	.08	.21**	.22**	.44**	-		
9. Condition	-.02	.21**	.14**	-.03	.21**	.19**	-.02	.16**	-	
10. Gender	.07*	.06	.11*	-.01	.06	.10*	.07	.03	-.02	
11. Māori	-.06*	-.13**	-.05	.09**	.06	.07	.09	-.05	-.01	
12. Pasifika	.19**	.05	-.02	.21**	.07*	.05	.15**	-.04	-.03	
13. Asian	.10**	.11**	.12**	-.08*	-.06	.00	.10*	.09	-.05	-

Table 3

Linear and Growth Models Comparing Project K and Control participants on Academic and Social Self-Efficacy

Parameters	**Academic Self-Efficacy**				**Social Self-Efficacy**			
	Full Sample		Propensity Matched Sample		Full Sample		Propensity Matched Sample	
Intercept (γ_{000})	3.30	(0.04)***	3.37	(0.05)***	4.01	(0.04)***	3.97	(0.06)***
Decile (γ_{100})	-0.04	(0.01)***		-	0.00	(0.01)		-
Condition(γ_{010})	0.04	(0.05)	0.04	(0.10)	0.02	(0.05)	0.07	(0.09)
Gender(γ_{020})	0.10	(0.06)		-	-0.03	(0.06)		-
Māori(γ_{030})	-0.20	(0.06)**		-	0.16	(0.05)***		-
Pasifika (γ_{040})	0.32	(0.06)***		-	0.48	(0.07)***		-
Asian(γ_{050})	0.42	(0.11)***		-	-0.29	(0.13)*		-
Time(γ_{100})	0.12	(0.03)***	0.12	(0.3)***	0.27	(0.02)***	0.24	(0.03)***
Decile.Time (γ_{101})	0.02	(0.01)*		-	0.01	(0.01)		-
Condition.Time (γ_{110})	**0.22**	**(0.04)*** **	**0.16**	**(0.05)** **	**0.20**	**(0.04)*** **	**0.16**	**(0.05)** **
Gender.Time (γ_{120})	-0.06	(0.05)		-	0.08	(0.03)*		-
Māori.Time (γ_{130})	-0.02	(0.06)		-	-0.00	(0.04)		-
Pasifika.Time (γ_{140})	-0.14	(0.06)*		-	-0.19	(0.06)**		-
Asian.Time (γ_{150})	-0.02	(0.08)		-	0.12	(0.06)*		-

Note. * = $p < .05$; ** = $p < .01$; *** = $p < .001$; effects of interest are in bold.

Table 4

Two-Level Full Interaction Models for Post-Program Academic and Social Self-Efficacy and Parent Perceptions of Participant Interpersonal Skills

Parameters	**Time 1 Effects**					
Regression coefficients (fixed effects)	Academic Self-Efficacy		Social Self-Efficacy		Parents' Perceptions of Interpersonal Skills	
Intercept (γ_{00})	3.53	(0.03)***	4.31	(0.03)***	3.61	(0.03)***
Decile (γ_{01})	0.00	(0.01)	0.01	(0.01)	0.02	(0.01)*
Condition (γ_{10})	**0.44**	**(0.06)*** **	**0.36**	**(0.06)*** **	**0.24**	**(0.06)*** **
Decile.Cond (γ_{11})	-0.00	(0.02)	-0.01	(0.04)	-0.03	(0.02)
Gender (γ_{20})	0.07	(0.05)	0.10	(0.04)*	-0.00	(0.05)
Māori (γ_{30})	-0.24	(0.07)***	0.05	(0.06)	-0.15	(0.09)
Pasifika Peoples (γ_{40})	-0.04	(0.08)	0.02	(0.10)	-0.14	(0.07)
Asian (γ_{50})	0.28	(0.16)	-0.09	(0.09)	0.14	(0.07)*
Cond.Gender (γ_{60})	-0.04	(0.12)	-0.02	(0.08)	0.13	(0.10)
Cond.Māori (γ_{70})	0.10	(0.13)	-0.05	(0.10)	0.22	(0.19)
Cond.Pasifika (γ_{80})	-0.14	(0.18)	**-0.45**	**(0.18)***	-0.28	(0.19)
Cond.Asian (γ_{90})	-0.01	(0.27)	0.25	(0.18)	-0.15	(0.15)
Time 1 ASE0/SSE0/PIS0 (γ_{100})	0.36	(0.04)***	0.34	(0.07)***	0.41	(0.04)***
Cond.ASE0/Cond.SSE0/Cond.PIS0 (γ_{110})	-0.06	(0.07)	-0.07	(0.07)	0.06	(0.07)

Note. * = $p < .05$; ** = $p < .01$; *** = $p < .001$; effects of interest are in bold.

Table 5

Effect sizes representing standardized mean differences of from Time 0 to Time 1 and Time 0 to Time 2 for Project K compared to Control

Project K	**Effect Size**	**95% Confidence Interval**
Paired Variables		
Academic Self-Efficacy (Time 0 – 1)	.50	[.44 – .56]
Academic Self-Efficacy (Time 0 – 2)	.37	[.29 - .45]
	.39	**[.31 - .48]**
Social Self-Efficacy (Time 0 – 1)	.64	[.59 – .70]
Social Self-Efficacy (Time 0 – 2)	.87	[.81-.94]
	.79	**[.71-.87]**
Parents' Perceptions of Interpersonal Skills (Time 0 – 1)	.50	[.45-.55]
Control		
Academic Self-Efficacy (Time 0 – 1)	.00	[-.02 – .06]
Academic Self-Efficacy (Time 0 – 2)	.04	[-.04 – .13]
	.06	**[-.02 – .15]**
Social Self-Efficacy (Time 0 – 1)	.03	[-.03 – .09]
Social Self-Efficacy (Time 0 – 2)	.34	[.26 – .42]
	.33	**[.24 – .42]**
Parents' Perceptions of Interpersonal Skills (Time 0 – 1)	.12	[.04 – .20]

Note. Values in bold are derived from the propensity matched sample.

Table 6

Two-Level Full Interaction Models for One Year Post-Program Academic and Social Self-Efficacy for Original and Propensity Matched Samples

Parameters	**Academic Self-Efficacy**				**Social Self-Efficacy**			
Regression coefficients (fixed effects)	Original Sample		Propensity Matched Sample		Original Sample		Propensity Matched Sample	
Intercept (γ_{00})	3.50	(0.05)***	3.51	(0.05)***	4.42	(0.05)***	4.39	(0.05)***
Decile (γ_{01})	0.01	(0.02)	0.02	(0.02)	0.02	(0.02)	0.00	(0.02)
Condition (γ_{10})	**0.25**	**(0.08)****	**0.27**	**(0.08)*****	**0.25**	**(0.06)*****	**0.29**	**(0.07)*****
Decile.Cond (γ_{11})	**-0.05**	**(0.02)***	**-0.07**	**(0.03)***	-0.01	(0.02)	-0.02	(0.03)
Gender (γ_{20})	0.20	(0.08)*	0.16	(0.09)	0.16	(0.05)**	-0.14	(0.07)*
Māori (γ_{30})	-0.08	(0.11)	-0.12	(0.12)	0.09	(0.10)	0.01	(0.09)
Pasifika Peoples (γ_{40})	-0.26	(0.12)*	-0.16	(0.15)	-0.03	(0.09)	-0.18	(0.11)
Asian (γ_{50})	0.20	(0.15)	0.14	(0.17)	0.10	(0.10)	-0.09	(0.10)
Cond.Gender (γ_{60})	**-0.34**	**(0.16)***	-0.28	(0.15)	-0.12	(0.13)	-0.13	(0.13)
Cond.Māori (γ_{70})	-0.22	(0.18)	-0.30	(0.19)	**-0.36**	**(0.16)***	**-0.31**	**(0.16)***
Cond.Pasifika (γ_{80})	**-0.47**	**(0.21)****	**-0.75**	**(0.26)****	-0.21	(0.18)	-0.07	(0.26)
Cond.Asian (γ_{90})	0.22	(0.24)	0.26	(0.30)	0.34	(0.20)	**0.75**	**(0.22)*****
Time 1 ASE0/SSE0 (γ_{100})	0.37	(0.05)***	0.37	(0.05)***	0.36	(0.03)***	0.36	(0.04)***
Cond.ASE0/Cond.SSE0 (γ_{110})	**-0.24**	**(0.09)****	**-0.27**	**(0.09)****	-0.14	(0.08)	-0.15	(0.08)

Note. * = $p < .05$; ** = $p < .01$; *** = $p < .001$; effects of interest are in bold.

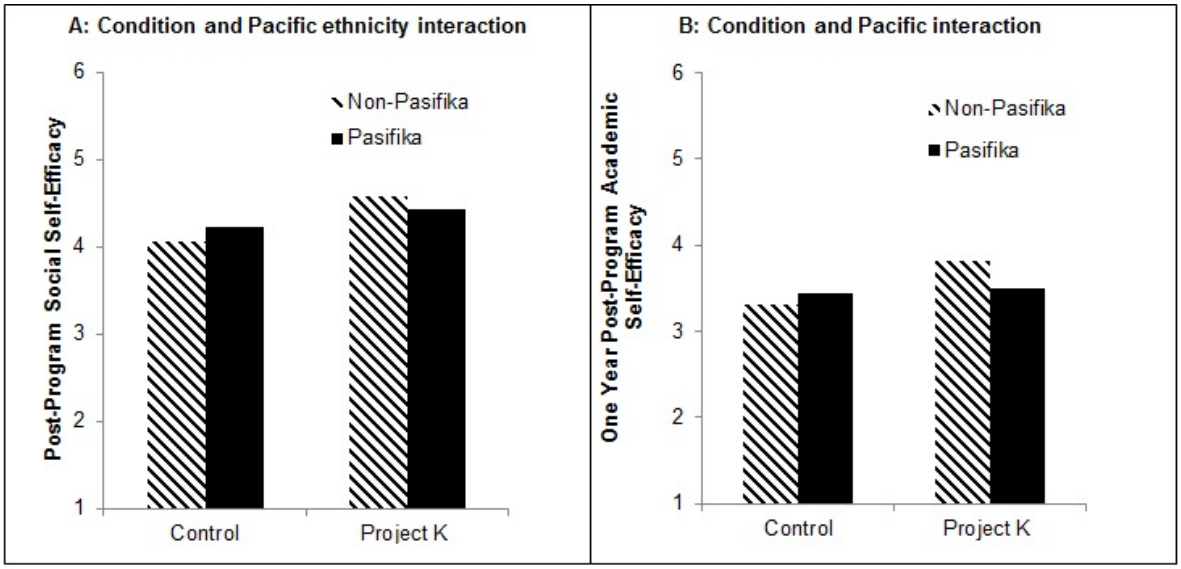

Figure 1. Condition x Pacific ethnicity interaction effects for Time 1 social self-efficacy and Time 2 academic self-efficacy.

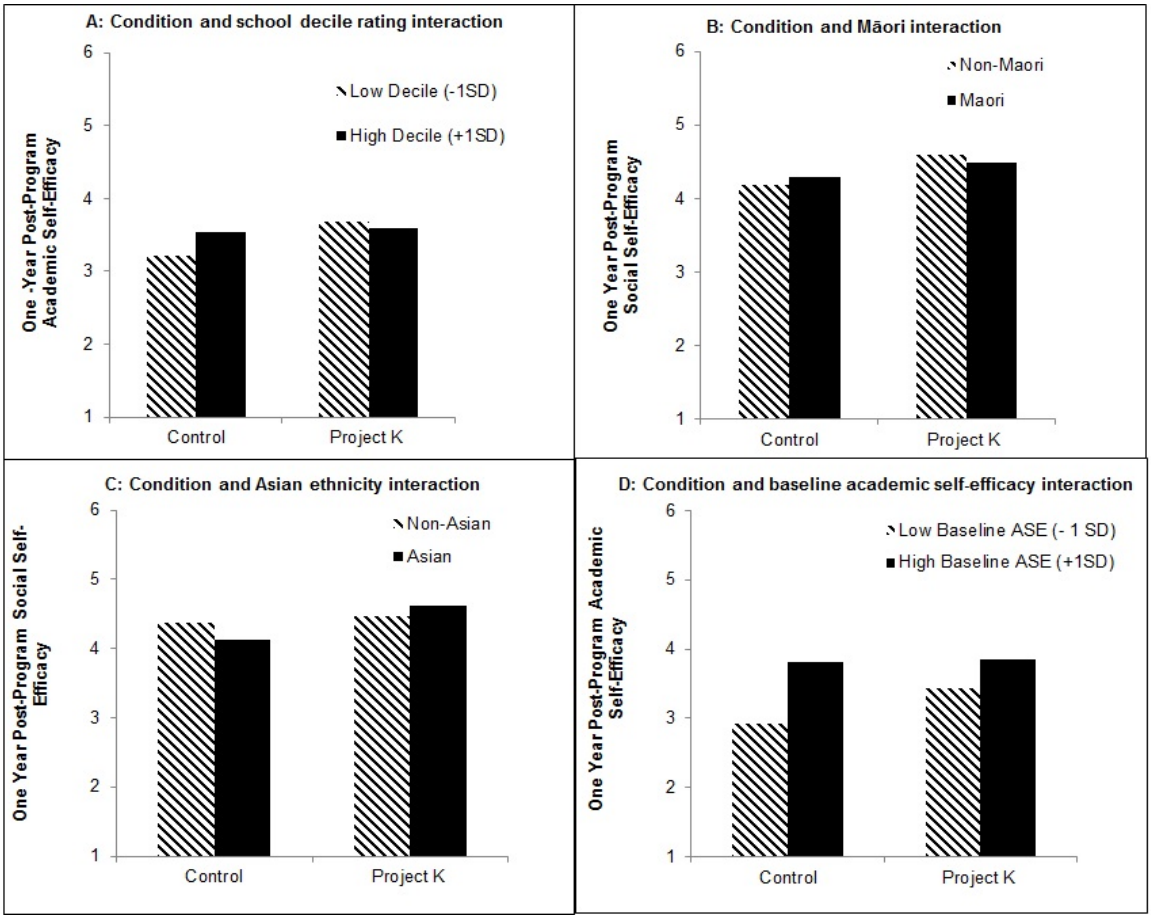

Figure 2. Interaction effects obtained at Time 2 for Condition x Māori ethnicity, Condition x School Decile, Condition x Asian ethnicity, and Condition x Baseline Academic Self-Efficacy.

Supporting Information for Online Publication Only

Appendix S1 – Tests of Model Fit, Convergent and Discriminant Validity and Invariance Tests

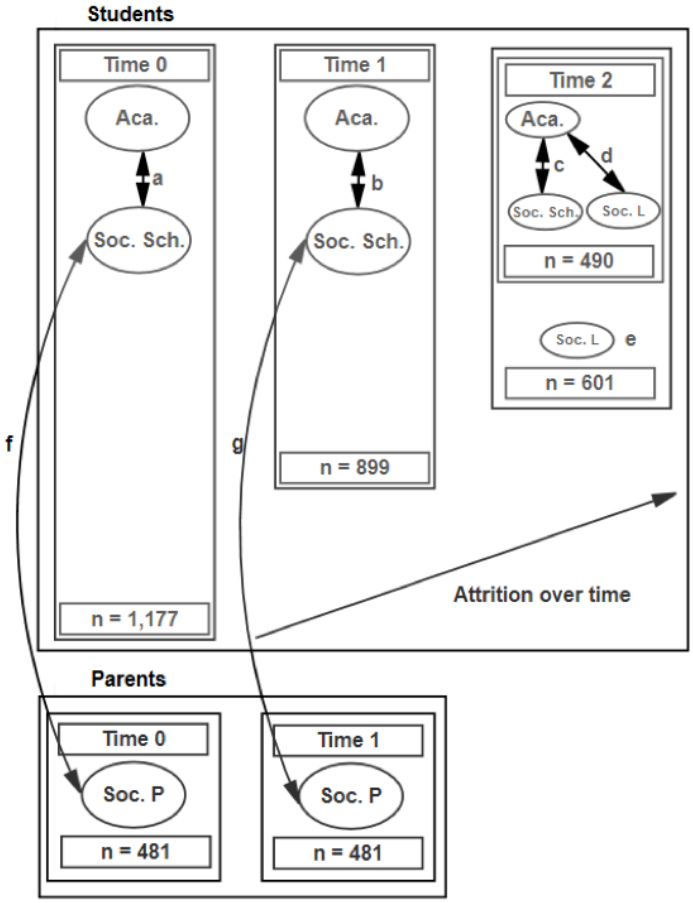

Figure S1. Confirmatory Tests of Model Fit, Convergent Validity, and Discriminant Validity for Social and Academic Efficacy at T0, T1, and T2. Note. Soc. Sch = Social Efficacy factor for school students; Acad. = Academic Efficacy factor for school students; Soc. L = Social Efficacy factor for school leavers; Soc. P = Perception of Students' Interpersonal Skills from Parents; a, b, c, d, f, g, represent confirmatory factor analyses and associated tests between two factors; e represents a single factor model and associated test for convergent validity; see model fit results in Table 1.

Table S1

Model Fit for Social and Academic Self-Efficacy and Parent Perceptions Measurement Models at Time 0, 1 and 2.

		Model Fit Indices								
Test	*N*	χ^2	*df*	χ^2/df	*p*	RMSEA	CFI	SRMR	$\hat{g}$	Pass?
					Students					
a	1,149	554.36	103	5.38	.02	**.06**	**.92**	**.05**	**.95**	Yes
b	899	585.34	103	5.68	.02	**.07**	**.91**	**.06**	**.94**	Yes
c	490	353.33	103	3.43	.06	**.07**	**.92**	**.05**	**.94**	Yes
d	490	317.81	103	3.09	.08	**.07**	**.93**	**.05**	**.95**	Yes
e	601	101.49	20	5.07	.02	**.08**	**.94**	**.04**	**.97**	Yes
					Parents					
f	481	278.23	76	3.66	.06	**.07**	**.92**	**.05**	**.94**	Yes
g	481	300.29	76	3.95	.05	**.08**	**.90**	**.05**	**.94**	Yes

Note. χ^2 = Chi-square; *df* = degrees of freedom; *p* = statistical significance associated with χ^2/df ratio; RMSEA = root mean square error of approximation; CFI = comparative fit index; SRMR = standardized root mean square residual; $\hat{g}$ = gamma hat; values in bold meet standards for acceptable fit (RMSEA ≤ .08; CFI ≥ .90; SRMR ≤ .08; $\hat{g} \geq .90$); values in bold and underlined meet standards for good fit (RMSEA ≤ .05; CFI ≥ .95; SRMR ≤ .05; $\hat{g} \geq .95$) (Fan & Sivo, 2007).

Table S2

Convergent and Discriminant Validity of Social and Academic Self-Efficacy and Parent Perception Factors at Time 0, 1, and 2

			Factorial AVE				Discriminant Validity Tests			
Test	Time	*N*	Soc. Sch.	Aca.	Soc. L	Soc. P	SV	AVE>SV	$HT/MT_{(ratio)}$	$HTMT_{\leq .85}$
	Students									
a	T0	1,149	.42	.48	-	-	.27	yes	$.23/.44_{(.52)}$	yes
b	T1	899	.46	.54	-	-	.26	yes	$.24/50_{(.48)}$	yes
c	T2	490	.50	.53	-	-	.33	yes	$.27/.50_{(.54)}$	yes
d	T2	490	-	.53	.52	-	.26	yes	$.26/.51_{(.51)}$	yes
e	T2	601	-	-	.50	-	-	-	-	-
	Parents									
f	T0	481	.44	-	-	.50	.04	yes	$.11/.47_{(.52)}$	yes
g	T1	481	.45	-	-	.50	.06	yes	$.12/.47_{(.52)}$	yes

Note. T0 = Time zero, T1 = Time one, T2 = Time three; AVE = average variance extracted; Soc. Sch = Social Efficacy factor for school students; Acad. = Academic Efficacy factor for school students; Soc. L = Social Efficacy factor for school leavers; Soc. P = Perception of Students' Social Efficacy from Parents; SV = shared variance between factors; HT = heterotrait; MT = monotrait; estimations made to three decimal places but rounded to two ($HTMT_{.85}$ and AVE-SV tests carried out in accordance with Voorhees, Brady, Calantone, & Ramirez, 2016).

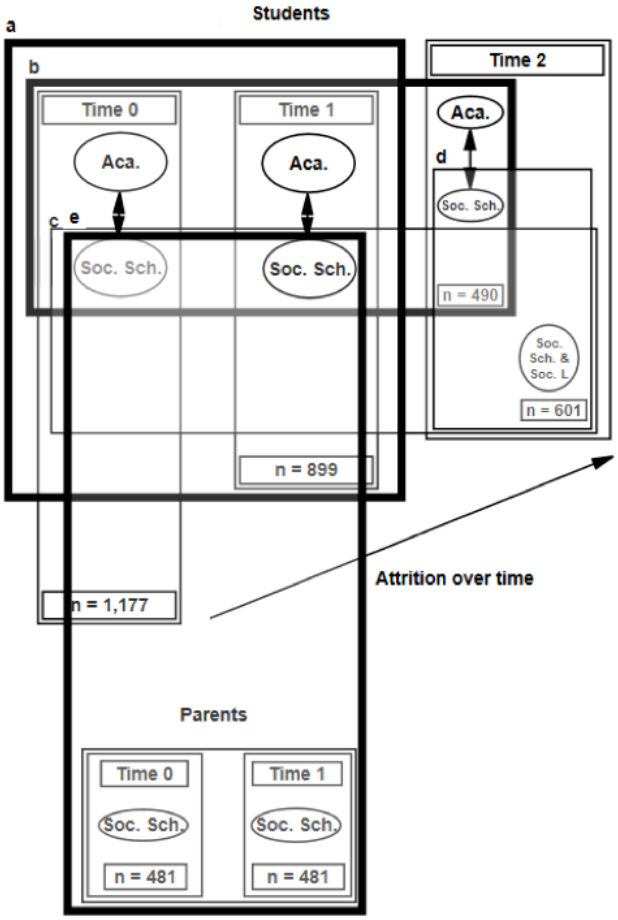

Figure 2. Invariance Tests for Social and Academic Efficacy Factors at T0, T1, and T2. *Note.* Soc. Sch = Social Efficacy factor for school students; Acad. = Academic Efficacy factor for school students; Soc. L = Social Efficacy factor for school leavers; Soc. P = Perception of Students' Interpersonal Skills from Parents; test a, b, c, d, or e designated above top left rectangle encompassing test data; tests a, b, c, & e are longitudinal tests of invariance (test e involves a total sample of n = 481), whilst test d is a single-factor multi-group test of invariance comparing 490 students with 111 leavers; an emboldened test frame denotes that the model met the requirements of measurement invariance; see summary of associated test results in Table 3.

Appendix S2 – Equation Specifications for Hierarchical Linear Models

Three-level hierarchical linear models.

The equation specifications for the linear models are provided below for the full Intention to Treat sample (using academic self-efficacy as the example outcome):

Level 1: 1.0 $ASE_{tij} = \pi_{0ij} + \pi_{1ij}(Time) + e$

Level 2: 1.1 $\pi_{0ij} = b_{00j} + b_{01j}(Condition_{ij}) + b_{02j}(Gender_{ij}) + b_{03j}(Māori_{ij}) + b_{04j}(Pasifika_{ij}) + b_{05j}(Asian_{ij}) + r_0$

Level 3: 1.2 $b_{0j} = \gamma_{000} + \gamma_{001}(DECILE) + u_{00}$

1.3 $b_{01j} = \gamma_{010} + u_{01}$

1.4 $b_{02j} = \gamma_{020}$

1.5 $b_{03j} = \gamma_{030}$

1.6 $b_{04j} = \gamma_{040}$

1.7 $b_{05j} = \gamma_{050}$

1.8 $b_{10j} = \gamma_{100} + \gamma_{101}(DECILE) + u_{10}$

1.9 $b_{11j} = \gamma_{110} + u_{11}$

2.0 $b_{12j} = \gamma_{120}$

2.1 $b_{13j} = \gamma_{130}$

2.2 $b_{14j} = \gamma_{140}$

2.3 $b_{15j} = \gamma_{150}$

The equation specifications for the linear models are provided below for the propensity matched sample (using academic self-efficacy as the example outcome):

Level 1: 2.4 $ASE_{tij} = \pi_{0ij} + \pi_{1ij}(Time) + e$

Level 2: 2.5 $\pi_{0ij} = b_{00j} + b_{01j}(Condition_{ij}) + r_0$

2.6 $\pi_{1ij} = b_{11j} + b_{01j}(Condition_{ij}) + r_0$

Level 3: 2.7 $b_{00j} = \gamma_{000} + u_{00}$

2.8 $b_{010j} = \gamma_{010} + u_{01}$

2.9 $b_{100j} = \gamma_{100} + u_{10}$

3.0 $b_{110j} = \gamma_{110} + u_{11}$

Two-level hierarchical linear models

The equation specifications for the two-level models are provided next using Time 1 (post-program) academic self-efficacy as the example outcome (the same specifications were used for academic self-efficacy at Time 2 (one year post-programme) and for both social self-efficacy outcomes with the exception that the *ASE0* and *Cond.ASE0* variables (the grand mean centred baseline academic self-efficacy score and Condition interaction term) were replaced with SSE0 and Cond.SSE0 (the grand mean centred baseline social self-efficacy score and Condition interaction term), the specifications were the same for social self-efficacy at Time 1 and Time 2. The same specifications were also used for parents' perceptions of the participants' interpersonal skills with ASE0 replaced by PIS0 and Cond.ASE0 was replaced with Cond.PIS0:

Level 1: 3.1 $ASE1_{ij} = b_{0j} + b_{1j}(Condition_{ij}) + b_{2j}(Gender_{ij}) + b_{3j}(Māori_{ij}) + b_{4j}(Pasifika_{ij}) + b_{5j}(Asian_{ij}) + b_{6j}(Condition.Gender_{ij}) + b_{7j}(Condition.Māori_{ij}) + b_{8j}(Condition.Pasifika_{ij}) + b_{9j}(Condition.Asian_{ij}) + b_{10j}(ASE0_{ij}) + b_{11j}(Cond.ASE0_{ij}) + r_{ij}$

Level 2: 3.2 $b_0 = \gamma_{00} + \gamma_{01}(DECILE) + u_{0j}$

3.3 $b_1 = \gamma_{10} + \gamma_{11}(DECILE) + u_{1j}$

3.4 $b_2 = \gamma_{20}$

3.5 $b_3 = \gamma_{30}$

3.6 $b_4 = \gamma_{40}$

3.7 $b_5 = \gamma_{50}$

3.8 $b_6 = \gamma_{60}$

3.9 $b_7 = \gamma_{70}$

4.0 $b_8 = \gamma_{80}$

4.1 $b_9 = \gamma_{90j}$

4.2 $b_{10} = \gamma_{100j}$

4.3 $b_{11} = \gamma_{110j}$

Author Biographies

Kelsey L. Deane, PhD is a Lecturer in the School of Counselling, Human Services and Social Work, Faculty of Education & Social Work at the University of Auckland. She received her doctorate in Psychology from the University of Auckland. Her major research interests include positive youth development, youth mentoring and program evaluation research.

Niki Harré, PhD is an Associate Professor in the School of Psychology at the University of Auckland, New Zealand. She received her doctorate in Psychology from the University of Auckland. Her recent research projects have focused on positive youth development, sustainable communities and schools, and political activism.

Julie Moore, MA (Hons) is the Research & Evaluation Manager for the Graeme Dingle Foundation (previously known at the Foundation for Youth Development or FYD). She received her Masters degree in Psychology from the University of Auckland. Her research interest include positive child and youth development and youth mentoring, and the sharing of evidenced-based knowledge to support the delivery of effective programmes.

Matthew G. R. Courtney, PhD is a research fellow in Quantitative Data Analysis and Research Unit (Quant-DARE) at The University of Auckland, New Zealand. He received his doctorate in Education from the University of Auckland. His research interests include the integration of social media and modern technologies in secondary and tertiary sectors, and the application of sophisticated quantitative research methodologies in educational contexts.

Author Contributions

NH and JM conceived of the study and all authors contributed to aspects of the research design. JM developed the measures under the supervision of NH. MC performed the statistical analysis for the measurement models and wrote the relevant section of the manuscript. KD took primary responsibility for the writing of the manuscript with NH and conducted the remainder of the statistical analyses. All authors reviewed and approved the final manuscript.

Conflicts of Interest

JM is the Research and Evaluation Manager for the Graeme Dingle Foundation (previously Foundation for Youth Development or FYD), the organization that owns the Project K program. She was involved in the study design, data collection and editing of the method section; however, she had no involvement in the data analysis or reporting of the findings.

KD, NH and MC report no conflicts of interest.

Acknowledgements

This study formed a part of the first author's PhD thesis that was supported by a New Zealand International Doctoral Research Scholarship from the New Zealand government and a University of Auckland International Doctoral Research Scholarship. We would also like to acknowledge the Foundation for Youth Development (now known as the Graeme Dingle Foundation) and their Community Partners for their part in implementing the evaluation and otherwise supporting the research as well as the New Zealand Ministry of Social Development for funding eight of the program deliveries included in this evaluation and for auditing the data collection process for these deliveries. We would like to thank Drs Nickola Overall, Chris Sibley and Kane Meissel, Shannon Johnston, Larissa Isted, Jade LeGrice and Diane Osborne for assistance with various aspects of the project. Finally, we would like to thank the editor of the *Journal of Youth and Adolescence* and the three blind reviewers for their incredibly helpful feedback on earlier versions of this manuscript.