FINAL REPORT

Inclusive Leadership Training pilot program

Report prepared by Dr. Michelle Stratemeyer and Prof. Robert Wood.

Proposed author list if submitted to Biorxive/ for publication

Michelle Stratemeyer, Joanne Hildebrand, Sharon Griffin, Elizabeth Bremer, Gabriella Brumatti, Nathalie Braussaud, Karen Dogget, Kate Keech, Rosemary Keogh, Lauchlan Simpson, Lucy Sullivan, Julie Bernhardt, Sarah Russel and Robert Wood

This study was designed by Michelle Stratemeyer and Prof. Robert Wood with the assistance of Julie Bernhardt, Elizabeth Bremer, Gabriella Brumatti, Nathalie Braussaud, Karen Dogget, Kate Keech, Rosemary Keogh, Lauchlan Simpson, Lucy Sullivan, Sarah Russell, Alice Tinning and Joanne Hildebrand.

The Inclusive Leadership Training modules were designed by Sharon Griffith in consultation with Dr. Michelle Stratemeyer and Prof. Robert Wood, Dr. Joanne Hildebrand, Prof. Sarah Russell and Alice Tinning.

The Inclusive Leadership Training workshops were presented by Sharon Griffith

This study and training course were made possible by funding from the Department of Industry, Innovation and Science through the Women in STEM and Entrepreneurship program, (WISE53456), and the support of the people and culture departments of the Florey Institute of Neuroscience and Mental Health, Walter and Eliza Hall Institute, Murdoch Children's Research Institute, The Peter MacCallum Cancer Centre and The Doherty Institute.

Produced by the Centre for Ethical Leadership in partnership with Women in Science Parkville Precinct (WiSPP)

Executive Summary

The Centre for Ethical Leadership, in partnership with Women in Science Parkville Precinct (WiSPP), undertook a two-year pilot program testing the efficacy of inclusive leadership training across five participating medical research institutes. This work formed part of a broader program of initiatives to increase the representation and utilisation of women in the leadership, team and work cultures of MRIs in the Parkville Precinct. The inclusive leadership training pilot program included three components, namely a qualitative evaluation of current workplaces practices, the delivery of an inclusive leadership training program tailored to the specific medical context, and the evaluation of the leadership training program. The ultimate aim of the program was to increase the retention, utilisation and development of all scientists, both female and male, at the lower, middle and upper levels of the scientific hierarchy.

The study employed a quasi-experimental, cross over design with non-randomised comparison groups who completed an Inclusive Leadership Training program based on an analysis of diversity and inclusion practices in the five participating MRIs. The total sample was made up of 30 teams (n= 220 staff registered across the length of the program). Fifteen teams completed the training in Semester 1 and 15 in teams in Semester 2. The study included pre and post training measures of opportunities for growth, retention, health and well-being, and work attitudes as well as perceptions of control over work, workplace flexibility, job satisfaction, organisational commitment, and development opportunities.

Implementation of the study encountered difficulties typical of ‘real life’ field studies, including selection effects in the composition of the comparison groups due to the difficulty of randomisation and attrition of participants across the different waves of pre and post measures. These limitations reduce confidence in inferences drawn regarding the impacts of the training. However, several interesting findings with implications for the leadership of teams within the MRIs, were found in the descriptive statistics, covariance analyses and group comparisons between men and women and part-time and full-time staff.

The main findings were:

- Existing levels of culture and work attitudes reported by study participants were consistent with the expected result for a positive work culture.
- The reported levels of stress and work life conflict were higher than expected.
- Researchers who reported greater control over decisions about how, when and where their work was done also reported more positive work attitudes, including job satisfaction.
- Researchers who reported greater access to flexible work arrangements also reported lower levels of work family conflict and lower stress.
- Male researches reported having greater control over work and greater personal growth at work than their female counterparts.
- Part time researchers, who were predominantly women, reported fewer opportunities for personal development and more work-family conflict.

These findings and others reported in the full text, can serve as a starting point for team discussions of how to increase inclusion, wellbeing and productivity of researchers. Despite the limitations presented by the study design, these exploratory findings offer the opportunity for potentially productive engagement.

Introduction

Why diversity and inclusion?

Diversity and inclusion are two core concepts in organizational psychology. Diversity refers to the collective differences among group members, at either the surface level e.g., age and gender, or at a deeper, less immediately observable level, such as cognitive abilities and experiences (Harrison, Price, & Bell, 1998). Inclusion refers to the degree to which individual workers feel they belong, are respected for their uniqueness, appropriately utilised, and are psychologically safe in the workplace (Janssens & Zanoni 2008; Nishii, 2013; Roberson, 2006; Shore et al. 2011). The benefits of diversity are realised through inclusive practices and cultures that enable diverse individuals, particularly the most diverse and those in the minority to contribute effectively to teams and organizations.

There is a strong business case for improving organizational diversity and inclusion. At the organizational level, Fortune 500 companies with gender diverse boards show 42% greater return on sales and 53% higher return on equity (Joy, Carter, Wagner, & Narayanan, 2007). Companies with more than one woman on their board of executives outperformed those without women by 26%. More gender and racially diverse boards had a 53% higher return on investment and business margins were also 14% higher on average. Organizational gender and racial diversity improve sales revenue, relative profit, market share, and grow the customer base (Herring, 2009). Finally, having diverse employees can be beneficial in an increasingly global economy where organisations must respond to the diverse needs of their customer bases (Konrad, 2003). Teams with a member who shares the ethnicity of the client base are 152% more likely than other teams to understand the needs of the client (Hewlett, Marshall, & Sherbin, 2013).

At the team level, which is the focus of the WiSPP Inclusive Leadership Training program, diversity has been found to have a range of benefits. Diverse groups have a higher likelihood of possessing a variety of different skills, values, beliefs, approaches, and knowledge bases, which lead to improved problem solving, creativity and innovation (Cox, 2001; Roberge & van Dick, 2010; Shin, Kim, Lee, & Bian, 2012). Teams with high diversity are more creative (Shin et al., 2012), and are better at challenging conventional assumptions (Gurin, Nagda, & Lopez, 2004). Cognitive diversity in teams improves decision making and problem solving (Hong & Page, 2001) over and above the performance of teams comprised of best-performing staff (Hong & Page, 2004).

Alongside the benefits of diversity sit the risks of fault lines, which lead individual and groups to divert productive energies into potentially destructive behaviours in order to maintain existing power structures, status differentials, and identities. Capturing the benefits of diversity requires inclusive practices and cultures, supported by policies and processes that bolster the psychological safety and personal efficacy of all individuals. Leadership is critical in creating the conditions of psychological safety and team efficacy that enable teams to realise the potential benefits of diversity.

In summary, having a diverse workforce is not just socially and ethically important, but is strongly supported by a robust business case based on the demonstrated benefits at team and organizational levels. Diversity, now more than ever, needs to be an integrated and prioritised aspect of workplace culture. At the team level inclusive leadership can make the difference between the productive and destructive effects of diversity.

Diversity and Inclusion in STEMM domains

Academia faces similar challenges to corporate workplaces. Academics who collaborate frequently and prolifically are generally more successful (Arora, Mittal, & Pasari, 2011). In STEMM fields, over 90% of publications are collaborative, with collaborative research leading to higher impact publications and commercial realisations of products (Bozeman & Boardman, 2014). Effective medical research teams are characterised by having multiple staff members with different functions, including academic, administrative, managerial, and support staff within and across institutions.

Research has shown that underutilization of diverse workers is a problem in academia, and specifically in STEMM (SAGE, 2016). In Australia, women outnumber men in all undergraduate degrees, with the exception of STEMM fields where 49% of students enrolled in a bachelor's degree are women (SAGE, 2016). Although this figure is almost at parity, removing medicine from the mix (and analysing STEM cohorts) reduces female representation to 33% of enrolled students. In medicine and health sciences, women outnumber men in student cohorts, as well as in early-career roles. But, as seniority increases, women drop in numbers until men become the majority of leaders at levels C and D. The attrition of women as one moves from less senior to more senior academic roles is affected by a wide range of factors, including work family conflicts. However, the experiences of single women at work often do not differ from those of colleagues in relationships, with or without children. They also have much in common with the experiences of ethnic minorities, both male and female, that of feeling excluded and under supported in their career, especially in the lack of structural support for their leadership ambitions.

The loss through attrition and underutilization of medical research practitioners, administrators, and researchers, whose training is funded by government resources, represents a significant drain on the innovation and productivity of medical research institutes. It also has a potentially negative impact on the well-being of Australian society due to the reduced scientific contribution of medical researchers and the breakthroughs that they can produce.

What, then, causes the disparities in progression and opportunities for diverse employees in medical research? First, the perception of research leaders is still biased in favour of males. The perfect academic is still one perceived to have no outside interests or responsibilities (Bailyn, 2003) and who is unencumbered by the burden of family life (Williams, 2000). The ideal lab head is seen as an "all-rounder" who is a single-minded genius (Lucht, 2014), characteristics that are less often ascribed to women and ethnic minority groups (Storage, Horne, Cimpian, & Leslie, 2016).

Second, unconscious biases and beliefs may have an impact on career progression. These effects are well documented for women and ethnic minorities. The Diversity Council of Australia (O'Leary & Tilly, 2013) found Asian employees feel disadvantaged in seeking out leadership roles due to incongruency between cultural values and expectations about how leaders should behave. Asian participants noted that they were less likely to be self-promoting but felt that this disadvantaged them in an environment where being able to talk up one's abilities resulted in promotions. Similarly, values around deference to authority and respect for the decisions of leaders mesh poorly with the positive associations in Western workplaces of challenging leadership, asking for career growth opportunities, and being a vocal contributor in team meetings.

Third, minority employees are more likely to network with others who are similar to them (e.g., Rothstein & Davey, 1995), potentially restricting their access to information about promotions, training, and other opportunities. Relatedly, STEM students report that it was important for them to have a mentor who was of the same race and/or gender to them, feeling this offered them more help than having a mentor from a different background (Blake-Beard, Bayne, Crosby, & Muller, 2011). This finding is consistent with evidence that individuals learn more from role models who are perceived as similar (Bandura, 1997). Nonetheless, having a mentor or sponsor of the same demographic background as the protégé can limit access to diverse ideas and information and reinforce the same patterns of achievement as existing leaders.

Finally, research shows that academia can be characterised by unequal divisions of labour and resource allocation that unfairly penalise women and ethnic minority workers, who are more likely to be assigned pastoral and administrative duties that are weighted less heavily than research and teaching in promotions. The gendered division of labour apparent in academia, particularly in STEM is only ameliorated when a critical mass of women faculty is reached (Carrigan, Quinn, & Riskin, 2011). Qualitative work supports a commonly held belief that women receive fewer resources and privileges than men in STEM organisations (Greene, Stockard, Lewis, & Richmond, 2010). After controlling for scholarly productivity, women attain tenure more slowly than men do. This cannot be explained by lower performance, as women publish work which is more highly cited than men, showing a higher standard of performance and academic rigour (Hewlett, 2002).

We have thus far elaborated upon the benefits of organizational and team-based diversity, and the positive impact of diversity in STEMM contexts. But, as stated in the introduction, diversity is not a universally positive benefit to organisations and teams. With the benefits of diversity comes specific difficulties that may negatively impact team efficacy. For example, while diverse teams may bring different forms of knowledge, skills, and values to their work, the downside of this is that different approaches can lead to misunderstandings, suspicion and workplace conflicts (Bassett-Jones, 2005). Diverse workplaces with prejudiced individuals embedded within them can result in communication problems (Parrotta, Pozzoli, & Pytlikova, 2014). Low levels of trust between diverse individuals can lead to poorer knowledge transfer and productivity losses (Alesina & La Ferrara, 2002). The end results of these tensions include employee absenteeism, loss of morale, and reduced competitiveness (Bassett-Jones, 2005).

The challenge for leaders, then, is "how do I capture the benefits of diversity without introducing negative team-based outcomes?" The answer is "through inclusive leadership practices." Diversity includes the valuing of an individual for the unique perspective and skill set that they bring to a team. Inclusion refers to the degree to which individuals feel they belong to, and are valued within, a team. Inclusive cultures are those where individuals feel they are respected and psychologically safe, and where fair treatment is evident across all social groups.

Inclusive Leadership for improving diversity management in medical research

Inclusive leadership practices have been proposed as a potential solution to issues of diversity in workplaces, particularly in dealing with gender diversity. Deloitte (2012) found that employees in organisations with high commitment to diversity but low inclusion felt less engaged than those with high inclusion and low commitment to diversity. This suggests that diversity initiatives alone are insufficient for getting the most out of individual staff members with different backgrounds, approaches, and skills. Recommendations from the Deloitte report for improving inclusion encompassed: regular mechanisms for sharing what each team member is working on, periodically reviewing quality and range of work assigned to flexible and non-flexible workers, improved mentoring programs, displaying inclusive behaviours exemplifying organisational values, and finding opportunities for diverse team members to problem-solve together.

The WiSPP Inclusive Leadership pilot program and evaluation

Devine et al. (2017) note the paucity of evidence-based interventions work for increasing gender and racial diversity. As such, this study provides a promising opportunity to collect data on the efficacy of inclusive leadership training on diversity outcomes in a STEMM context. A key feature of the intervention will be to link inclusive leadership with the creation of productive research cultures, as well as increasing the representation of women and other minority employees in senior roles through greater utilisation, development and retention of female scientists.

The research to test the efficacy of the inclusive leadership program included two phases. In the first phase we conducted semi-structured interviews with representative team members from across the five participating institutes (The Peter Doherty Institute for Infection and Immunity, The Florey Institute of Neuroscience and Mental Health, Murdoch Children's Research Institute, Peter MacCallum Cancer Centre, and the Walter and Eliza Hall Institute of Medical Research). In the second phase, information from the phase one interviews was used to develop an inclusive leadership program that accounted for the identified weaknesses in the current workplace culture. This inclusive leadership training program was run across both semesters in 2018, with a lagged cross-over design allowing for testing changes both between groups and across time. Outcomes were self-reported responses to the training program, with variables chosen to reflect the biases and weaknesses present in the current organisational climate of participating teams.

Method

The study was conducted in two phases; a qualitative phase including semi-structured interviews with employees of participating medical research institutes, followed, in the second phase, by the deployment and statistical evaluation of the inclusive leadership training program.

Qualitative Phase

Before designing the training evaluation survey, we undertook a series of semi-structured interviews with a subset of the program participants. The interviews were designed to better understand the culture and needs of the organisations and to tailor the program to employees in medical research institutes. We met with volunteers from five teams, one from each of the institutes involved in the study. Each team provided three to four volunteers, representing different status, experience and roles in the team. For example, most teams volunteered a team leader or manager, an employee with an established track record in the team, such as a postdoc or lab technician, a PhD student and a research assistant.

This set of interviews revealed a number of key areas that could be leveraged for inclusion in the training program. First, feedback was a consistent point of contention for most participants, from the most junior to most senior team members. Second, many team members felt there were limited opportunities for development, partly hampered by the properties of academic work, such as fixed term contracts, an over-reliance on grants, and limited opportunities for new leaders to emerge without others retiring. Third, many team leaders had no formal training in leadership and felt underprepared for the task of managing a team of staff. This was reflected in feedback of junior staff who felt that leaders performed well as academics, but perhaps lacked the motivation or skills to excel at leadership. As a result of the pilot testing, we introduced several key additions to the content of the inclusive leadership training program.

Evaluation Phase

Participants

Each participating organisation selected six teams to participate in the training and evaluation, providing a total of 30 teams. Teams ranged in size from five members (one leader, four subordinates) to 40 members, with more complex leadership structures. In total, 220 staff registered to attend Modules 1 and 2 across the duration of the program. Teams were drawn from across institutes and included academic, clinical, and administrative/support teams. Partner organisations were responsible for nominating teams that differed in function, age of leader, gender composition, and other key traits. Team leaders were invited to agree to their team participating, however there was no requirement that all team members engage with the training. Therefore, participants represented most, but not all, members of nominated teams across the institutes.

The nominated teams were divided into two groups, such that one group of 15 teams undertook training in Semester 1 and the remaining 15 teams completed the training in Semester 2. The training itself was comprised of two modules. There were repeat sessions for each of the modules throughout each semester. Participants only attended training for the semester they were enrolled in the training program. Approximately 300 participants took part in the training modules over the course of the year. Most, but not all, participants undertook both modules within their semester.

Two modules were developed for the purposes of this pilot program. Module 1 was developed around background concepts such as unconscious bias and privilege. Participants reflected on their own understanding of these terms, and were introduced to some of the barriers for minority group members working in medical research. Module 2 focused on skills and practice, with participants engaging in exercises aimed at improving targeted areas within the cultures of the participating institutes e.g., giving and receiving feedback.

Survey completion rates were relatively high but dropped off in consecutive surveys (see Table 1). Due to anonymity concerns, we did not track participant names across time points and were unable to distribute the surveys in a way that only allowed participants who had completed T1 surveys to participate in T2, and for participants who completed T1 and T2 surveys to participate in T3. We therefore had a dataset which included participants who did not complete all three surveys.

Table 1: *Participant numbers by semester allocation and time*

	Semester 1 *n*	**Semester 2 *n***	**Total**
Time 1	90	66	185
Time 2	49	43	92
Time 3	42	42	84

Measures

A total of four outcome variables and seven predictor variables were included in the survey. All measures, unless otherwise indicated, were assessed on a 5-point Likert scale from 1 (completely disagree) to 5 (completely agree).

Predictor measures

Opportunities for development is a 6-item scale which assesses the participant's perceived opportunities to improve their skills, training and expertise to grow their career e.g., 'I have attended training programs or workshops that have fostered my career'. Higher scores indicate more opportunities.

Control over work life is a 4-item scale that assesses the participant's ability to control what type of work they do and how they do it e.g., 'In my job, I have freedom to decide how I work'. Higher scores indicate greater control.

Flexibility comprises 5 items that ask participants about their ability to work flexibly through practices such as varying their hours of employment, changing start/finish times, and perceived impact of working flexibly e.g., ‘My colleagues are supportive of flexible work arrangements’. Higher scores indicate greater flexibility.

Work climate comprises of 10 items focusing on the team environment, including concepts such as trust, communication, and psychological safety e.g., ‘In my team, people openly and freely share relevant information and ideas’. Higher scores indicate a more positive, open work climate.

Work-family conflict includes 5 items that assess the extent to which there is conflict between the participant’s work and family obligations e.g., ‘I go home too tired to do other things I’d like to do after work’. Higher scores indicate a greater degree of conflict between work and family.

Work stress is a 5-item scale which focuses on participant’s stressors and dissatisfaction at work e.g., ‘I have excessive responsibilities in my position’. Higher scores indicate a greater experience of stress from work.

Leader-member exchange is a 10-item scale used to assess the degree to which subordinates feel that their supervisor has a good working relationship with them, including their understanding of the subordinate’s role, work quality, and feedback e.g., ‘I always know how satisfied my supervisor is with what I do’. Higher scores indicate a better working relationship with one’s supervisor.

Outcome measures

Perception of growth was measured using 4 items e.g., ‘I am going to be able to meet my career goals in this organisation’. Higher scores indicate a higher expectation of career growth at the organisation.

Retention was measured using 5 items e.g., 'I plan to stay with this organisation for the rest of my career'. A higher score indicated the participant felt they were more likely to stay at the organisation.

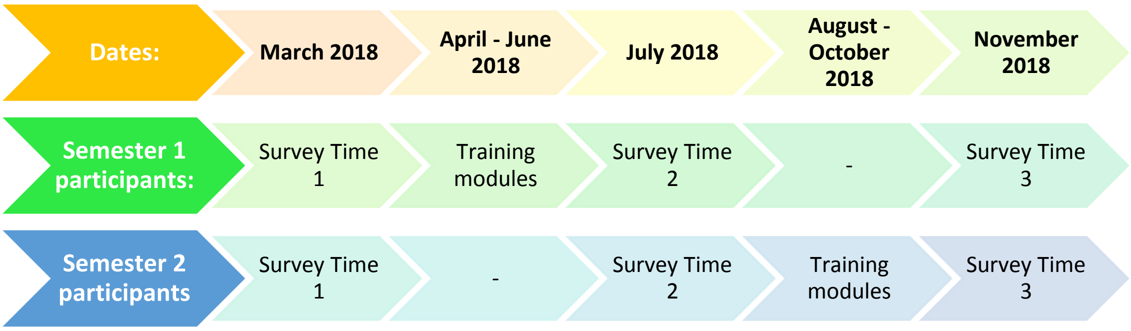
Health and wellbeing was measured using 6 items e.g., 'I often find myself worrying about work issues'. The items were created to reflect poor health and wellbeing; therefore, the scoring of this measure was reversed such that higher scores reflected better health and wellbeing outcomes.

Work attitudes comprised two subscales: job satisfaction (3 items) e.g., 'All things considered, I am satisfied with my job' and organisational commitment (2 items) e.g., 'I enjoy working in this organisation'. Higher scores indicate a more positive attitude towards the workplace. Each of the two subscales did not have sufficient internal validity, so an average score was calculated for all five items. A single job satisfaction item, 'All things considered, I am satisfied with my job', was also used for specific analyses, given that the remaining two job satisfaction items did not show a high level of correlation with this item, which was most representative of the construct.

Design

The content of the inclusive leadership program was developed and delivered by a consulting partner who was not involved in the evaluation design, data collection, analyses and interpretation of the data. Participants attended two two-hour modules, for a total of four hours of training. The first and second modules were held approximately one month apart, however each module had repeat sessions so the time between modules 1 and 2 could vary from three weeks to over two months. The first module focused on leadership, the role of unconscious bias in the workplace, modern challenges to efficacy in medical research, and other topics relating to leadership. The second module explored how leadership worked in a dynamic environment of medical research and how to provide effective feedback within teams. Approximately half the participants (15 teams) undertook the leadership training program in Semester 1 (April – June 2018) while the other 15 teams undertook leadership training in Semester 2 (August – October 2018).

The efficacy of the leadership training modules was tested using a lagged cross-over evaluation model. Participants completed an identical survey at three time points. The timings of the surveys and training for participants who undertook training in Semester 1 and Semester 2 are shown in Figure 1. Semester 1 participants completed one pre-training and two post training surveys. Semester 2 participants completed two pre training and one post training surveys. We therefore had three comparable time points for the two conditions, allowing us to evaluate scores before training (T1), after Semester 1 participants (but not Semester 2 participants; T2) had completed training, and after all participants had completed training (T3). Such a design enabled us to test for changes across time and between groups as a function of receiving training.



Dates:	**March 2018**	**April - June 2018**	**July 2018**	**August - October 2018**	**November 2018**
Semester 1 participants:	Survey Time 1	Training modules	Survey Time 2	-	Survey Time 3
Semester 2 participants	Survey Time 1	-	Survey Time 2	Training modules	Survey Time 3

Figure 1

Experimental design including survey and training time points.

Results

Data consolidation

Due to difficulties with the recruitment and retention of participants across all three time points, we collapsed scores between time points to allow us to only have two scores for each variable: pre-training and post-training. To do this, we collapsed T2 and T3 scores for participants in Semester 1, as both time points were after training was completed. We retained T2 scores, and populated missing scores with T3 scores where available. A similar process was used to collapse T1 and T2 scores for Semester 2 participants, such that T2 scores were prioritised, and missing scores were populated with T1 data where available. As a result, our sample size was boosted (see Table 2).

Table 2. *Participant numbers after collapsing data across time points*

	Pre-training sample	**Post-training sample**
Semester 1	81	67
Semester 2	75	42
Total	156	109

Demographic information

Participants represented a relatively broad cross-section of society. Gender was skewed towards women, with women comprising approximately three quarters of the sample at both pre- and post-training. The majority (around 70-80%) of the sample spoke English as a first language, although only just over half of participants were born in Australia. Around three quarters of the sample identified as heterosexual. The mean age of the sample was mid-thirties.

Table 3. Demographic details of the participants at pre- and post-training intervals

Demographic category	**Pre-training**	**Post-training**
Male	38	32
Female	140	116
Undisclosed gender	6	4
English as a first language	133	112
First language other than English	48	40
Undisclosed first language	3	3
Australian born	99	84
Overseas born	75	63
Undisclosed COB	6	6
Heterosexual	153	112
Non-heterosexual	28	37
Undisclosed sexual orientation	3	4
Mean age (SD)	36.91 (9.72)	36.97 (9.51)

Descriptive statistics

In general, teams reported relatively high scores on positive outcomes. All positive outcomes were above the midpoint of the scale i.e., scores of 2.5 or higher (on a 5-point scale), while negative outcomes were generally below the midpoint. However, negative outcomes were more likely to indicate a problem in the workplace. The three negatively framed outcomes were: (poor) health and wellbeing, work stress, and work family conflict. All these returned scores around the midpoint of the scale, which – when they are reversed such that high scores indicate a positive outcome – would be considerably lower than the other outcomes considered in this pilot test. These variables are therefore identified as areas that participating organisations may want to put additional resources towards.

Table 4. Variable means, standard deviations, and reliability scores at pre-training and post-training.

	Pre-training (*n*=156)			**Post-training (*n*=109)**		
Scale	**M**	**SD**	**α**	**M**	**SD**	**α**
Personal growth	3.28	.60	.72	3.19	.72	.78
Retention	3.36	.84	.87	3.31	.86	.90
Health and wellbeing (reversed)	2.84	.87	.88	2.84	.83	.89
Work attitudes	3.78	.55	.72	3.77	.67	.82
Opportunities for development	3.71	.71	.84	3.86	.71	.87
Flexible work arrangements	3.88	.61	.72	3.92	.63	.83
Control over work life	4.02	.68	.84	4.09	.69	.90
Work climate	3.74	.51	.82	3.76	.62	.88
Work family conflict	2.41	.71	.80	2.37	.74	.77
Work stress	2.48	.64	.70	2.50	.73	.78
Leader-member exchange	3.58	.57	.85	3.57	.65	.86

Group Differences

A number of differences were observed in the pre-training means between the Semester 1 and Semester 2 cohorts, suggesting inadequate randomization of participants. These differences tended to favour Semester 1 participants, with higher average scores for positive predictors and outcomes reported by Semester 1 compared to Semester 2 participants. Personal growth, work attitudes, opportunities for development, control over work, and leader-member exchange scores were all higher in the Semester 1 group compared to Semester 2 (see Table 5), at both the pre- and post-training time points. Flexible work arrangements scores were also higher for Semester 1 participants, but only during the pre-training stage as scores improved for Semester 2 participants in the post-training data collection. The differences observed between groups before undertaking training poses significant difficulties in interpreting subsequent results. In Table 5, we display the results of independent-samples t-tests for the Semester 1 and Semester 2 participants' scores on the pretraining survey, noting those that are significant.

Table 5. Independent samples t-tests comparing pre-training scores for Semester 1 and Semester 2 participants.

Scale	**Sem 1 *M***	**Sem 2 *M***	***t***	***df***	***Sig.***
Personal growth	3.38	3.16	2.32	154	.02*
Retention	3.46	3.24	1.66	154	.10
Health and wellbeing	2.93	2.75	1.29	154	.20
Work attitudes	3.87	3.69	2.00	154	.05*
Opportunities for development	3.91	3.49	3.78	154	<.001*
Flexible work arrangements	3.99	3.76	2.38	154	.02*
Control over work life	4.18	3.85	3.09	154	<.01*
Work climate	3.78	3.70	.99	154	.33
Work family conflict	2.43	2.38	.47	154	.64
Work stress	2.53	2.43	.93	154	.35
Leader-member exchange	3.69	3.45	2.70	154	<.01*

Training Analyses

To determine if the inclusive leadership training experience influenced participant scores on our variables of interest, we ran a series of univariate mixed-model ANOVAs with *Time* (pre-training vs. post training) as a within-person variable and *Condition* (Semester 1 or Semester 2 participation) as a between-subjects variable. The analysis was run twice; once as described, and once with the addition of gender, part time/full time work, and control over work life[1] as covariates.

Personal Growth. There was evidence that participants in Semester 1 and Semester 2 differed in their self-perceived personal growth, *F*(1,80) = 4.54, *p*=.04. This difference was stable over time, *F*(1,80) = 1.05, *p*=.31, and did not interact with the training, *F*(1,80) = 1.56, *p*=.22. There was a marginally significant interaction of gender, *F*(1,80) = 3.85, *p*=.05; men reported higher personal growth scores than women.

Retention. There was no evidence that retention scores differed between participants in Semester 1 and Semester 2, *F*(1,80) = 1.78, *p*=.19. Although scores did not significantly change over time, *F*(1,80) = 2.66, *p*=.11, there was a significant interaction of time by condition, *F*(1,80) = 4.56, *p* = .04. Participants who undertook training in Semester 2 showed a significant increase in their desire to remain at their current organisation; however, there was no significant change for Semester 1 participants. Gender, work status, and control over work life were not significant covariates.

Health and Well-being. We did not find any evidence of differences between the two conditions, *F*(1,80) = .00, *p* = .97. Neither did we find evidence of change over time, *F*(1,80) = .17, *p* = .68, or time by condition, *F*(1,80) = .84, *p* = .36. Gender, work status, and control over work life were not significant covariates.

Job Satisfaction. Job satisfaction was extracted as a single item from the Work Attitudes scale. This decision was made on the basis of low reliability for the Work Attitudes scale, as well as

[1] Control over work life was highly correlated across the sample and so was included as a covariate.

the unique content of the single item relating to job satisfaction. When we used this single-item measure, we found no significant differences based on condition [*F*(1,80) = .07, *p* = .79], time [*F*(1,80) = .23, *p* = 63], nor their interaction [*F*(1,80) = .80, *p* = .38]. There was also no effect of gender [*F*(1,73) = .00, *p* = .99] as a covariate, however control over work [*F*(1,73) = 8.64, *p* = <.01] was significantly related to job satisfaction.

Opportunities for Development. This variable differed significantly between Semester 1 and Semester 2 participants, *F*(1,80) = 7.04, *p* = .01; this difference was similar at pre- and post-training. There was no change over time, *F*(1,80) = .84, *p* = .36, nor was there an interaction between time and condition, *F*(1,80) = .01, *p* =. 93. Gender was not a significant covariate. However, full-time employees reported significantly higher opportunities for development compared to their part-time colleagues, *F*(1,80) = 4.09, *p* = .05.

Flexible Work Arrangements. Those who completed training in Semester 1 reported significantly higher opportunities for flexible work arrangements, discretion over time etc. compared to those who completed training in Semester 2, *F*(1,80) =4.10, *p*=.05). No differences occurred over time, *F*(1,80) = .07, *p* = .80, nor did time interact with condition, *F*(1,80) = .32, *p* =.58. Gender, work status, and control over work life were not significant covariates.

Control Over Work. This variable was, on average, higher for those in the Semester 1 group, *F*(1,80) = 5.84, *p* = .02. However, this variable did not change over time, *F*(1.80) = .03, *p* = .87, nor did time interact with condition, *F*(1,80) = .17, *p* = .68. Men self-reported greater control over their work compared to women, *F*(1,78) = 5.38, *p* = .02. We explored this variable by including flexible work arrangements as a covariate; although these practices are conceptually similar, they are not synonymous. We theorised that participants who had more control over the time they spent at work, their start/finish times, and perceived less negative consequences of engaging in flexible work practices were more likely to also feel they had discretion over decisions in the workplace including structuring and initiating tasks. Flexible work arrangements at pre-training was a significant predictor of control over work, $F(1,77) = 38.76$, $p <.001$. When flexible work arrangements was included in the model, differences between the two conditions were no longer significant, $F(1,77) = 2.73$, $p = .10$.

Work Climate. There was no significant difference between participants in each condition, $F(1,80) = .95$, $p = .33$. We also did not find evidence of change over time, $F(1,80) = 1.89$, $p = .17$, or interactions between time and the semester the participants were trained in, $F(1,80) = .06$, $p = .82$. Gender was not a significant covariate. However, when we included control over work in the model, we found this to be a significant predictor of work climate perceptions, $F(1,77) = 30.48$, $p < .001$.

Work-family conflict. We did not find evidence of a significant difference between the Semester 1 and Semester 2 participant group, $F(1,80) = .63$, $p = .43$. Neither did we see any change over time, $F(1,80) = 1.81$, $p = .18$, or time by condition interactions, $F(1,80) = 1.24$, $p = .27$. Gender was not a significant covariate of this model, $F(1,80) = .01$, $p = .94$; however, we found that participants employed in full-time roles experienced higher levels of work-family conflict than part-timers, $F(1,75) = 10.24$, $p = <.01$. In addition, responses to flexible work arrangements at pre-training was a significant covariate, $F(1,80) = 7.68$, $p = <.01$, suggesting that participants who had access to flexible work arrangements without negative judgements experienced a lower degree of work-family conflicts.

Work Stress. There were no significant differences on the basis of condition, $F(1,80) = .07$, $p = .79$. Similarly, no significant effects emerged for time, $F(1,80) = 1.65$, $p = .20$ or for the interaction of condition and time, $F(1,80) = 2.04$, $p = .16$. Gender was not significant, $F(1,78) = .05$, $p = . 82$, however those with access to flexible work arrangements reported experiencing less work stress, $F(1,75) = 8.39$, $p = <.01$.

Leader-Member Exchange. We found that LMX marginally varied by condition, $F(1,80) = 3.24$, $p = .08$; Semester 1 participants had slightly higher ratings of LMX compared to Semester 2 participants. There was a significant time by condition effect, $F(1,80) = 4.00$, $p = .05$, qualified by

Semester 2 participants reporting lower LMX scores post-training compared to pre-training while Semester 1 participants did not vary in their LMX scores. Gender did not prove to be a significant covariate, *F*(1,80) = 1.07, *p* = .31.

Participant Feedback

Finally, participants also provided qualitative feedback about the program. The program was generally well received by both groups of participants. When asked if they would recommend the program to others, Semester 1 participants returned a score of 7.61 on a 10-point scale (1 = not at all, 10 = absolutely), and Semester 2 participants returned a slightly higher recommendation score of 7.68. Participants also reported on their perceptions of the most successful and unsuccessful components of the program. Three core themes emerged from the success comments. First, participants appreciated the opportunity to learn about the ways that bias, privilege, and power play out in their workplaces, and felt the opportunity to reflect and raise awareness was important. Second, participants enjoyed the interactive components of the training, which allowed them to take the content they had been taught and put it into practice. Third, participants felt they benefited from having a structured feedback exercise that provided a framework for giving and receiving effective feedback in their teams.

Opportunities for improvement of the training program were more varied, and there was less consistency in participant feedback. Some participants felt that the length of training sessions needed to be extended so that content could be explored more fully. Others noted that the two modules were relatively independent, and wanted more explicit connections made between the introduction of concepts in Module 1 and putting these concepts into practice in Module 2. Participants also commented on the gender composition of participants and felt that the sessions may have been ‘preaching to the choir’, given that teams self-selected into the program and were therefore most likely to already agree with the principles and values put forth.

Participants who attended Semester 1 training were asked additional questions about which practices they had put into place, having had several months between attending the training modules and completing the third survey. Although there were relatively few responses for this section, those who did respond reported more engagement with concepts around inclusivity and respecting differences, demonstrated through practices such as improving the format of group discussions, delegating important growth opportunities such as external stakeholder engagement, and proactively seeking feedback from team members.

Discussion

The results of the analyses showed that the cultures within the participating medical research institutes were evaluated positively by the study participants. Scores were generally above the midpoint of the scale for positive attitudinal measures. However, the results for the health and wellbeing, work stress, and work-family conflict measures indicate room for improvement. Group composition differences between the Semester 1 and Semester 2 participants presented difficulties in the interpretation of the data, however we found some significant findings especially with regard to women working in STEMM. We also had a significant drop-out rate between sampling points, which resulted in us collapsing data across time points, potentially obscuring more nuanced changes across time.

The discovering of several pre-existing differences between participants in Semester 1 and Semester 2, meant that the two groups of teams were not equivalent for comparison purposes. Therefore, any changes at the post-training interval cannot be entirely attributed to the training program, and could potentially be caused by higher levels of positive attributes in the Semester 1 participants.

This pre-training differences of the two groups may have been caused any of several factors. First, participants were recruited by participating institutes with an aim to represent a spectrum of team differences, including in leader seniority, and team cultures. Recruitment was conducted by representatives from each of the institutes with identical instructions. However, recruiting six teams proved problematic in some cases institutes. Some leaders who were approached for the training did not agree to participate, whereas other teams were eager to engage. Finding the last few teams required stretching the guidelines for our definitions of a team; for example, rather than taking an entire lab group, we included participants who represented a subsample of a lab group under the leadership of a senior postdoctoral researcher. Due to difficulties in finding the last few groups to participate, the Semester 1 participants began undertaking the first survey before we had completed the recruitment process.

This selection process may have resulted in some self-selection bias. Teams eager to participate were likely to be those with better leadership cultures and fewer existing cultural issues. In contrast, teams which were reluctant to participate may have been those that did not want their internal culture examined or to be challenged by a training program. This may also have meant that teams that were signed up earlier – therefore being more likely to be placed in the Semester 1 group – were those with healthier existing cultures, represented by Semester 1 participants being more likely to have higher scores on positive indices of team culture. We also noted that scores for Semester 1 participants tended to be stable across time, whereas Semester 2 participant scores were more likely to change; typically, this change was a reduction, suggesting less positive attitudes towards team culture variables following training. Rather than indicating a failure of training, this may instead reflect a growing awareness among Semester 2 participants of existing problems within their team culture that they were not aware of before the training.

Second, the high dropout rates between times points may have affected the efficacy of the evaluation program. At Time 3, less than half of the Time 1 participants completed the survey. The sample size was too small to effectively find small or medium effect sizes, so we collapsed data such that we did not have a cross-over design, but instead only had a pre-and-post within-participant design. Nonetheless, the drop out rate causes problems with accounting for patterns of variance amongst participants that dropped out of the survey process. Given concerns about anonymity, we were also not able to collect data on which institute each participant worked for, so we were unable to determine if there was a pattern of survey discontinuation that was correlated with specific workplaces. Potentially, those that continued with the survey through all time points were those who were most committed to the principles and content of the training, which may have led to generally high, consistent scores for team culture variables.

Despite the limitations due to problems in the implementation of the research design, there are a number of key recommendations that can be taken from the evaluation of the inclusive leadership training program. First, although many of the outcomes tested in this study returned relatively high scores on average, variables that were negatively phrased returned results that could be improved upon. These include health and wellbeing (high scores indicate more health complaints), work stress, and work-family conflict. These variables represent the consequences of poorly designed workplaces where staff are overloaded, feel conflicted in balancing their home and work lives, and experiencing negative health outcomes as a result. Moving forward, the participating institutes may wish to consider the impact of these outcomes on their staff and ways to mitigate these effects.

Similarly, aspects of the results showed that participants with specific support structures fared better in their workplace attitudes and outcomes. In particular, those that had greater control over their work life and flexible work arrangements had higher scores on several work attitudes . Control over work life – being able to make decisions about how, when, and where their work will be done – increased job satisfaction and perceptions of work climate, whereas flexible work arrangements – having access to flexibility in organising one’s time without prejudice from colleagues – was related to feeling more control over work, reduced work-family conflict, and reduced work stress. Given the particularly low scores for the latter two constructs, flexible work arrangements may be of particular importance in improving diversity and inclusion culture across the institutes.

Finally, some additional differences were detected between participants of different demographics. Men reported greater feelings of personal growth and control over work, suggesting that women may feel more constrained in the work they do and in their ability to continue growing their career in their current organisation. Similarly, participants in part-time working arrangements reported fewer opportunities for development and more work-family conflict. Part-time workers may experience being marginalized or given non-core business to work on, rather than feeling they are given opportunities to grow their careers and continue climbing the seniority ladder. Those with children may feel that they are not nominated as often for opportunities that involve travel due to stereotypical beliefs about travelling with (or away from) small children. Those with caring responsibilities are more likely to choose a part-time working arrangement, resulting in greater feelings of work-life conflict. The institutes may wish to consider the impact that part time work has on the opportunities staff have to continue flourishing in their careers, rather than feeling sidelined by their leaders.

In particular, given the intersection between women and part-time workers reporting less development and career opportunities, the institutes may be in danger of 'mommy tracking' their female staff by reducing the quality and importance of the work that they engage in while engaging in carer roles outside of work. Similar findings have been observed for other career-driven women (Ely, Stone, & Ammerman, 2014), who – despite having similar levels of career ambitions to their male counterparts find themselves in unsatisfying, dead-end roles after taking maternity leave. Take heed, though: career-driven women who are mommy-tracked will not stay in the role they have been allocated to. The vast majority will leave to take up another role elsewhere where their contributions will be valued. Rather than presume women and/or part time workers would prefer less challenging work, leaders should be engaging in conversations with their team to ensure that each team member has access to the development and growth opportunities that align with their long-term career goals.

CONCLUSION

The current study reflects many of the difficulties that can arise in the conduct of quasi experimental field studies. These include selection effects in the composition of the comparison groups due to lack of randomisation and attrition of participants across the different waves of pre and post measures. The resulting limitations in the study design limit the confidence in the inferences that can be drawn regarding the impacts of the training results. That said, while the study does not support causal inferences, the data and the results do identify several significant differences between groups that are of interest and provide a set of conclusions that can be further tested in discussions within medical research teams. In particular, teams should explore opportunities for increasing individual staff control over the decisions as to how, when and where their work will be done and providing flexible work arrangements, without negative reactions from peers or the forfeiting of development opportunities.

References

Alesina, A., & La Ferrara, E. (2002). Who trusts others?. *Journal of Public Economics*, *85*(2), 207-234.

Arora, A., Mittal, A., & Pasari, R. (2011). What makes a good researcher. *Social and Information Network Analysis*. Retrieved from https://www.researchgate.net

Bailyn, L. (2003). Academic careers and gender equity: Lessons learned from MIT. *Gender, Work and Organizations,* 10(2) 137-153.

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

Bassett-Jones, N. (2005). The paradox of diversity management, creativity and innovation. *Creativity and Innovation Management*, *14*(2), 169-175.

Blake-Beard, S., Bayne, M. L., Crosby, F. J., & Muller, C. B. (2011). Matching by race and gender in mentoring relationships: Keeping our eyes on the prize. *Journal of Social Issues*, *67*(3), 622-643.

Bozeman, B., & Boardman, C. (2014). Assessing research collaboration studies: A framework for analysis. In *Research collaboration and team science* (pp. 1-11). New York: Springer, Cham.

Carrigan, C., Quinn, K., & Riskin, E. A. (2011). The gendered division of labor among STEM faculty and the effects of critical mass. *Journal of Diversity in Higher Education*, *4*(3), 131-146.

Cox, T. (2001). *Creating the multicultural organization: A strategy for capturing the power of diversity.* San Francisco, CA: Jossey-Bass.

Deloitte. (2012). *Waiter, is that diversity in my soup? A new recipe to improve business performance.* Retrieved from https://www.humanrightscommission.vic.gov.au

Devine, P. G., Forscher, P. S., Cox, W. T., Kaatz, A., Sheridan, J., & Carnes, M. (2017). A gender bias habit-breaking intervention led to increased hiring of female faculty in STEMM departments. *Journal of Experimental Social Psychology*, *73*, 211-215.

Ely, R. J., Stone, P., & Ammerman, C. (2014). Rethink what you “know” about high-achieving women. *Harvard Business Review*, *92*(12), 100-109.

Greene, J., Stockard, J., Lewis, P., & Richmond, G. (2010). Is the academic climate chilly? The views of women academic chemists. *Journal of Chemical Education*, *87*(4), 381-385.

Gurin, P., Nagda, B. R. A., & Lopez, G. E. (2004). The benefits of diversity in education for democratic citizenship. *Journal of social issues*, *60*(1), 17-34.

Harrison, D. A., Price, K. H., & Bell, M. P. (1998). Beyond relational demography: Time and the effects of surface-and deep-level diversity on work group cohesion. *Academy of management journal*, *41*(1), 96-107.

Herring, C. (2009). Does diversity pay?: Race, gender, and the business case for diversity. *American Sociological Review*, *74*(2), 208-224.

Hewlett, S. (2002). *Baby Hunger: The New Battle for Motherhood*. London, UK: Atlantic Books.

Hewlett, S. A., Marshall, M., & Sherbin, L. (2013). How diversity can drive innovation. *Harvard Business Review*, *91*(12), 30.

Hong, L., & Page, S. E. (2001). Problem solving by heterogeneous agents. *Journal of Economic Theory*, *97*(1), 123-163.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385-16389.

Janssens, M., & Zanoni, P. (2008, November). What makes an organization inclusive? Organizational practices favoring the relational inclusion of ethnic minorities in operative jobs. In *Organizational Practices Favoring the Relational Inclusion of Ethnic Minorities in Operative Jobs (November 9, 2008). IACM 21st Annual Conference Paper*.

Joy, L., Carter, N. M., Wagner, H. M., & Narayanan, S. (2007). The bottom line: Corporate performance and women's representation on boards. *Catalyst*, *3*(1). Retrieved from http:// www.catalyst.org/knowledge/bottom-line-corporate-performance-and-womens-representation-boards

Konrad, A. M. (2003). Special issue introduction: Defining the domain of workplace diversity scholarship. *Group & Organization Management*, *28*(1), 4-17.

Lucht, P. (2014). *De-Gendering STEM - Lessons Learned from an Ethnographic Case Study of a Physics Laboratory.* Conference paper presented at the 2nd Network Gender and STEM conference, 3-5 July, Berlin, Germany.

Mak, W. W., Law, R. W., Alvidrez, J., & Pérez-Stable, E. J. (2007). Gender and ethnic diversity in NIMH-funded clinical trials: Review of a decade of published research. *Administration and Policy in Mental Health and Mental Health Services Research*, *34*(6), 497-503.

Nishii, L. H. (2013). The benefits of climate for inclusion for gender-diverse groups. *Academy of Management Journal*, *56*(6), 1754-1774.

O'Leary, J., & Tilly, J. (2013). Cracking the cultural ceiling: Future proofing your business in the Asian century.

Parrotta, P., Pozzoli, D., & Pytlikova, M. (2014). The nexus between labor diversity and firm's innovation. *Journal of Population Economics*, *27*(2), 303-364.

Roberge, M. É., & Van Dick, R. (2010). Recognizing the benefits of diversity: When and how does diversity increase group performance?. *Human Resource Management Review*, *20*(4), 295-308.

Roberson, Q. M. (2006). Disentangling the meanings of diversity and inclusion in organizations. *Group & Organization Management*, *31*(2), 212-236.

Rothstein, M. G., & Davey, L. M. (1995). Gender differences in network relationships in academia. *Women in Management Review*, *10*(6), 20-25.

Science in Australia Gender Equity. (2016). Gender equity in STEMM. Retrieved from https://www.sciencegenderequity.org.au/gender-equity-in-stem/

Shore, L. M., Randel, A. E., Chung, B. G., Dean, M. A., Holcombe Ehrhart, K., & Singh, G. (2011). Inclusion and diversity in work groups: A review and model for future research. *Journal of Management, 37*(4), 1262–1289.

Shin, S. J., Kim, T. Y., Lee, J. Y., & Bian, L. (2012). Cognitive team diversity and individual team member creativity: A cross-level interaction. *Academy of Management Journal*, *55*(1), 197-212.

Storage, D., Horne, Z., Cimpian, A., & Leslie, S. J. (2016). The frequency of “brilliant” and “genius” in teaching evaluations predicts the representation of women and African Americans across fields. *PloS one*, *11*(3), e0150194.

Williams, J. 2001. *Unbending Gender: Why Family and Work Conflict and What to Do about It.* Oxford: Oxford University Press.