Workplace Project Portfolio
Masters in Biostatistics

Evaluating the success of a workplace health and wellbeing intervention using a small group of repeat-respondents from a large repeated cross-sectional survey

Kate Chappell
October 2015

Contents

Preface

My role

I completed this project while working part-time for the *partnering* Healthy@Work (*p*H@W) research group at the Menzies Institute for Medical Research, University of Tasmania, between August 2014 and October 2015. The *p*H@W research group were responsible for the evaluation of a Tasmanian State Service (TSS) health and wellbeing initiative-the Healthy@Work (H@W) intervention. The evaluation was conducted using data collected from two large cross-sectional surveys. My role in the group was to manage the survey data (for the final two years of the five-year evaluation project), and to support three PhD students in accessing, cleaning and analysing the data. An opportunity to undertake this individual research project arose when the group decided that it would be useful for someone to analyse data from a "cohort" of repeat-respondents to the two main cross-sectional surveys, as a sensitivity analysis for the outcomes of the broader cross-sectional analyses. The research questions for this project were:

(1) Did the health of employees in the TSS change over the three year implementation period of the H@W intervention, and

(2) Was there evidence of increased engagement in workplace health and wellbeing as a result of the intervention?

The project was initiated by my supervisor at Menzies, Professor Alison Venn, and supervised by Menzies biostatistician Associate Professor Leigh Blizzard. I was entirely responsible for the project, which involved extracting relevant data from the two survey datasets, cleaning the data, creating new variables to adequately define exposures and outcomes, analysing the data and reporting the results. Assoc. Prof. Leigh Blizzard, my statistical supervisor, assisted me with various aspects of the analysis and interpretation of results and revision of the final document. The results of this analysis have been presented as a scientific manuscript, with the intention of submitting the document to a peer-reviewed journal in the near future. A statistical appendix gives further details of the statistical methods used.

Reflections on learning

Communication

Undertaking a project within a research team enabled me to develop skills in communicating statistical concepts and explaining outcomes without using overly technical statistical terminology - both informally within our research team and more broadly when presenting my results to a wider group of Menzies researchers and project partners. Organising meetings with my supervisor and communicating clearly the difficulties I was experiencing also required some development of communication skills, which improved as I became more familiar with the modelling methods I was using.

Work patterns/planning

I was given time at work to undertake this project, which meant that I had 1-2 days per week over several months to complete it. It was important to set my own timelines for completion of various intermediate tasks, as I was working independently, without any external pressure to finish the project. At times important workplace tasks interrupted my work, or I was unable to access my supervisor for help with analyses, which slowed my progress. Forward planning and flexibility were required to arrange meeting times, and sometimes I headed off in a less-than-useful direction with my analyses while awaiting advice. However, as I had intended this to be the first of two projects, meaning that it was not required to be submitted after only one semester, I felt that I had some flexibility in my timeline for completing the project. As a result, I took some extra time working on specific aspects of the analysis. For example I decided to re-analyse the data using log multinomial regression, when I realised late in 2014 that my count data model was not a good fit to the data. This exploration of modelling techniques was beneficial for my overall statistical understanding, but did delay submission of the project.

Statistical principles and methods

The research questions and scope of the project evolved somewhat from beginning to end. My initial brief was rather vague, as in "*see what the cohort data shows…*" , but without clear research questions it was not possible to sensibly present an analysis of the data with a meaningful statistical interpretation. There was also the temptation to indulge in data-dredging, which we (in the Biostatistics Collaboration of Australia (BCA) statistics courses) have been taught to avoid! Although the project could perhaps have been completed earlier if I had had clear research questions early on, the process of developing research questions and using relevant statistical methods to answer them was a good learning experience. Given the breadth of data that had been collected in the *p*H@W surveys and the differing research interests in the group, some initial vagueness in the scope of my research project was perhaps inevitable.

I reviewed the literature and investigated the differences in statistical techniques for analysing cohort versus cross-sectional data. This study was not a typical cohort study in which a group was recruited and followed through time, but rather an "accidental cohort" study using the repeated measures available for the group who were surveyed twice purely by chance and who chose to respond both times. The low response percentage (15%) raised the issue of selection bias, or ‘response bias’. I needed to address this issue if my results were to be interpreted as applicable to all employees in the Tasmanian State Service. I considered the two main techniques described in the literature for managing missing data: Multiple Imputation and Inverse Probability Weighting. I was not previously familiar with either of these techniques. Other members of the research group had used Inverse Probability Weighting for non-response, and my investigations led me to decide that this was the most sensible solution for my study also.

I spent a large amount of time categorising and scaling the variables used to indicate availability of and participation in various aspects of the health and wellbeing intervention. Because of differences in the survey questions between the two surveys, which were taken three years apart, it was difficult to create consistent measures of availability and participation that could be compared between the two time points. I relied on advice from my supervisor and other group members and some common sense when developing these measures, but I did feel that the evaluation of the intervention was limited by some methodological issues to do with the survey development, which pre-existed my employment on the project and which I had no choice but to work with. Design of health indicators and surveys is obviously an important topic, and unfortunately I have not had time to study the existing BCA subject - I think it would be very useful!

Because I was analysing change over time using repeated measures on individuals, it was important to be aware of statistical issues relating to correlated and longitudinal data. Paired t-tests and methods for analysing the difference in correlated proportions were used to investigate changes in continuous and categorical health measures for individuals. These methods had been touched on briefly in several BCA subjects, but I learned more through my own research for this project. The BCA subject Longitudinal and Correlated Data, which I undertook as I was finishing this project, added much more to my understanding of longitudinal and correlated data.

There were several options for analysing the association between exposure to the intervention and the participation outcome. The outcome could be considered as a count, binary, ordinal or multinomial variable, and so I needed to consider which of the various models available would provide the clearest and most accurate results. This decision required a consideration of whether or not it would be beneficial to categorise the participation outcome variable. I used techniques learned in Linear Models (LMR) and Categorical Data Analysis (CDA) for this part of the project, but the log multinomial model, developed by my supervisor and colleagues, was an extension to my prior learning in the BCA subjects. Model diagnostics are important to check the fit of regression models, and I needed to learn some new goodness-of-fit techniques to assess the log multinomial model. Other issues I considered in regard to regression analyses were choosing which covariates to include in the model, and assessing interactions. I developed solutions to these issues through frequent reference to the relevant literature and my LMR and CDA course notes.

In making all of the above decisions I relied heavily on the course notes I had gathered during the past 3 years, while studying the Master of Biostatistics coursework units. Since this Masters course is my first foray into public health research, I had no other prior knowledge to work with. During this independent project I think I used principles and methods from every BCA unit that I had studied, except for Survival Analysis.

Statistical computing

I used Stata version 12.1 for all of my analyses. Stata is the statistical software currently preferred by many of the staff at Menzies, as well as the software most commonly used in the BCA units I have been studying, so this was a good match. I became much more proficient in using Stata during the course of my Workplace Project Portfolio, as I cleaned and manipulated data, ran regression models and created graphs and tables. In particular, I became proficient at using Stata's help documents and the online support provided by the Institute for Digital Research and Education, University of California, Los Angeles.

Teamwork

Communicating with other team members

This was an independent workplace project which was undertaken during my employment as a Project Officer in the *p*H@W research group. Within the group there was ongoing discussion and collaboration in terms of data cleaning needs, exposure and outcome variable development and comparison of results from different analyses. It was a supportive environment within which to undertake this type of project, and communication with other team members was never difficult.

Negotiating roles and responsibilities

As far as this project was concerned, there was no negotiation needed, as it was my responsibility alone. My statistical supervisor assisted me with some of the analysis and interpretation of results, which utilized Stata code and a modelling technique that he had developed. He was happy to assist, but because of his busy schedule I often had to wait for several weeks to get advice.

Working within timelines

The project took longer than I had initially expected, in part because I allowed myself extra time to investigate several different modelling options for my main analysis, and partly because of my lack of experience in developing papers for submission to journals- there is a lot of formatting and attention to detail to take care of. The process of writing and revision of the manuscript was slowed during the final months by my supervisor's other commitments and unexpected health issues. I also experienced delays in accessing my supervisor for comment and sign-off at completion of the project. These delays were frustrating but entirely outside my control.

Helping others to understand statistical issues-teaching

There was only a small amount of teaching that related specifically to my project, and this was mostly helping other team members with Stata coding and some simple statistical techniques.

Ethical considerations

NHMRC ethics guidelines & confidentiality issues

I read the National Health and Medical Research Council (NHMRC) ethics guidelines- they are mostly not relevant here as the data I used was de-identified. I didn't create any tables with small cell numbers that could be used to identify individuals. In creating the flowchart I had to access some of the original files, which contained name and address data. These files were only accessed at Menzies on my desktop computer, and none of the personal information was saved anywhere other than on the original locked, password protected disk. All analyses were undertaken at work within the password-protected secure server.

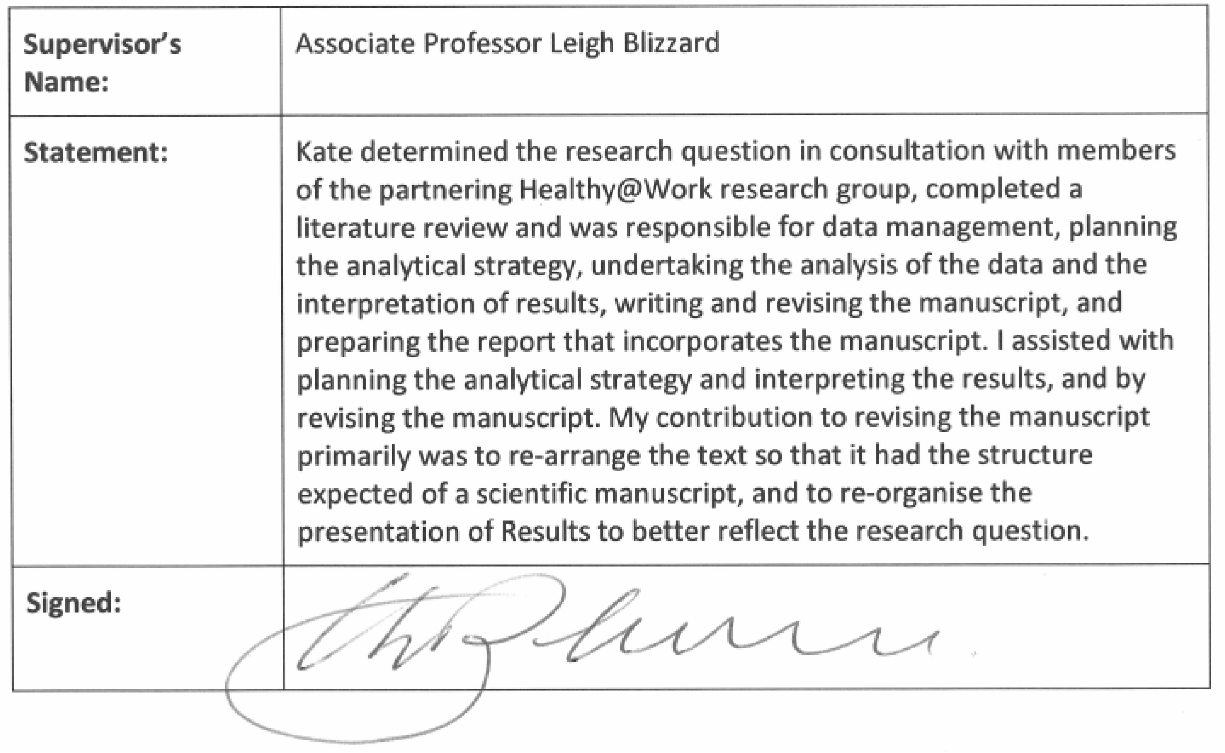
Professional responsibility

It was important to represent the *p*H@W team and Menzies Institute for Medical Research as a professional, and to present the results of the project accurately and with an understanding of the broad context within which the research was conducted. I hope I have succeeded in doing this.

Evaluating the success of a workplace health and wellbeing intervention using a small group of repeat-respondents from a large repeated cross-sectional survey

Location and Dates:	Menzies Institute for Medical Research, University of Tasmania, August 2014-August 2015
Context:	This project was undertaken during my part-time employment as a Project Officer with the research group 'partnering Healthy@Work' (*p*H@W). *P*H@W was formed in 2009 by researchers at the Menzies Institute for Medical Research, University of Tasmania, with the support of a 5-year NHMRC Partnership Project grant, to evaluate a large public-sector workplace health and wellbeing intervention in the Tasmanian State Service (TSS) - the Healthy@Work project. The *P*H@W evaluation included two surveys of TSS employees, in 2010 and 2013, that were conducted using a repeated cross-sectional design. The surveys included 580 participants who responded to both surveys. My project was intended to add to the research outcomes of the group. In particular, by analysing data collected from the cohort of repeat-respondents, I had the responsibility of investigating change within individuals over time.
Contribution of student:	I completed all data cleaning and manipulation, undertook a review of the literature, analysed the data and interpreted the results, prepared the draft manuscript and revised its content. The log-multinomial modelling method was developed by (and utilized Stata code written by) Assoc. Prof. Leigh Blizzard and colleagues. Assoc. Prof. Leigh Blizzard contributed to the interpretation of the data and revision of the manuscript.
Statistical issues involved:	• Cross-sectional versus cohort study designs • Selection and representation (classification/scaling) of appropriate outcome and exposure factors • Inverse probability weighting for non-response • Choice of models for multi-variable analysis: o The Poisson model as an option for a binary outcome with robust standard errors to account for the misspecification of the error distribution

	o Modelling count data: investigating the fit of Poisson and Negative Binomial models o Log multinomial modelling
Declaration:	I declare that I have undertaken this work independently, with direction and assistance provided by my project supervisor. I have not submitted this work for previous academic credit.
Signed:	



Supervisor's Name:	Associate Professor Leigh Blizzard
Statement:	Kate determined the research question in consultation with members of the partnering Healthy@Work research group, completed a literature review and was responsible for data management, planning the analytical strategy, undertaking the analysis of the data and the interpretation of results, writing and revising the manuscript, and preparing the report that incorporates the manuscript. I assisted with planning the analytical strategy and interpreting the results, and by revising the manuscript. My contribution to revising the manuscript primarily was to re-arrange the text so that it had the structure expected of a scientific manuscript, and to re-organise the presentation of Results to better reflect the research question.
Signed:	

Introduction

Chronic diseases such as heart disease, cancer and stroke are the leading causes of death in Western countries (1), including Australia (2). Non-fatal chronic conditions such as psychological disorders and musculoskeletal disorders are major sources of disability (1). Health-related factors that have been identified as contributing to chronic disease include lifestyle factors such as smoking, harmful alcohol intake, physical inactivity and poor diet, and their pathophysiological outcomes including obesity, hypertension, elevated glucose and high cholesterol (1, 3, 4).

There has been an increasing awareness in recent years of both the costs to workplaces of chronic disease among employees, and the influence that workplace conditions can have on employee health and wellbeing (3, 4). Workplaces have therefore been identified as key venues to address multiple individual health risk factors in the large segment of the population that is employed (1). As a result there has been an increase in workplace health promotion interventions through the implementation of health and wellbeing programs, particularly in the United States but also in Australia (1, 3).

The health and wellbeing intervention "Healthy@Work" (H@W) was implemented during the period 2009 to 2012 among the public sector workforce in Tasmania, Australia. The goals of Healthy@Work were to support the development of effective, integrated health and wellbeing programs within each of the 15 Tasmanian State Service (TSS) agencies, with the ultimate aim of improving the health and wellbeing of employees (5). "*Partnering* Healthy@Work" (*p*H@W), a partnership of government and university researchers based at the Menzies Research Institute Tasmania (now renamed as the Menzies Institute for Medical Research), was formed to evaluate the intervention. Because the intervention was implemented across the whole TSS, there was no internal control population with which to compare outcomes. The use of an external control population (such as the state service of another state or territory of Australia) would have required more resources than were available, and was considered to be infeasible and beyond the scope of the project. A full-scale cohort follow-up, using participants as their own controls in before and after comparisons, was not able to be contemplated for privacy reasons. Therefore, a repeated cross-sectional survey design was used for the evaluation. Two cross-sectional surveys of employees in the TSS were taken 3 years apart: one near the beginning and one at the end of the intervention implementation period. The surveys assessed employee health profiles and engagement with workplace health and wellbeing programs. Around 44% of the (approximately) 27,000 TSS employees were surveyed each time, using stratified random sampling, with a response percentage of around 28%.

Purely by chance, approximately 14% of TSS employees were invited to participate in both *p*H@W surveys. The group of employees who were surveyed twice and who responded to both surveys are defined as “the cohort” in this report. The aim was to undertake an evaluation of the success of the H@W intervention using the repeated survey responses of the cohort. In particular, the repeated measures on individuals enabled us to identify changes in availability of health and wellbeing activities and participation in these activities by individual employees, and to investigate associated changes in health-related factors over the implementation period of Healthy@Work.

Methods

Subjects

There were approximately 27,000 employees working across the 15 agencies of the TSS in 2010 and 2013. Stratified random sampling was used to select a sample of employees to participate in each of the two *P*H@W surveys. Stratification was by employment category (permanent versus fixed term/casual), employment condition (fulltime versus part-time) and agency. For each survey, a random sample of 44% of employees in each stratum was selected. In total, approximately 12,000 employees were selected for each survey. Paper questionnaires were sent to selected employees at their work addresses. There were 3410 respondents in 2010 (28% response) and 3228 in 2013 (27% response). Purely by chance, 3844 people were sent questionnaires in both 2010 and 2013. Of these, 580 (15%) were repeat-respondents, that is, they were surveyed twice, and they responded twice. This is the group we call the “cohort”. A further 539 (14%) responded in 2010 but not in 2013, and 495 (13%) responded in 2013 but not in 2010. A total of 2230 employees who received both surveys (58%) did not respond at either time point. The selection and recruitment process is summarised in the flowchart (Figure 1).

Measurements

Availability of and participation in health and wellbeing activities

Availability of and participation in activities were self-reported. TSS employees were given a list of several different types of activities, and asked to indicate whether each type had been available to them at their workplace and, if available, how many times they had participated. In the 2010 survey, employees were asked to indicate what activities were available and how many times they had participated during the last year. In 2013, they were asked what activities were available and how many times they had participated during the last three years. The lists of activities differed in the two surveys due to the grouping together of several activities for the second survey, and the removal of some activities and the addition of others. For consistency, we reduced the activities to 5 main “activity types” that could be matched consistently between the two surveys: physical activities, mental health activities, injury prevention/rehabilitation, health assessments and health education.

Availability of amenities

Amenities were structural or organisational supports that facilitated participation in health and wellbeing activities. They included bike racks, showers and change facilities, accessible stairwells, fruit bowls, healthy vending machines and drinking water. Employees were asked to indicate from a fixed list of possible amenities those that were available at their workplace. Ten of these amenities were listed in both surveys, and these are the amenities used in the analysis of change in amenities between surveys. A further four amenities were added in 2013, and so there were 14 possible amenities in the analysis of the association between the number of amenities in 2013 and participation in 2013. An “other” option was available for employees to list amenities that did not fit into the categories provided.

Health-related factors

Body-mass index (BMI), daily serves of fruit and vegetables, leisure-time physical activity, hours spent sitting at work, psychological distress, alcohol intake and smoking status are the ‘health-related factors’ used in this study. The World Health Organisation defines a ‘Health Risk Factor’ (HRF) as any attribute, characteristic or exposure of an individual that increases the likelihood of developing a disease or injury (6). The HRFs in this study were therefore the levels of each health-related factor that could be considered likely to increase the risk of ill-health: being overweight or obese; current daily smoking; inadequate physical activity; risky alcohol consumption; low fruit and vegetable intake; prolonged sitting at work; and high or very high psychological distress. These are defined below:

Overweight/obese: employees reported their height and weight. These were used to calculate BMI as weight in kilograms divided by height in metres squared (kg/m^2). BMI was used to group respondents into 4 categories of underweight (BMI less than 18.5), normal weight (BMI of 18.5 to 24.9), overweight (BMI of 25 to 29.9) and obese (BMI of 30 or more). For analysis, the categories were collapsed into a binary classification of underweight/normal versus overweight/obese.

Smoking: information on lifetime smoking was collected. Responses were categorised as current daily smoker versus not (combining ex-smokers and never-smokers).

Physical activity and prolonged sitting at work: physical activity and time spent in sedentary behaviours were measured using the long version of the International Physical Activity Questionnaire (IPAQ-Long) (7). Physical activity was categorised as high risk if respondents reported less than the 150 mins/week of moderate-vigorous leisure-time physical activity recommended for health benefit (8). No guidelines currently exist that quantify a maximum 'safe' level of engagement in sedentary behaviours by adults. In the absence of any established recommendations, participants were classified as high risk in this study if they reported sitting for six or more hours per day, on average (9).

Alcohol: alcohol intake was assessed using the three-item AUDIT-C (Alcohol Use Disorders Identification Test) (10). The AUDIT-C measures the frequency of alcohol intake, the typical quantity consumed on a day when drinking, and instances where five or more standard drinks are consumed on one occasion. Each of the three AUDIT-C items has 5 response options scored from 0 to 4, with a total score between 0 and 12. "Risky drinking" was defined according to Royal Australian College of General Practitioners guidelines as a score of 4 or more for men, and a score of 3 or more for women. Those scores are considered to indicate hazardous drinking or active alcohol use disorders (unless the points are all from question 1, which measures frequency of drinking but not quantity consumed)(11).

Diet: low fruit intake was defined as eating one serve or less of fruit per day. Low vegetable intake was defined as eating four serves or less of vegetables per day. These definitions were based on the National Health and Medical Research Council's Australian Dietary Guidelines (12). The health risk factor "low fruit or vegetable intake" was defined as eating insufficient fruit and/or insufficient vegetables.

Psychological distress: psychological distress was measured using the Kessler Psychological Distress Scale (K10) (13, 14), which uses 10 items to assess the level and severity of distress. Questions relate to anxiety and depression symptoms experienced during the previous four weeks. Five-level responses range from "None of the time" (score of 1) to "All of the time" (score of 5), and total scores range from 10 to 50. Higher scores indicate higher levels of psychological distress. Psychological distress total scores were dichotomised as low (K10 total score 10-21) or high (22-50) (15).

All information on health-related factors was self-reported by respondents.

Statistical Analyses

Because only 15% (580/3844) of those selected to participate in both surveys responded both times, there was a concern that the results of this study could be biased if the association between study exposures and outcomes (number of activity types available in association with participation in activities, and participation in activities in association with health risk factors) differed between those who responded to both of the surveys and those who declined to respond to either or both surveys. To address the likelihood of non-response bias, non-missing data were weighted in all analyses, using the inverse of the estimated probability of response for each respondent (16). The probability of response was estimated using logistic regression on the dataset of all employees surveyed twice, with response as the binary outcome variable and age, sex, service length, employment category (permanent/fixed term), employment condition (fulltime/part-time) and agency as the covariates. The selected covariates were the stratification factors (employment category, employment condition and agency) and other variables (age, sex and service duration) on which data was available for all subjects. Inverse probability weighting allows inferences to be drawn for the initially sampled population, which for this study is the Tasmanian State Service.

The descriptive statistics reported are means and standard deviations of continuous variables, and percentages and relative frequencies of categorical variables. Paired t-tests were used to assess the mean within-person differences in continuous factors between the two surveys, and the difference in correlated proportions of categorical factors was assessed using the standard error given by Fleiss, Levin and Paik (17) and with a continuity correction applied in the calculation of 95% confidence intervals. Percentages and relative frequencies of respondents moving from the lower to the higher risk category of health-related factors, and of respondents moving from the higher to the lower risk category in each case, are reported. The longitudinal contribution of aging to the change in health-related factors between the surveys, which were separated in time by three years, was estimated from the cross-sectional association between each health-related factor and age in the 2010 survey.

The distribution of participation counts when combined across the disparate activity types was extremely right-skewed. This resulted in a poor fit of the Poisson and negative binomial regression models that are commonly used to model count data. Therefore the data on 'number of times participated' were grouped into three categories (never, 1-5 times, 6+ times) for analysis. These three categories were considered a reasonable classification of the participation data into categories indicative of the respondents' engagement with their workplace health and wellbeing program. Approximately one third of respondents chose not to participate at all, around half participated occasionally (1-5 times), and the remainder participated more often (between 6 and 300 times). The constraints imposed in fitting logit-link (18) and log-link (19) ordinal regression models resulted in statistically significant loss of model fit, so ordinal regression was not used and the response was treated as a nominal outcome with three attributes. Log multinomial regression (20) was used to estimate prevalence and ratios of prevalence of participation in health and wellbeing activities classified into the three categories at levels of study factors. Respondents who indicated that none of the five main activity types was available to them, and who were therefore unable to participate, were excluded from the analysis.

Results

The characteristics of the subjects who participated in both surveys are summarised in Table 1. A higher percentage of the cohort respondents were women than men. The majority of respondents were married or in a de facto relationship, and had a university degree or some other type of post-year 12 educational qualification (diploma/certificate or trade qualification). The mean age of the respondents was 46.6 years in 2010. More respondents were in fulltime employment than part-time or casual employment, and a regular Monday-Friday working week was the most common work schedule.

Table 2 shows mean levels or prevalence of health-related factors in the cohort in 2010 and 2013, and changes between the two surveys. There were small but statistically significant increases in BMI and daily serves of vegetables, and a small but statistically significant decrease in the proportion of smokers. Based on the cross-sectional association with age in the 2010 data, 64% of the increase in BMI could be attributable to the three additional years of age. There was no cross-sectional association between age and vegetable intake or age and smoking. For the other health-related factors (daily serves of fruit, daily hours sitting at work, leisure-time physical activity, psychological distress and risky alcohol use), there was no change in mean levels between surveys. For these factors the number of cohort participants who moved from the high risk category to a lower risk category was approximately balanced by the number who moved from a lower-risk category to the high-risk category.

The mean number of activity types available per person increased by 0.57 in number (around a third) between the 2010 and 2013 surveys (see Table 3). Overall, 21% (120/580) of cohort respondents reported that none of the 5 main activity types was available to them in 2010, whereas only 14% (84/580) reported that none of these activities was available to them in 2013. Of those to whom nothing was available in 2010, 73 people (13% of the cohort) reported that at least one activity type was available in 2013. In comparison, only 40 people (7% of the cohort) reported activities being available to them in 2010 but not in 2013. The mean number of amenities available per person increased slightly from 3.9 in 2010 to 4.1 in 2013.

The proportion of respondents who had participated at least once if activities were available increased from 66% in 2010 to 72% in 2013 (see Table 3). The mean number of times each respondent reported participating (excluding those for whom no activities were available) was 7.2 (Standard Deviation (SD)=19.5) in 2010 and 6.8 (SD=22.6) in 2013. These means are not directly comparable because the 2010 survey required respondents to report their participation over the previous year, whereas the 2013 survey required respondents to report their participation over the previous 3 years. Figure 2 illustrates the percentage of respondents in each of three participation categories (0, 1-5 and 6+ times participated) in 2010 and 2013.

Table 4 shows prevalence of participation in 2013, and ratios of prevalence, at levels of relevant study factors. For these analyses, the number of times participated has been grouped as 0, 1-5 and 6+ times, and the analyses are restricted to those for whom at least one activity type was available in 2013. The results are adjusted for participation in 2010. Prevalence of participation 6+ times in 2013, conditional on participation in 2010, was higher for men than women and higher when a greater number of amenities or activity types were available. Importantly, none of the seven health-related factors (BMI, fruit and vegetable consumption, hours of weekly leisure activity, hours of daily sitting at work, psychological distress, risky alcohol use and current smoking) were associated with higher levels of participation in 2013.

Table 5 shows the results of analyses of each of the three statistically significant predictors of prevalence of participation (sex, number of amenities available in 2013 and number of activity types available in 2013) with mutual adjustment for the other two factors. There was a statistically significant interaction between the number of activity types available in 2013 and the effect of having participated in 2010 (p<0.001). Participation in 2010 had a modifying effect on the association between the number of activity types available in 2013 and participation in the 1-5 times and 6+ times categories in 2013. The prevalence ratios for participation in the 1-5 times category increased with increasing number of activity types available in 2013 for those who had not participated in 2010 (trend p<0.001) but not for those who had participated in 2010. The prevalence of participation in the 6+ times category increased with increasing number of activity types available in 2013 for both those who had and those who had not participated in 2010, but the association was only statistically significant for the respondents who had participated in 2010 (trend p<0.01). There was not strong evidence of an association with sex or the number of amenities, in either of the categories of participation in 2013, after adjusting for each other factor and for participation in 2010. Including a covariate in the model for any of the other factors in Table 4 did not change the coefficients of the factors in Table 5 by more than 10%. As a sensitivity analysis we distributed the missing data from the 3264 non-respondents into differing categories of availability of activities and participation. We found that it would take more than a 50% change across the categories of either activity availability or participation to remove the observed trends.

Discussion

This study investigated whether factors related to the health of employees in the Tasmanian State Service were improved by participation in health and well-being activities during the first three years of the H@W intervention. There was no evidence that they were. Nevertheless the intervention appears to have been well implemented, with survey respondents reporting increased availability of health and wellbeing activities and supportive amenities over the three years, and an increased proportion of respondents reporting participation in activities. The importance of availability of activities and amenities was revealed in associative analyses, because those enabling factors were predictive of greater participation. The strongest predictors of participation in 2013 were having already participated in workplace health and wellbeing activities in 2010, before the implementation of Healthy@Work, and having a wide range of activity types available. These factors were not associated with participation independently of each other: respondents who were already participating in 2010 were more likely to report participating 6 or more times in 2013 if a greater range of activity types was available, whereas respondents who were *not* already participating in activities in 2010 were more likely to report participating 1-5 times in 2013 if a greater range of activity types was available. This suggests that by increasing the range of activity types available, H@W was successful in encouraging non-participants in 2010 to participate in 2013, and in encouraging those already participating in 2010 to either continue to participate or to participate more often in 2013.

Healthy@Work was unique in that it was a large public sector health intervention, implemented across a demographically and geographically diverse population divided between 15 separate agencies of the Tasmanian State Service. Despite the inherent challenges involved in delivering comprehensive workplace health and wellbeing programs to a large and diverse employee population, the programs developed by the TSS agencies were considered to be well-resourced and underpinned by the best available evidence (21). Previous studies have found that workplace health and wellbeing programs, when well developed and integrated into an organisation's management structure, can play a key role in improving employee health outcomes (22). However, changes in health outcomes take time to emerge. In a recent review of the effectiveness of workplace health promotion programs, Goetzel *et al.* (22) argued that a study duration of at least 3 years is necessary to detect population health effects. In this context it is not surprising that no changes in health outcomes in response to the H@W intervention were found after just 3 years, in either the cross-sectional (23) or cohort analyses. In addition, because of the diversity of starting points across different agencies in terms of existing programs and inevitable delays in implementation of new activities after the initiation of the intervention, the full range of activities was unlikely to have been available in all workplaces for the full three years (21).

A strength of this study was the substantial cohort of 580 repeat respondents that enabled us to investigate factors associated with individual-level changes in health-related factors and the respondents' engagement with the H@W intervention. Because the variation in measures within subjects between time points is generally less than the variation between subjects at a given time point, cohort analyses have more statistical power than cross-sectional analyses of the same population, increasing the likelihood of modest effects of an intervention being identified (24). In addition, with the cooperation of the Tasmanian State Service we had access to administrative data that provided demographic and employment information for every employee, enabling us to use inverse probability weighting for non-response. Our findings are therefore generalisable to all employees in the Tasmanian State Service.

This study has limitations. Inverse probability weighting will provide valid estimates if the missing data are ignorable (that is, the values of the missing data are not related to the reason they are missing). This is a requirement that cannot be verified from the data. However our sensitivity analyses suggested that the reported trends in association between availability of and participation in activities are robust to moderately differential patterns of missingness with up to 50% change in proportions in participation categories and categories of availability of activities.

Because the data on health-related factors and on perceived availability of and participation in activities were collected by retrospective self-report, another possible source of bias is systematic errors of recall. The use of standardised and validated measures, such as the IPAQ-Long questionnaire for physical activity, and the Kessler psychological distress scale, increases the likelihood of accurate measurement of health-related factors. However, although we have no reason to believe this to be the case, it is possible that those reporting lesser participation in workplace activities mistakenly reported fewer activity types being available to them. Finally, the differences between the 2010 and 2013 surveys in the survey questions on participation and on availability of activities and amenities made it difficult to accurately compare exposure to the intervention and participation in workplace activities between surveys. Consistent measures of exposure and participation at the different time points would have enabled us to exclude differential measurement error in the assessment of changes over time.

Conclusion

The Healthy@Work intervention in the Tasmanian State Service was responsible for increased availability of and participation in health and wellbeing activities, but there was little evidence of improvement in health-related factors for this group of respondents over the three year period of this study. Changes in the health-related factors were expected outcomes of the intervention but a study duration of just three years is possibly too short to allow change to be manifest.

References

1. Sorensen G, Landsbergis P, Hammer L, Amick BC, 3rd, Linnan L, Yancey A, et al. Preventing chronic disease in the workplace: a workshop report and recommendations. Am J Public Health 2011;101 Suppl 1:S196-207.

2. Partnership NPH. Blueprint for nation-wide surveillance of chronic diseases and associated determinants. In: Partnership NPH, editor. Melbourne, Australia; 2006.

3. Kilpatrick M, Sanderson K, Blizzard L, Nelson M, Frendin S, Teale B, et al. Workplace health promotion: what public-sector employees want, need, and are ready to change. J Occup Environ Med 2014;56(6):645-51.

4. Richmond R, Wodak A, Bourne S, Heather N. Screening for unhealthy lifestyle factors in the workplace. Aust N Z J Public Health 1998;22(3 Suppl):324-31.

5. Sanderson K. Health and Wellbeing of the Tasmanian State Service. Summary of findings from the online health and wellbeing survey conducted for Healthy@Work.: Menzies Research Institute Tasmania; 2009.

6. World Health Organisation. Preventing chronic diseases: A vital investment. Geneva, Switzerland: World Health Organization; 2005.

7. Craig CL, Marshall AL, Sjostrom M, Bauman AE, Booth ML, Ainsworth BE, et al. International physical activity questionnaire: 12-country reliability and validity. Medicine and Science in Sports and Exercise 2003;35(8):1381-1395.

8. Haskell WL, Lee IM, Pate RR, Powell KE, Blair SN, Franklin BA, et al. Physical activity and public health - Updated recommendation for adults from the American college of sports medicine and the American heart association. Circulation 2007;116(9):1081-1093.

9. Ford ES, Caspersen CJ. Sedentary behaviour and cardiovascular disease: a review of prospective studies. International Journal of Epidemiology 2012;41(5):1338-1353.

10. Gual A, Segura L, Contel M, Heather N, Colom J. AUDIT-3 and AUDIT-4: Effectiveness of two short forms of the Alcohol Use Disorders Identification Test. Alcohol and Alcoholism 2002;37(6):591-596.

11. Practitioners TRACoG. Smoking, nutrition, alcohol, physical activity (SNAP): A population health guide to behavioural risk factors in general practice, 2nd edn. . Melbourne; 2015.

12. National Health and Medical Research Council. Australian Dietary Guidelines. In. Canberra; 2013.

13. Furukawa TA, Kessler RC, Slade T, Andrews G. The performance of the K6 and K10 screening scales for psychological distress in the Australian National Survey of Mental Health and Well-Being. Psychological Medicine 2003;33(2):357-362.

14. Slade T, Grove R, Burgess P. Kessler Psychological Distress Scale: normative data from the 2007 Australian National Survey of Mental Health and Wellbeing. Australian and New Zealand Journal of Psychiatry 2011;45(4):308-316.

15. Australian Bureau of Statistics. Australian Health Survey: First results, 2011-12. In. Canberra: Australian Bureau of Statistics; 2012.

16. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res 2013;22(3):278-95.

17. Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions. 3 ed: Wiley; 2013.

18. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. Communications in Statistics-Theory and Methods 1980;9(10):1043-1069.

19. Blizzard CL QS, Canary JD and Hosmer DW. Log-Link Regression Models for Ordinal Responses. Open Journal of Statistics 2013;3:16-25.

20. Blizzard L, Hosmer DW. The log multinomial regression model for nominal outcomes with more than two attributes. Biom J 2007;49(6):889-902.

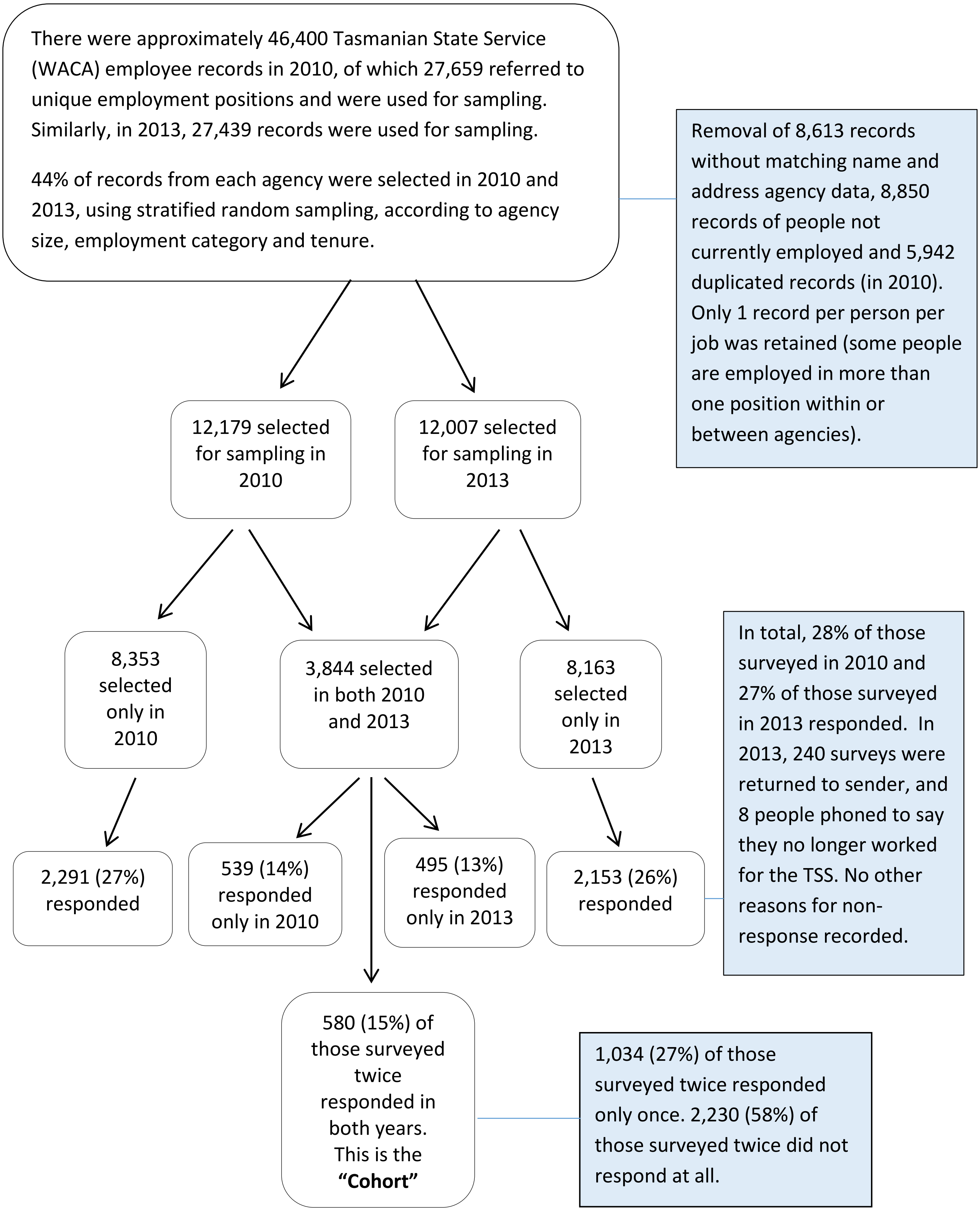
21. Kilpatrick M SK, Blizzard L, Teale B, and Venn A Factors associated with availability of, and employee participation in, comprehensive workplace health promotion (WHP) in a large and diverse Australian public-sector setting: a cross-sectional survey. Journal of Occupational and Environmental Medicine In press.

22. Goetzel RD*ea*. Do Workplace Health Promotion (Wellness) Programs Work? JOEM 2014;56(9):927:934.

23. Kilpatrick M, Sanderson, K., Blizzard, L., Teale, B., & Venn, A. Benefits of workplace health promotion in a large, diverse Australian public-sector setting: a repeated cross-sectional study. (under preparation)

24. Atienza AA, King AC. Community-based health intervention trials: an overview of methodological issues. Epidemiol Rev 2002;24(1):72-9.

Figure 1. Flowchart of sampling process and response proportions.



There were approximately 46,400 Tasmanian State Service (WACA) employee records in 2010, of which 27,659 referred to unique employment positions and were used for sampling. Similarly, in 2013, 27,439 records were used for sampling.

44% of records from each agency were selected in 2010 and 2013, using stratified random sampling, according to agency size, employment category and tenure.

Removal of 8,613 records without matching name and address agency data, 8,850 records of people not currently employed and 5,942 duplicated records (in 2010). Only 1 record per person per job was retained (some people are employed in more than one position within or between agencies).

12,179 selected for sampling in 2010

12,007 selected for sampling in 2013

8,353 selected only in 2010

3,844 selected in both 2010 and 2013

8,163 selected only in 2013

In total, 28% of those surveyed in 2010 and 27% of those surveyed in 2013 responded. In 2013, 240 surveys were returned to sender, and 8 people phoned to say they no longer worked for the TSS. No other reasons for non-response recorded.

2,291 (27%) responded

539 (14%) responded only in 2010

495 (13%) responded only in 2013

2,153 (26%) responded

580 (15%) of those surveyed twice responded in both years. This is the **"Cohort"**

1,034 (27%) of those surveyed twice responded only once. 2,230 (58%) of those surveyed twice did not respond at all.

Figure 2. Histograms of counts and categories of participation in health and wellbeing activities during the year prior to the 2010 survey and the three years prior to the 2013 survey (excluding those to whom no activities were available).

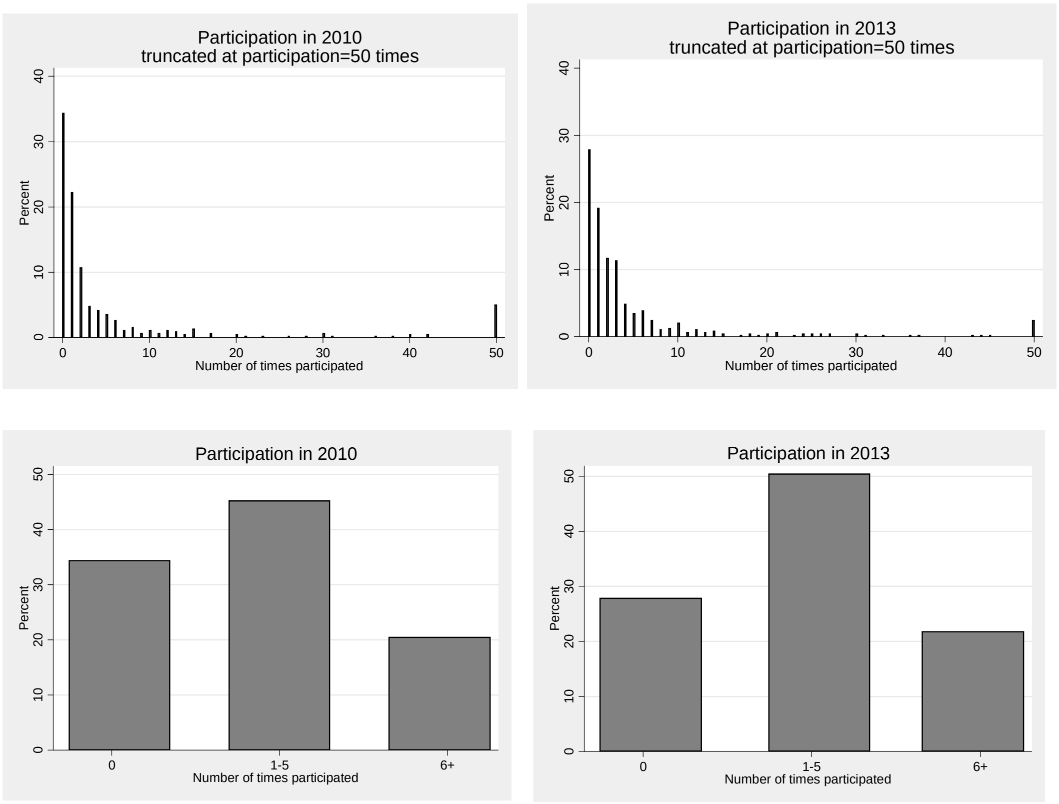


Table 1. Characteristics of the 580 cohort respondents in 2010 and 2013

	2010	2013
Age (years), mean (SD*)	46.6 (8.9)	49.7 (8.9)
Age group % (n/N):		
<40 years	22% (130/580)	15% (88/580)
40-49 years	37% (217/580)	29% (166/580)
50-59 years	34% (200/580)	46% (264/580)
60+ years	6% (33/580)	11% (62/580)
Marital status, % (n/N):		
Single	10% (59/569)	9% (52/579)
Separated	11% (62/569)	10% (60/579)
Widowed	2% (9/569)	2% (9/579)
Married	64% (364/569)	66% (381/579)
De facto	13% (75/569)	13% (77/579)
Highest level of education, % (n/N):		
Year 12 or less	23% (134/575)	21% (120/580)
Post year 12 (University degree, diploma/certificate or trade qualification)	77% (441/575)	79% (460/580)
Occupation category, % (n/N):**		
Blue collar	17% (99/573)	17% (97/573)
White collar/professional	44% (253/573)	42% (242/573)
Services	39% (221/573)	41% (234/573)
Employment condition, % (n/N):		
Fulltime employment	64% (369/580)	60% (347/580)
Part-time/casual	36% (211/580)	40% (233/580)
Work schedule, % (n/N):		
Regular Mon-Fri every week	60% (347/578)	58% (336/577)
Set days but not Mon-Fri	25% (144/578)	27% (153/577)
Days vary from week to week	15% (87/578)	15% (88/577)

*SD=Standard Deviation

**Grouped according to similar industrial awards. White collar/Professional includes management.

Table 2. Mean levels or prevalence of health-related factors in 2010 and 2013, and change in means or prevalence between surveys

Health-related factors:	**2010**		**2013**		**Difference* (95% CI**)**	**Changed from lower to higher risk % (n/N)**	**Changed from higher to lower risk % (n/N)**
BMI*, mean (SD)**	26.2	(4.4)	26.5	(4.7)	**0.28 (0.07, 0.50)**	8 (40/502)	5 (23/502)
Daily serves vegetables[†], mean (SD)	2.8	(1.2)	2.9	(1.3)	**0.10 (0.00, 0.20)**	4 (25/577)	8 (45/577)
Daily serves fruit[†], mean (SD)	1.9	(0.9)	1.9	(0.9)	0.01 (-0.06, 0.07)	11 (65/576)	11 (64/576)
Hours of leisure activity/week[‡], mean (SD)	3.5	(3.7)	3.3	(3.4)	-0.19 (-0.49, 0.12)	15 (89/578)	17 (99/578)
Hours sitting at work/day[#], mean (SD)	4.7	(2.6)	4.6	(2.6)	-0.06 (-0.24, 0.12)	10 (54/564)	11 (61/564)
Psychological distress[§] (K10 score), mean (SD)	14.9	(4.9)	14.7	(4.7)	-0.21 (-0.57, 0.16)	6 (34/573)	6 (34/573)
Risky alcohol use[‖], % (n/N)	43%	(246/574)	43%	(243/564)	0.00 (-0.04, 0.04)	12 (67/558)	12 (68/558)
Current smoking[¶], % (n/N)	9%	(51/579)	7%	(39/578)	**-0.02 (-0.04,-0.01**)	1 (4/577)	3 (16/577)

*Data weighted using the inverse probability of response

**CI=Confidence Interval, SD=Standard Deviation

***BMI=Body Mass Index, Higher risk=BMI>25

[†] Higher risk=less than 4 serves of vegetables and/or 1 serve of fruit per day

[‡] Higher risk=less than 150 mins moderate-vigorous physical activity per week

[#] Higher Risk= sitting at work for 6 hours or more per day

[§] Higher risk=k10 score>21

[‖] From AUDIT-C and Royal Australian College of General Practitioners guidelines, see methods section

[¶] Current smokers are defined as being at higher risk than previous or never-smokers

*Table 3. Change in availability of activities and amenities and participation** in activities between the 2010 and 2013 surveys*

	2010	**2013**	**Difference (95% CI*) *****
Number of amenities, mean (SD*):	3.9 (2.2)	4.1 (2.1)	**0.18 (0.00, 0.36)**
Number of health & wellbeing activity types available, mean (SD):	1.7 (1.3)	2.2 (1.5)	**0.58 (0.46, 0.69)**
Proportion with no activities available, % (n/N):	21% (120/580)	14% (84/580)	**-0.07 (-0.10, -0.03)**
Participation in activities (if activities available), % (n/N):**			
No participation (0)	34% (158/460)	28% (138/496)	**-0.08 (-0.14, -0.02)**
Low participation (1-5 times)	45% (208/460)	50% (250/496)	0.06 (-0.01, 0.13)
Moderate-high participation (6+ times)	20% (94/460)	22% (108/496)	0.02 (-0.04, 0.08)

*SD=Standard Deviation, CI=Confidence Interval

**Participation was measured during the last year (2010) or the last 3 years (2013)

***Data weighted using the inverse probability of response

Table 4. Association of characteristics of the cohort respondents with participation in 2013, adjusted for participation in 2010.

	Did not participate *	Participated 1-5 times		Participated 6+ times	
	% (n/N)	% (n/N)	PR** (95% CI)	% (n/N)	PR** (95% CI)
Sex:					
Men	25 (34/137)	45 (61/137)	1.00	31 (42/137)	1.00
Women	29 (104/357)	53 (188/357)	1.20 (0.94, 1.53)	18 (65/357)	**0.63 (0.44, 0.89)**
Age group:					
<40	27 (22/82)	50 (41/82)	1.00	23 (19/82)	1.00
40-49	30 (45/148)	51 (76/148)	1.07 (0.79, 1.47)	18 (27/148)	0.84 (0.50, 1.42)
50-59	23 (52/222)	54 (120/222)	1.10 (0.82, 1.48)	23 (50/222)	0.95 (0.59, 1.51)
60+	43 (19/44)	30 (13/44)	0.70 (0.41, 1.20)	27 (12/44)	1.07 (0.56, 2.05)
Trend			p=0.76		p=0.89
Marital status:					
Single/divorced	28 (29/105)	51 (54/105)	1.00	21 (22/105)	1.00
Married/de facto	28 (109/390)	50 (196/390)	1.08 (0.83, 1.40)	22 (85/390)	0.87 (0.58, 1.32)
Trend			p=0.58		p=0.52
Education:					
No post-year 12	30 (28/94)	49 (45/94)	1.00	22 (21/94)	1.00
Post year 12	27 (110/402)	51 (205/402)	1.02 (0.79,1.32)	22 (87/402)	0.95 (0.63, 1.44)
Trend			p=0.87		p=0.87
Occupation group:					
Blue collar	33 (27/81)	51 (41/81)	1.00	16 (13/81)	1.00
White collar/professional	25 (52/212)	51 (109/212)	1.06 (0.78, 1.45)	24 (51/212)	1.00 (0.59, 1.68)
Services	29 (56/196)	49 (96/196)	1.02 (0.74, 1.39)	22 (44/196)	0.98 (0.58, 1.67)
Employment condition:					
Fulltime employment	28 (85/302)	50 (150/302)	1.00	22 (67/302)	1.00
Part-time/casual	27 (53/194)	52 (100/194)	1.09 (0.89, 1.34)	21 (41/194)	0.82 (0.57, 1.17)
Trend			p=0.40		p=0.28
Work schedule:					
Mon-Fri	27 (78/292)	51 (149/292)	1.00	22 (65/292)	1.00
Days vary week to week	29 (22/75)	45 (34/75)	0.97 (0.72, 1.31)	25 (19/75)	0.98 (0.62, 1.54)
Other set days	29 (37/127)	52 (66/127)	1.04 (0.82, 1.32)	19 (24/127)	0.82 (0.53, 1.26)
No. of amenities available:					
0-3	34 (30/88)	53 (47/88)	1.00	13 (11/88)	1.00
4-7	30 (77/253)	53 (134/253)	0.94 (0.72, 1.22)	17 (42/253)	1.30 (0.65, 2.61)
8-14	20 (31/155)	45 (69/155)	0.82 (0.61, 1.12)	35 (55/155)	**2.24 (1.13, 4.45)**
Trend			p=0.18		**p=0.001**
No. of activity types available:					
1	49 (59/121)	45 (55/121)	1.00	6 (7/121)	1.00
2	30 (37/122)	52 (64/122)	**1.44 (1.04,1.99)**	17 (21/122)	1.86 (0.77, 4.40)
3	13 (15/120)	59 (71/120)	**1.40 (1.00,1.97)**	28 (34/120)	**3.78 (1.71, 8.36)**
4	23 (21/92)	50 (46/92)	1.27 (0.88, 1.84)	27 (25/92)	**3.04 (1.31, 7.07)**
5	15 (6/41)	34 (14/41)	0.90 (0.52, 1.56)	51 (21/41)	**5.38 (2.34, 12.38)**
Trend			p=0.97		**p<0.001**
BMI*:**					
Underweight/normal	33 (65/197)	49 (96/197)	1.00	18 (36/197)	1.00
Overweight/obese	22 (55/255)	54 (137/255)	1.13 (0.91, 1.39)	25 (99/452)	1.22 (0.83, 1.78)
Fruit and vegetable intake:					
Sufficient	32 (14/44)	41 (18/44)	1.00	27 (12/44)	1.00
Insufficient	27 (123/451)	51 (232/451)	1.33 (0.84, 2.11)	21 (96/451)	0.86 (0.49, 1.50)
Leisure activity:					
Sufficient	26 (69/269)	51 (137/269)	1.00	23 (63/269)	1.00
Insufficient	30 (69/227)	50 (113/227)	1.03 (0.84, 1.26)	20 (45/227)	0.97 (0.69, 1.36)
Sitting at work:					
<6hrs/day	29 (76/259)	48 (124/259)	1.00	23 (59/259)	1.00
>=6hrs/day	26 (60/233)	54 (125/233)	1.10 (0.90, 1.34)	21 (48/233)	0.90 (0.64, 1.27)
Psychological distress:					
Low/moderate	29 (128/445)	50 (221/445)	1.00	22 (96/445)	1.00
High/very high	20 (10/50)	58 (29/50)	1.15 (0.86, 1.54)	22 (11/50)	1.03 (0.60, 1.77)

Alcohol use:					
Not risky	28 (76/271)	54 (145/271)	1.00	18 (50/271)	1.00
Risky	27 (58/212)	48 (101/212)	0.91 (0.74, 1.12)	25 (53/212)	1.20 (0.84, 1.70)
Smoking:					
Not a current smoker	27 (125/463)	51 (237/463)	1.00	22 (101/463)	1.00
Current smoker	39 (12/31)	39 (12/31)	0.93 (0.60, 1.46)	23 (7/31)	0.77 (0.36, 1.63)

* in the 3 years between 2010-2013, excluding those to whom no activities were available.

**PR(95% CI)=prevalence ratio (95% confidence interval) estimated from data weighted by the estimated inverse probability of response and adjusted for participation in 2010.

***BMI=Body Mass Index

Table 5. Prevalence of participation in the three years prior to the 2013 survey in three categories, and ratios of prevalence, for increasing number of amenities and activity types available.*

		Categories of participation in 2013				
		Did not Participate*	Participated 1-5 times		Participated 6+ times	
		% (n/N)	% (n/N)	PR (95% CI)**	% (n/N)	PR (95% CI) **
Sex:	Male	24.8 (34/137)	44.5 (61/137)	1.0	30.7 (42/137)	1.0
	Female	29.1 (104/357)	52.7 (188/357)	1.10 (0.79, 1.53)	18.2 (65/357)	0.45 (0.17, 1.18)
Number of amenities available in 2013:						
	low (0-3)	31.0 (22/71)	56.3 (40/71)	1.0	12.7 (9/71)	1.0
	medium (4-7)	29.1 (62/213)	52.6 (112/213)	0.89 (0.69, 1.15)	18.3 (39/213)	1.40 (0.79, 2.48)
	high (8-14)	16.6 (23/139)	44.6 (62/139)	0.81 (0.58, 1.14)	38.9 (54/139)	1.43 (0.74, 2.77)
	trend			p=0.25		p=0.36
Non-participants in 2010 Number of activity types available in 2013:						
	1	65.7 (23/35)	25.7 (9/35)	1.0	8.6 (3/35)	1.0
	2	40.6 (13/32)	46.9 (15/32)	**2.37 (1.17, 4.81)**	12.5 (4/32)	2.22 (0.50, 9.90)
	3	24.1 (7/29)	65.5 (19/29)	**3.30 (1.69, 6.44)**	10.3 (3/29)	1.11 (0.23, 5.22)
	4	22.9 (8/35)	62.9 (22/35)	**3.20 (1.63, 6.29)**	14.3 (5/35)	2.14 (0.56, 8.23)
	5	10.0 (1/10)	70.0 (7/10)	**3.88 (1.78, 8.44)**	20.0 (2/10)	3.27 (0.48, 22.12)
	Trend			**p<0.001**		p=0.23
Participants in 2010 Number of activity types available in 2013:						
	1	37.5 (18/48)	54.1 (26/48)	1.0	8.3 (4/48)	1.0
	2	23.3 (17/73)	58.9 (43/73)	1.07 (0.77, 1.49)	17.8 (13/73)	2.54 (0.85, 7.54)
	3	5.2 (4/77)	57.1 (44/77)	0.99 (0.65, 1.50)	37.7 (29/77)	**7.32 (2.12, 25.28)**
	4	22.2 (12/54)	40.7 (22/54)	0.86 (0.55, 1.35)	37.0 (20/54)	**3.10 (1.07, 9.12)**
	5	13.3 (4/30)	23.3 (7/30)	0.49 (0.22, 1.08)	63.3 (19/30)	**5.42 (1.85, 15.96)**
	trend			p=0.18		**p=0.009**

*Excluding those to whom no activities were available.

** PR (95% CI)=Prevalence Ratio (95% Confidence Interval). Data used in estimating prevalence ratios are weighted using inverse probability of response.

Statistical appendix

A1: Data Management

The datasets used in this study were the two *p*H@W cross-sectional survey datasets of information collected in 2010 and 2013. There were 3410 respondents in 2010 and 3228 in 2013. Individual employees were assigned a unique employee number, and so the cohort group was easily identified as those with an employee number that was included in both datasets (n=580). The relevant variables included in this study were demographic and workplace related variables, variables measuring availability of amenities and activities and participation in activities, and measures of health-related factors.

A large and time-consuming part of the study consisted of cleaning the data, and creating exposure variables (availability of activities and amenities) and an outcome variable (participation in activities) that were meaningful and consistent between surveys. The main issues encountered (and solutions reached) when developing these variables were:

1. Were employees asked about the same activities and amenities in both surveys?

   No. The activities asked about in 2013 were changed in response to information gathered in the 2010 survey in order to better differentiate between different activities. For this reason, the activities were grouped into five main "activity types", which could be considered relatively consistent between the two surveys.

2. Should we combine participation across all activity types (physical activities, mental health activities, education, injury prevention/rehab, health assessments), or look at participation in each type separately?

   There are large differences in the frequency of activities offered in each 'activity type', with 'physical activities' potentially being offered several times per week while 'health education' or 'health assessments' are only run occasionally. The health impacts of participating once in each of these disparate activities may not be comparable. Nonetheless, in this study we consider the counts of participation in different activities to be additive, with the sum of 'number of times participated' across all activities considered as a particular 'dose' of the H@W intervention. This may be to some extent an artificial construction of an exposure variable. In addition, predictors of participation in the different types of activities may differ. For this reason other studies using the entire pH@W data have focused on specific activity *types such as mental health activities, or “SNAP” (smoking, nutrition, alcohol and physical activity) related activities. Nonetheless, for this cohort study, by focusing instead on the broader patterns of association with participation across the entire suite of activities, it was hoped that the overall success or otherwise of the H@W intervention would not be obscured.*

3. What should we do about the difference in time frames between the two surveys, when considering responses to the questions about availability of and participation in activities?

The pH@W surveys asked employees to record the activities and amenities available in their workplace, and the number of times they had participated during “the last year” in 2010 and during “the last 3 years” in 2013. One possibility was to divide the 2013 participation count by three to obtain a mean yearly participation count. This would have resulted in an annual average participation count that was significantly lower than the 2010 average count. This difference could be partly explained by the use of an inappropriate divisor, when some and possibly many subjects were not exposed to workplace activities for the entire period of three years because they were recent entrants to the State Service, or the activities were available only towards the end of the period. There is also likely to be greater recall bias when reporting infrequent events over a three year period than reporting over a one year period, and so we can be less certain of the accuracy of reported participation from the 2013 survey. Thus comparing availability and participation between the two surveys remains problematic.

A2: Missing data, and the use of inverse probability weighting to correct for non-response.

There was a low response proportion in this study. Only 15% (580/3844) of those selected to participate in both surveys responded both times (Table A2.1).

Table A2.1 Response status of those invited to participate in both the 2010 and 2013 surveys

	Invited to participate in both 2010 and 2013		
	Responded 2013	Did not respond 2013	Total
Responded 2010	580	539	1119
Did not respond 2010	495	2230	2725
Total	1075	2769	3844

Missing data are a common problem in longitudinal studies designed to collect data on each participant at multiple time points. The validity of the statistical methods used to analyse incomplete data depend on the missing data mechanisms. Rubin (1) introduced a hierarchy to formalise this dependency. Data are missing completely at random (MCAR) if the probability that each response is missing is unrelated to the observed and missing responses (and covariates), missing at random (MAR) if the probability depends on the observed data but is unrelated to the missing data, and missing not at random (MNAR) otherwise.

(1) Missing Completely at Random (MCAR)

If the missing data were "missing completely at random" (MCAR), the associations between study exposures and outcomes would not differ between those who declined to respond and those who did respond and there would therefore be no bias if we analysed the data using the information on respondents only. This is known as complete-case analysis (2).

(2) Missing at Random (MAR)

Because we have no exposure or outcome measures for those who did not respond, we cannot determine whether the missing data are MCAR. We do have complete information on some demographic and employment-related characteristics for all TSS employees, irrespective of whether they responded to the surveys. There is complete information on agency, permanent/non-permanent employment category and fulltime/part-time employment condition, age, sex and service length. We will refer to these variables as the "fully observed variables". The proportion of those surveyed at both time points who responded to both surveys differed according to observed levels of these variables: for example, women were more likely to respond to both surveys than men, and older employees were more likely to respond to both surveys than younger employees. If we can assume that the probability of response within the categories of the fully observed variables is not associated with the main study outcomes, then the missing data are said to be "Missing at Random" (MAR). When missing data are MAR, maximum likelihood estimation of the likelihood function for the observed data, or Bayesian estimation of a statistical model that includes an assumed prior distribution for the measurements, yield valid estimates. In addition, frequentist methods such as multiple imputation or inverse probability weighting can be used to adjust the analyses to account for the differences between data for respondents and the missing data for non-respondents (3, 4).

(3) Missing Not at Random (MNAR)

The MAR assumption may not be valid. It is quite possible that within the categories of the fully observed variables, the probability of response is associated with the main study outcomes. The missing data in this situation are termed "Missing Not at Random" (MNAR), and the resulting bias cannot be measured or corrected by inverse probability weighting or any of the other commonly used methods. In this case, more complex analyses are needed to model the missingness mechanism as well as the observed data (4). These techniques will not be discussed further here.

Two methods are commonly used to address non-response bias in analyses if the missing data are MAR:

(1) Multiple imputation (MI)

Multiple imputation (MI) specifies a model for the distribution of missing values given the observed data. Values randomly generated from this model are used to replace the missing values and to create a complete set of data. This process is repeated so that there are several datasets created. The analyses are run on each of these in turn and the resulting estimated parameters are averaged over the datasets (2).

(2) Inverse probability weighting (IPW)

The data from the respondents in the analyses can be weighted using the inverse of the estimated probability of response (2). This is called inverse probability weighting (IPW). IPW requires the specification of a model for the probability that an individual is a respondent. A logistic regression model is commonly used, although this is for mathematical convenience rather than because there is a particular reason to believe this model is correctly specified. Although MI is more efficient than IPW, in a situation such as this where the missing data are due to non-participation, and data on many variables are missing, it is difficult to correctly specify the model for the joint distribution of the many variables that is required for MI. In this situation, IPW is usually the preferred approach (2), and it is the method used here: the probability of response was estimated using logistic regression on the dataset of all employees selected to participate in both surveys, with response on both occasions as the binary outcome variable and the fully observed variables- age, sex, service length, employment category (permanent/fixed term), employment condition (fulltime/part-time) and agency- as the covariates. The selected covariates were the stratification factors (employment category, employment condition and agency) and other variables (age, sex and service duration) on which data were available for all subjects.

Sensitivity analyses

Inverse probability weighting is not guaranteed to remove non-response bias if the data are not Missing at Random (MAR), or if the model used to estimate the relationship between the fully observed variables and the probability of response is not correctly specified. We undertook a sensitivity analysis to explore the potential influence on our main results of differential patterns of availability of activities and participation among the non-respondents. It appeared that the weighted estimates and trends reported using the data from the respondents were robust to moderately differential patterns of association in the missing data. An additional problem of unstable weights can emerge if the model predicting the probability of response yields very small fitted probabilities, and thus very large weights, for some individuals. The estimation of the analysis models is then dominated by a few very large weights, reducing effective sample size. To ensure that we have created stable weights that accurately reflect the missingness in the data, Seaman and White (2) recommend that the weights be checked. In particular, it is important that the mean predicted probability of response for the individuals with the largest weights corresponds to the actual response proportion for these individuals. A further suggested check is that the sum of the largest 10% of weights for individuals with non-missing data is less than half the total sum of weights for individuals with non-missing data. Both of these checks were carried out on this dataset and the weights appeared to be stable. It is also recommended that the weighted results be compared to unweighted results, to check that the weighting does not result in a 'large' change in either the estimates or standard errors of the main results. In the weighted analysis, the coefficient for the association between sex and participation in the 6-plus times category was reduced by 57%, and the estimated regression coefficients for several of the factor levels and standard errors of the other main covariates changed by up to 33%, although the overall trends and the main findings were unchanged. Therefore although the weighting significantly changed some of the regression coefficients and standard errors, the weighted results were not qualitatively different to the unweighted results.

A3: Investigating change in health outcomes and health risk factors over the three year duration of H@W.

Aim

To investigate whether health-related factors had changed between 2010 and 2013, and if so, to assess whether there was an association between participation in health and wellbeing activities and change in health-related factors.

Rationale

A strength of a longitudinal study design is that the analyses generally have more statistical power than cross-sectional analyses of the same population, increasing the likelihood of modest effects of an intervention being identified (5). This is because the size of the differences between means or proportions at different time points that can be identified by the study is dependent on the standard errors of these differences. The standard errors in a repeated cross-sectional survey are large whenever there are large variations between individuals. In longitudinal analyses, such as this cohort study, the differences within subjects are generally smaller than the differences between subjects, and the standard errors are smaller in consequence (6).

The overall aim of the H@W intervention was to improve the health of the employees in the Tasmanian State Service. Although the pH@W surveys were only undertaken in the first 3 years of the intervention, analysing the data to determine if there had been any change in health–related factors was considered important. Earlier analyses of the repeated cross-sectional survey data had failed to find any change in health-related factors between the two surveys. Of interest was whether the cohort study had the power to detect small differences in health-related factors over time that the analyses of the cross-sectional survey data could not.

Overview of analyses

Paired t-tests were used to investigate within-subject differences in continuous health-related factors between surveys, and the difference in correlated proportions of categorical factors was assessed using the standard error given by Fleiss, Levin and Paik (7) with a continuity correction applied in the calculation of 95% confidence intervals. The numbers of respondents moving into each higher risk category were compared with the numbers moving out of each higher risk category to summarise net movement between risk states over the three years. Linear regression methods were used to investigate whether changes in health-related factors could be attributed to the aging of the cohort over the three year period between surveys, because several of the health outcomes are known to be associated with age. In these data, BMI and fruit and vegetable intake were positively correlated with age, while leisure time physical activity, time spent sitting at work and psychological distress were negatively correlated with age. Because all participants aged by the same amount over the three years of the study, it was not possible to adjust for aging when measuring change in the health-related factors over time. As an approximate method of accounting for age, the cross-sectional association of each health-related factor with age was estimated using the 2010 survey data and the projected longitudinal effect of an additional three years of age was estimated from the cross-sectional regression coefficient.

A4: Exploring the relationship between availability of health and wellbeing activities and supportive amenities, and participation in health and wellbeing activities.

Aim:

To model the association between the availability of activities and amenities, baseline participation in activities, and participation in activities in 2013, while adjusting for potential confounders.

Rationale:

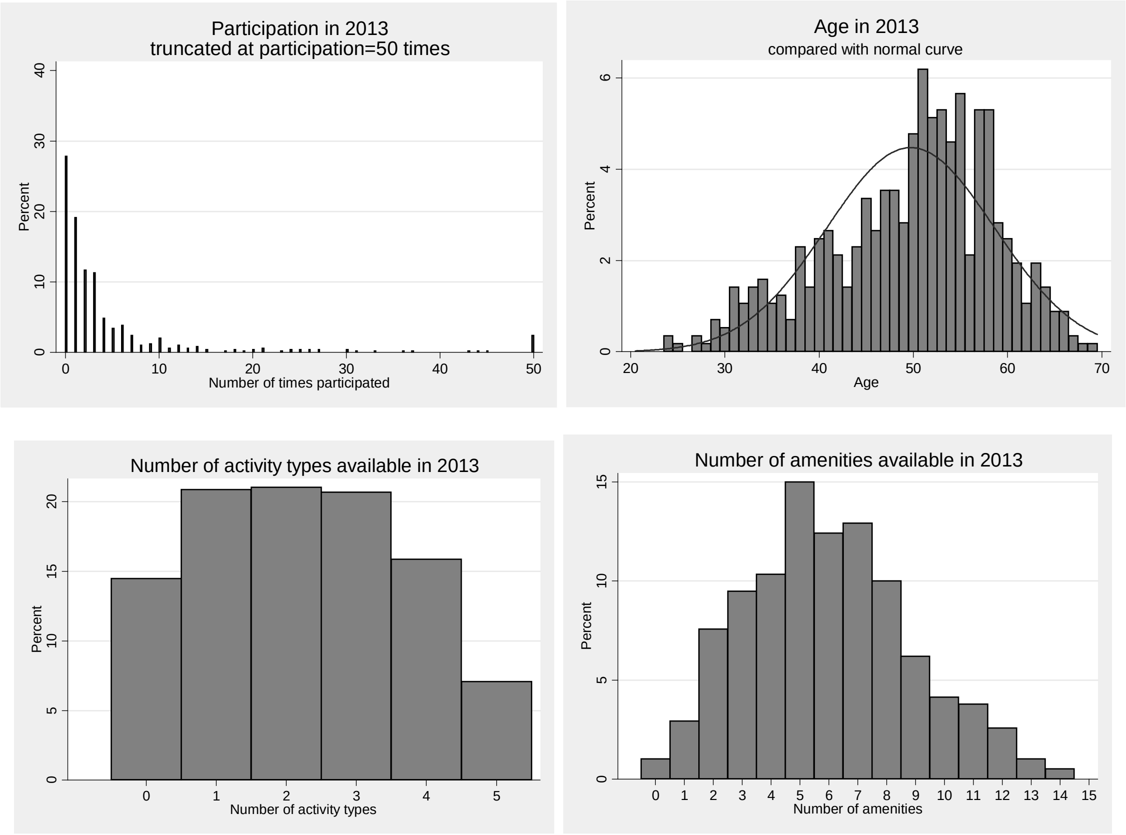
When evaluating the success of a workplace health and wellbeing program, participation can be considered an intermediate step between availability of the program and change in health-related factors. For this study, participation was modelled first as an outcome of a program that provides activities and amenities, and then as a predictor of change in health-related factors.

Overview of analysis steps taken:

(1) *Assessing the distribution of each variable:*

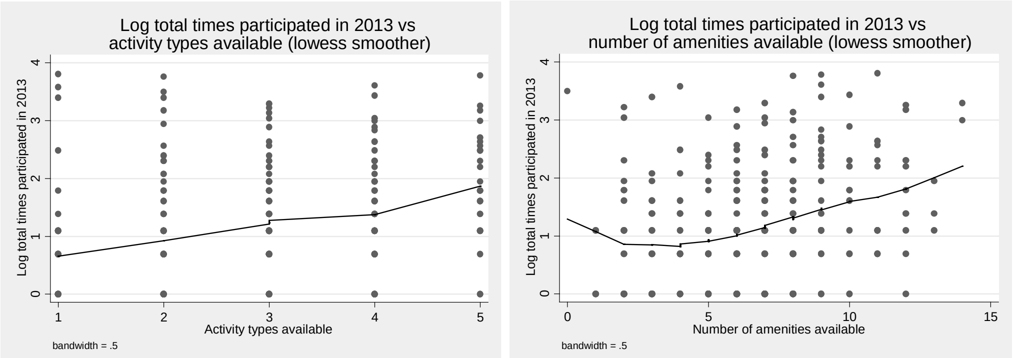
The distribution of each variable was assessed, using histograms for continuous variables and cross-tabulations for categorical variables. Histograms for the outcome “number of times participated”, and the covariates age, number of activity types available and number of amenities available are presented in Figure A4.1 below. Numbers in each category of the categorical variables are presented in tables 3 & 4 above.

Figure A4.1 Assessment of variables:



(2) Checking whether a log link is appropriate for modelling the relationship between participation and the main study factors:

Figure A4.2 Plots of log participation versus each main study factor.



The log of the participation count appears to have a linear association with both the number of amenities and the number of activity types available.

(3) Selecting an appropriate regression model

(i) Modelling the participation data as a count variable

Initial attempts to model the relationship between the participation data and the main study factors used regression models for count outcomes. The count outcome "number of times participated" was extremely right-skewed, with a mean of 9.4 but a variance of 683 due to a small number of very high counts (the highest was 301). Although these high counts were obvious outliers, they were valid data points. The unevenness of the data was due in part to the "grab-bag" nature of the participation variable, which included participation in very different types of activities. For example, each day of cycling or walking to and from work would count as two "participation" events, whereas participation in the Global Corporate Challenge or in a mental health or health education session would count as a single "participation". This meant that the few people who undertook regular activities had very high participation counts, whereas 90% of respondents participated in 10 or fewer activities over the 3 year period.

The mean number of events expected over a period of time is often modelled using a Poisson distribution. The probability distribution function for a Poisson distribution is:

$$\mathrm{P}(Y{=}y){=}\frac{\mu^{y}e^{-\mu}}{y!}\text{ , for } y{=}0{,}1{,}2\ldots,$$

where Y is the number of events and μ is the mean value of Y. The distribution is completely characterised by its mean μ and a key property of the distribution is that the variance is equal to the mean. That is, $\mathrm{Var}(Y)=\mathrm{E}(Y)=\mu$. Because the variance was so much greater than the mean in this study, the associations between the "number of times participated" variable and the main study factors were not well modelled using a Poisson model.

The negative binomial distribution includes a "variance inflation factor" that can be used to adjust the variance independently of the mean, and can be used as an alternative to the Poisson distribution to model over-dispersed count data. The standard negative binomial regression model can be derived as either a Poisson-gamma mixture model, or as a member of the exponential family of distributions (8), which can be modelled under the framework of generalised linear models (GLMs). The derivation of the negative binomial distribution as a member of the exponential family begins with the PDF:

$$\mathrm{P}(Y=y)=\begin{pmatrix} y+r\text{-}1 \\ r\text{-}1 \end{pmatrix} p^r(1\text{-}p)^y,\ \text{for } y=0,1,2\ldots, \text{ and r>0},$$

where *Y* is the random variable 'number of trials before r events have occurred', in a series of Bernoulli trials, and p is the probability of each event occurring. The positive integer *r* is known as the "shape" or "dispersion" parameter. The mean of this distribution is $\mu = b'(\theta) = \frac{r(1-p)}{p}$, the variance is $V(Y) = b''(\theta) = \frac{r(1-p)}{p^2}$, and the variance V(*Y*) and mean *μ* are related by V(*Y*)=*μ+α μ²* , where the "variance inflation parameter" *α=1/r* must be positive.

Re-parameterising p and r in terms of μ and α gives $\frac{1-p}{\alpha p} = \mu,$ so that $p = \frac{1}{1+\alpha\mu}$. Given the defined values of μ and α we can re-parameterise the negative binomial PDF such that

$$f(y;\mu,\alpha) = \begin{pmatrix} y+1/\alpha-1 \\ 1/\alpha-1 \end{pmatrix} \left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y}$$

For this distribution, over-dispersion (the ratio of variance to mean) is 1+ *α μ*, and the extent of over-dispersion increases with the mean.

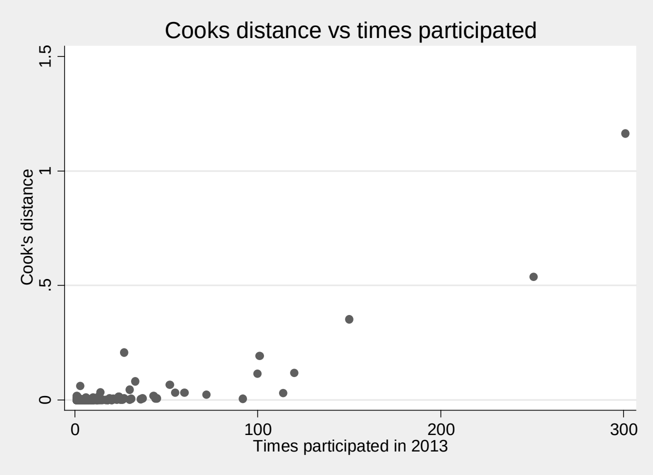
The negative binomial distribution appeared to be a better choice than the Poisson distribution for modelling the participation data. However, after fitting a negative binomial model to the participation data, the outlying observations with very high participation counts still had a large influence on the estimated coefficients. This was shown by the Cook's distance measure. Cook's distance is used to check for a disproportionate influence of any individual observations. It "measures the aggregate change in the estimated coefficients when each observation is left out of the estimation"(9). Cook's distance C_i for each observation i, with i=1,…n, is approximated for generalised linear models using:

$$C_i=(\boldsymbol{\beta}^*_{(i)}-\boldsymbol{\beta})^T\mathbf{I}(\boldsymbol{\beta}^*_{(i)}-\boldsymbol{\beta}) ,$$

where **I** is the negative Hessian, and $\boldsymbol{\beta}^*_{(i)}$ is an approximation to the estimated coefficient vector leaving out the ith observation (10). The approximation is described by Hardin and Hilbe (10) as a "one-step approximation" to the jackknife-estimated coefficient vector. The approximation uses the full observation estimate as a starting point, then takes only one Newton-Raphson iteration, instead of fitting to full convergence each of the n candidate models.

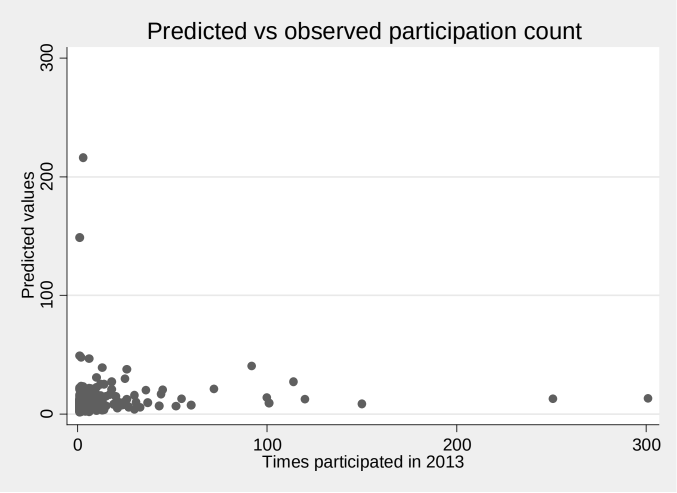
It is suggested that observations with Cook's distance greater than 4/n (where n=the total number of observations) should be considered problematic (10), and in this case 21 of the 349 observations (6% of the total) were greater than the cut-off of 0.011. The problematic observations were those for respondents who had participated many times in 2013, and the more times they participated, the greater the Cook's Distance (see figure A1.3 below). Because all of these participation counts were strictly valid under the measurement protocols, this suggested that the model did not fit the data well- it was excessively influenced by the 6% of observations with high participation counts.

Figure A4.3 Cook's distance: a measure of the aggregate change in the estimated coefficients when each observation is left out of the estimation.



A scatter plot of the predicted versus the actual participation counts also helps to illustrate the problems with this model. The model does not successfully predict the spread of participation counts in the data:

Figure A4.5 Predicted versus observed participation counts from the Negative Binomial model.



(ii) Modelling the participation data after categorisation

Because the count models were not providing an adequate description of the data, options for categorising the participation outcome variable were investigated.

One option was a binary classification (participated/did not participate). Those to whom no activities were available were excluded, so that we were modelling associations with participation outcomes for those who had chosen either to participate or not to participate, and not for those for whom no activities were available. Binary logistic regression could be used to estimate associations with the binary participation outcome, but logistic regression provides odds ratio estimates of relative risk that is the fundamental measure of effect for a closed cohort (11). As an approximation of the relative risk, the odds ratio overstates its target unless the outcome is rare in all exposure and confounder categories (11). To estimate relative risk, Poisson regression with robust standard errors (12) to correct for the misspecified error structure of the binary outcome was used. This model can result in estimated probabilities greater than unity, but this is not an issue if the only purpose of the model is to obtain valid prevalence ratios (13). Because exact predicted values for individuals were not required, this was considered a reasonable model. However, a large amount of information on participation would be lost if the only consideration is whether or not respondents participated, without using the available data on the frequency of participation.

Another option was to model the participation data as a categorical outcome. The categories 0, 1-5, and 6+ times participated were chosen because these categories each contained a substantial proportion of the respondents, and were considered to provide a reasonable distinction between those who were not engaged with workplace health and wellbeing at all, those who were engaged at a low level, and those who were engaged at a higher level. The three participation count categories were reported by 138, 250 and 108 respondents respectively (excluding 84 respondents to whom no activities were available). Because the categories 0, 1-5 and 6+ times participated have a natural order, we initially considered using a log-link ordinal model as described by Blizzard *et al.* 2013 (14). There are three logit-link regression models that are frequently used to deal with ordered categorical response data, which Blizzard *et al.* propose fitting with a log link to allow estimation of risk ratios (or prevalence ratios in cross-sectional studies). The log-link ordinal models are the adjacent-categories (AC) probability model, continuation-ratio (CR) probability model and the proportional probability (PP) model. Each is a constrained form of the log multinomial model, which is the log-link counterpart of the multinomial logistic model (15). The constraints impose a simple linear relationship on the slope coefficients, and it is possible to test whether the constraints result in substantial loss of model fit using asymptotically-equivalent Wald, score and likelihood ratio tests of the constraints (for more detail see Blizzard *et al.* 2013(14)). These tests indicated that each of the ordinal models resulted in significant loss of model fit (p<0.02 for all three tests for each model). On the grounds that the three-level outcome required less data reduction than the binary classification, a log multinomial model was used to estimate relative risk (15).

The log multinomial model for these data is:

$$\Pr(Y_j=1|\mathbf{x})=\pi_j(\mathbf{x})=\exp(\mathbf{x'\beta}_j),\ j=1,2,$$

where the binary (0/1) random variables Y_j, j=0,1,2 indicate which participation category was observed, with $Y_0=1-(Y_1+Y_2)$ and $\Pr(Y_0=1|\mathbf{x})=1-[\ \Pr(Y_1=1|\mathbf{x})+\Pr(Y_2=1|\mathbf{x})]$, and where the mean π_j of each Y_j depends upon a linear combination of the observed values $\mathbf{x}=(x_1, x_2, \ldots x_K)$ of K non-constant covariates $X_1, X_2, \ldots X_K$. In this case there are three outcome categories including the category corresponding to j=0, only two of which can be estimated because the sum of the means (π_j) of each of the binary outcomes Y_j must equal one. The excluded category was chosen to be the "participation=0" category but it could alternatively have been either of the other two. The linear predictor is $\mathbf{x'\beta}_j = \beta_{j0}+ \beta_{j1}x_1+ \beta_{j2}x_2+\ldots.+ \beta_{jk}x_k$, where $\mathbf{x}'=(1,x_1,x_2,\ldots,x_k)$ are the observed values of the constant and the K non-constant covariates $X_1, X_2,\ldots,X_k$, and where $\boldsymbol{\beta}'_j=(\ \beta_{j0,}\ \beta_{j1},\ldots.,\beta_{jk})$, j=1, 2 are parameters to be estimated.

Initially, single factor analyses were undertaken to assess which of the available demographic variables and the availability of activities and amenities were associated with participation in 2013, adjusting for participation in 2010. That is, to find which variables were associated with a change in participation over the three years of the study. For the single factor analyses a log multinomial model with participation categorised in 3 groups (0, 1-5 and 6+ times) was used, as described above. For each covariate we reported the proportions in each participation category, and prevalence ratio of the different levels of the covariate in the low participation and moderate-high participation categories (with 95% CI). To choose which covariates to include in the final adjusted model, we used the results of the single factor analyses. All three variables for which there was moderate evidence against the null hypothesis of no association with participation during the three years prior to the 2013 survey (at p<0.1) were included in the adjusted model. Potential confounders were then included in the model one-by-one. Potential confounders for which data were available were age, sex, education, marital status, permanent or fixed term/casual employment, part-time or fulltime employment, agency, occupation type and health risk factors. If inclusion of a covariate changed the estimated regression coefficient of a principal study factor by more than 10%, or if the variable itself was significantly associated with the outcome, the covariate was included in the final model in accordance with the change-in-parameter-estimate approach (16). Two-way interactions between the included variables were also tested in the full model, and were only included in the final model if there was evidence against the null hypothesis of no association with participation in 2013 at p<0.05.

The final model contained only baseline participation status (yes/no), number of activity types available in 2013, number of amenities available in 2013 (in 3 categories), sex, and an interaction between baseline participation and number of activity types available in 2013.

(iii) Model diagnostics

Having fitted a regression model, it is important to assess whether the fitted model adequately represents the data. Fagerland *et al.* examine the effectiveness of various goodness-of-fit tests for the multinomial logistic regression model (17). Blizzard *et al.* (18) extend this to the log multinomial model, assessing the performance of the goodness-of-fit tests using simulated data. They propose using an extension of the Hosmer-Lemeshow test statistic for binary logistic regression, in conjunction with the ungrouped standardised Pearson X^2 test, to assess goodness-of-fit of the log multinomial model.

Working with binary data, Hosmer and Lemeshow (19) combined the n observations into G groups based on ascending size of the estimated probabilities of the (binary) outcome. Fagerland *et al.* (17) grouped the observations in descending order based on the complement of the sum of fitted probabilities: $\hat{\mu}_{i1} = 1 - \sum_{j=2}^{J} \hat{\mu}_{ij}$. When there are tied values at the group boundaries, or n/G is not an integer, the groups will not be even sized and the value of the test statistic will depend on the way the observations are allocated in to groups. The (extended) Hosmer-Lemeshow test statistic is the Pearson X^2 statistic for the $G \times J$ table of observed and estimated frequencies:

$$X^2 = \sum_{g=1}^{G} \sum_{j=1}^{J} \frac{\left(O_{gj} - E_{gj}\right)^2}{E_{gj}} = \sum_{g=1}^{G} \sum_{j=1}^{J} \frac{\left(m_g \bar{y}_{gj} - m_g \bar{\mu}_{gj}\right)^2}{m_g \bar{\mu}_{gj}}$$

where m_g is the number of observations in each group, *g*=1,2,…*G* , with $m_g \geq 1$ and $\sum_{g=1}^{G} m_g = n$.

This statistic is expected to have an approximate chi squared distribution with (*G*-2)x(*J*-1) degrees of freedom. A disadvantage of the extended Hosmer-Lemeshow test is that the magnitude of the test is potentially dependent on the choice of grouping method. Three different methods of grouping are suggested by Blizzard *et al.*: the "cumulative target", "target" and "percentile" methods, that are implemented in SAS and Stata software (9). These can be used to cross-check that the results of the test are robust to the choice of grouping method.

In order to use the extended Hosmer-Lemeshow test to test the fit of the log multinomial model used here for the participation data, all three of the "cumulative target", "target" and "percentile" methods of grouping were used. These resulted in extended Hosmer-Lemeshow test statistics of 10.5, 10.6 and 11.3 respectively. Comparisons of these test statistics with a chi2 distribution with 16=(10-2)x(3-1) degrees of freedom results in failure to reject the null hypothesis that the model fits the data (p=0.58, 0.59, 0.75 respectively).

The standardised Pearson statistic is an alternative summary measure of goodness of fit. Osius and Rojek (20) presented asymptotic moments for a class of statistics that includes the Pearson chi-square statistic. With an asymptotic mean of $\hat{\mu} = H \times \left(J-1\right)$, the standardised Pearson statistic is

$$z = \frac{\left(X^2 - \hat{\mu}\right)}{\hat{\sigma}},$$

where the estimator of the asymptotic variance $\hat{\sigma}^2$ is evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_2, \hat{\beta}_3, \ldots \hat{\beta}_J\right)'$. The estimate is

$$\hat{\sigma}^2 = v^2\left(\hat{\boldsymbol{\beta}}\right) - Q\left(\hat{\boldsymbol{\beta}}\right)$$

where:

$$v^2\left(\hat{\boldsymbol{\beta}}\right) = 2P\left(J-1\right) + \sum_{p=1}^{P}\left[\frac{1}{m_p}\sum_{j=1}^{J}\left(\frac{1}{\hat{\mu}_{pj}}\right) - J^2 - 2\left(J-1\right)\right]$$

$$Q\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{c}'\left(\hat{\boldsymbol{\beta}}\right)\hat{\mathbf{V}}\left(\hat{\boldsymbol{\beta}}\right)\mathbf{c}\left(\hat{\boldsymbol{\beta}}\right)$$

$$\mathbf{c}\left(\hat{\boldsymbol{\beta}}\right) = \left[\sum_{p=1}^{P}\sum_{j=1}^{J}\frac{1}{\hat{\mu}_{pj}}\frac{\partial\hat{\mu}_{pj}}{\partial\hat{\alpha}_{jk}}\right]_{k=1,2,\ldots\left(J-1\right)\times\left(K+1\right)}$$

and $\hat{\mathbf{V}}\left(\hat{\boldsymbol{\beta}}\right)$ is the usual estimator of the covariance matrix of the estimated parameters.

The standardised Pearson X^2 test statistic (Z statistic) under the null hypothesis can be approximated by a standard normal distribution. The Z statistic for the log multinomial model of the participation data was -2.8, with a one-sided p-value of 0.99, which results in failure to reject the null hypothesis that the model fits the data.

References for Statistical Appendix

1. Rubin DB. Inference and missing data. Biometrika 1976;63(3):581-592.

2. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res 2013;22(3):278-95.

3. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med 2012;367(14):1355-60.

4. Agresti A. Categorical Data Analysis, third edition. Hoboken, New Jersey: John Wiley and Sons, Inc; 2013.

5. Atienza AA, King AC. Community-based health intervention trials: an overview of methodological issues. Epidemiol Rev 2002;24(1):72-9.

6. Yee JL, Niemeier D. Advantages and disadvantages: Longitudinal vs. repeated cross-section surveys. Project Battelle 1996;94:16.

7. Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions. 3 ed: Wiley; 2013.

8. Hilbe JM. Negative binomial regression: Cambridge University Press; 2011.

9. StataCorp. Stata 12 Base Reference Manual. . College Station, TX: Stata Press; 2011.

10. Hardin J, Hilbe J. Generalized Linear Models and Extensions. College Station, Texas: StataCorp LP; 2012.

11. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. American journal of epidemiology 1987;125(5):761-768.

12. Zou G. A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol 2004;159(7):702-6.

13. Knol MJ, Le Cessie S, Algra A, Vandenbroucke JP, Groenwold RHH. Overestimation of risk ratios by odds ratios in trials and cohort studies: alternatives to logistic regression. Canadian Medical Association Journal 2012;184(8):895-899.

14. Blizzard CL QS, Canary JD and Hosmer DW. Log-Link Regression Models for Ordinal Responses. Open Journal of Statistics 2013;3:16-25.

15. Blizzard L, Hosmer DW. The log multinomial regression model for nominal outcomes with more than two attributes. Biom J 2007;49(6):889-902.

16. Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health 1989;79(3):340-9.

17. Fagerland MW, Hosmer DW, Bofin AM. Multinomial goodness-of-fit tests for logistic regression models. Statistics in medicine 2008;27(21):4238-4253.

18. Blizzard L HD, Quinn S and Canary, J Goodness-of-fit tests for multinomial log-link regression models. In: Unpublished; 2014. p. 18.

19. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. Communications in Statistics-Theory and Methods 1980;9(10):1043-1069.

20. Osius G, Rojek D. Normal goodness-of-fit tests for multinomial models with large degrees of freedom. Journal of the American Statistical Association 1992;87(420):1145-1152.