Australian & New Zealand Journal of Statistics

Check for updates

*Aust. N. Z. J. Stat.* **65**(4), 2023, 287–308 doi: 10.1111/anzs.12399

# Comparisons of distributions of Australian mental health scores

D. Gunawan[1], William E. Griffiths[2*] and D. Chotikapanich[3]

*University of Wollongong, University of Melbourne and Monash University*

## Summary

Bayesian non-parametric estimates of Australian distributions of mental health scores are obtained to assess how the mental health status of the population has changed over time, and to compare the mental health status of female/male and Aboriginal/non-Aboriginal population subgroups. First-order and second-order stochastic dominance are used to compare distributions, with results presented in terms of the posterior probability of dominance and the posterior probability of no dominance. If a criterion for dominance is satisfied, then, in terms of that criterion, the mental health status of the dominant population is superior to that of the dominated population. If neither distribution is dominant, then the mental health status of neither population is superior in the same sense. Our results suggest mental health has deteriorated in recent years, that males' mental health status is better than that of females, and that non-Aboriginal health status is better than that of the Aboriginal population.

*Key words*: Aboriginal population; Bayesian non-parametric estimation; male and female populations; posterior probabilities; stochastic dominance.

## 1. Introduction

Improving the general level of health and reducing health inequality are major objectives of public policy. It is therefore important to assess whether improvements are being made over time, and in different subgroups of a population. To make such assessments, we need to sample the health status of individuals from the populations of interest, and to use those samples to make inferences about the populations. These inferences can take the form of comparing health status at different points in time or comparing the health status of different segments of the population. We focus on mental health, comparing the mental health status of the Australian population over the years 2001, 2006, 2010, 2014 and 2017, and on that for male/female and Aboriginal/non-Aboriginal population subgroups in the same years. We chose mental health status as an indicator of

*Author to whom correspondence should be addressed.

[1]School of Mathematics and Applied Statistics, University of Wollongong, Northfields Ave. Wollongong, NSW 2522, Australia.

[2]Department of Economics, University of Melbourne, Parkville VIC 3010, Australia. e-mail: wegrif@unimelb.edu.au

[3]Department of Econometrics and Business Statistics, Monash University, Caulfield East VIC 3145, Australia.

*Acknowledgements*. The authors thank three referees whose comments and suggestions have led to improvements in the paper. Open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australian University Librarians.

well-being because not having good mental health can have important negative effects on many lifestyle factors such as employment opportunities and social inclusion. Also, depression and anxiety can lead to a deterioration in other dimensions of health and overall health status (Schmits 2011). According to the World Health Organisation (WHO), many of our global health problems are attributable to mental disorders. McGorry (2005) suggests that one in two people will have at least one period of depression in their lifetime. Studies investigating the relationship between mental health outcomes and socioeconomic characteristics include Broom *et al.* (2006), Schmits (2011) and Bechtel *et al.* (2012). In recent years, the prevalence of poor mental health has attracted increasing attention, particularly in relation to difficulties resulting from COVID-19 lockdowns. Our sample does not cover the COVID period, but our results suggest mental health had already been deteriorating prior to that time. Interest has also centred on the status of female mental health relative to that of males, and a major Australian Government policy has been to narrow the gap between Aboriginal and non-Aboriginal health status (see www.closingthegap.gov.au for further details). We examine evidence on the relative health status of these population subgroups and how this evidence has changed over time. The chosen years were spaced at intervals sufficient to enable a large number of pairwise comparisons. The sample we use is the SF-36 health scores obtained from the Household, Income and Labour Dynamics in Australia (HILDA) survey (https://melbourneinstitute.unimelb.edu.au/hilda).

A major contribution of the paper is the novel criteria used to compare health scores. After estimating complete distributions of mental health scores using Bayesian non-parametric methods, these distributions are compared using stochastic dominance. In addition to using posterior densities to compare mean scores and single measures designed to capture acute mental illness, we find posterior probabilities of first (FSD)- and second-order stochastic dominance (SSD) for each pairwise comparison of distributions. Checking to see if one distribution of mental health scores dominates another involves comparing the cumulative distribution functions (CDFs) of the two sets of health scores. The CDF of the distribution gives the proportion of the population below each level of mental health. For first-order stochastic dominance (FSD), a distribution $A$ dominates a distribution $B$ (written $A_{FSD}B$) if the CDF for $A$ lies below the CDF for $B$; the proportion of population with a mental health score below any value $y$ is less in $A$ than it is in $B$. The two CDFs do not cross. For second-order stochastic dominance (SSD), $A$ dominates $B$ (written $A_{SSD}B$) if the area under the CDF between zero and any value $y$ is less for $A$, than it is for $B$. It implies the sum of all mental health scores less than any value $y$ is less for $A$ than it is for $B$. The existence of SSD does not rule out the possibility of CDFs that cross; FSD implies SSD, but the converse is not true. We present evidence on the relative standing of $A$ and $B$ in terms of three posterior probabilities: the probability $A$ dominates $B$, the probability $B$ dominates $A$ and the probability that neither distribution is dominant.

Likely to be of particular concern is the prevalence of severe mental illness—those in the left tail of the distribution. To provide evidence on this tail, we borrow concepts from poverty analysis, providing posterior information on the proportion of the population below a threshold and the severity of illness below that threshold. We also examine dominance results for those in the lower tail of the distribution.

In Section 2 we briefly review examples of studies that have converted categorical data from self-reported health surveys into a continuous variable, and then describe

the continuous SF36 mental health score used in this study. The observations on this variable are modelled non-parametrically using an infinite mixture of beta distributions. The Bayesian Markov Chain Monte Carlo (MCMC) methodology for estimating this model, and the quantities necessary for comparing the health score distributions, are described in a Supporting Information. The criteria used for comparing distributions are described in Sections 3 and 4, with Section 3 being devoted to single characteristics of the distributions, and Section 4 covering dominance criteria which involve comparing whole distributions. The mean is the single characteristic used for comparing complete distributions. For comparing the left tails of the distributions, where mental health is particularly poor, we use the headcount ratio and two values from the Foster–Greer–Thorbecke (FGT) measures introduced in Foster *et al.* (1984), concepts borrowed from poverty analysis. In Section 4, we define FSD and SSD, explain how MCMC draws are used to estimate probabilities of dominance, and describe how this approach differs from sampling theory tests for dominance. Our results are presented in Section 5. Briefly, we find poor mental health was more prevalent in 2017 than in 2001, more prevalent in females than males, and more prevalent in Aboriginal than in non-Aboriginal subgroups of the population. Some concluding remarks are provided in Section 6.

## 2. Specifying the health score distribution

Constructing a suitable distribution for mental health scores is not a simple task. It involves converting responses to questions designed to be mental health indicators into scores on a single variable, preferably a continuous one that facilitates further analysis. The scores that we utilise originate from the SF-36 survey, a multipurpose and short form health survey with 36 questions that provides one of the most widely used generic and continuous measures of health-related quality of life in clinical research and general population health (Ware *et al.* 1993). It has been translated and studied in more than 40 countries (Ware & Gandek 1998). The questions in the SF-36 survey are part of the HILDA survey which is a national representative longitudinal survey which began in Australia in 2001, and is conducted annually (Watson & Wooden 2012). It was initiated and is funded by the Australian Government Department of Social Services (DSS) and is managed by the Melbourne Institute of Applied Economics and Social Research (Melbourne Institute). Data on key variables concerning family and household structure, as well as data on education, income, health, life satisfaction and other variables relating to economic and subjective wellbeing are collected.

The questions in the SF-36 survey fall into eight categories with each category representing a different health dimension and with the responses in each category being aggregated to provide a score. We are concerned with the mental health dimension that consists of five multiple-choice questions that ask respondents about their perceptions of their mental health. The questions are (1) 'been a very nervous person', (2) 'felt down in the dumps', (3) 'felt calm and peaceful', (4) 'felt downhearted and blue' and (5) 'been a happy person'. Each respondent is asked to rate how often they felt in such a way in the past 4 weeks: (a) all of the time, (b) most of the time, (c) a good bit of the time, (d) some of the time, (e) a little of the time or (f) none of the time. The responses are scored using the scoring algorithm in Ware *et al.* (1993) and are provided as part of the HILDA data base. The final measure ranges between 0 and 100, where a score of 100

implies good mental health and a 0 represents a serious mental health problem. Individuals with scores below 50 are considered to have poor mental health. The developers of the SF-36 claim that scaling assumptions used to transform the ordered categorical responses into a continuous health measure can be interpreted as 'quasi-interval measurement scales' (Ware & Gandek 1998). They argue that such scales can consistently rank health status, and that the ratio of differences between scores has meaning. While such claims can be debatable, it is important to note that all techniques used to convert discrete category scores into a continuous variable will have some issues. Evidence provided by Butterworth & Crosier (2004) supports the validity of SF-36 data collected by the HILDA survey as general measures of physical and mental health status.

Having obtained a mental health score for everyone in the sample, we are faced with the problem of using the sample of scores to estimate a distribution. If we adopt a parametric approach, then, given that each mental health score lies in the interval [0,100], a convenient continuous distribution for representing the population of scores is the beta distribution, applied to a scaling of the scores to make them lie in the interval $[0, 1]$. Its probability density function (PDF) is given by

$$B(y|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1},$$

where $y$ is a randomly drawn mental health score and $\alpha$ and $\beta$ are parameters. However, using a parametric distribution with only two parameters is unlikely to be adequate to capture a wide variety of mental health distributions. As an alternative, whose results are likely to be less sensitive than those from specific parametric choices, we use a Bayesian non-parametric approach, modelling an infinite mixture of beta distributions via a Dirichlet process prior (Escobar & West 1995). In this context, the pdf for $y$ is given by

$$p(y|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{w}) = \sum_{k=1}^{\infty} w_k B(y|\alpha_k,\beta_k) \tag{1}$$

with parameter vectors $\boldsymbol{\alpha}^\top = (\alpha_1, \alpha_2, \dots)$, $\boldsymbol{\beta}^\top = (\beta_1, \beta_2, \dots)$ and $\boldsymbol{w}^\top = (w_1, w_2, \dots)$, where the $w_k$ represent the weights attached to each component of the mixture.

Distributions are estimated for each of the 5 years and for each of the population subgroups in those years, treating each set of observations as a cross-section, estimated with cross-sectional weights. These cross-sectional weights are provided by the HILDA dataset and used to correct the representativeness of the HILDA sample to the population. Details of an MCMC algorithm for estimating this model, along with results from convergence diagnostics for the MCMC draws, are provided in the Supporting Information. Draws from the posterior density of the parameters are taken at each iteration of the MCMC algorithm. Functions of these parameter draws are then used to estimate criteria for comparing distributions. The apparent need to sample an infinite number of parameters in (1) is avoided by using a device known as the slice sampler (Walker 2007). At each of $j = 1, 2, \dots, M$ iterations of the MCMC algorithm, this device stochastically truncates the infinite number of components in (1) to a finite number $K^{(j)} + 1$. Posterior sampling for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{w}$ proceeds using this finite number of elements; the draws $\left(\boldsymbol{\alpha}^{(j)}, \boldsymbol{\beta}^{(j)}, \boldsymbol{w}^{(j)}\right)$ are of dimension $K^{(j)} + 1$, with this dimension changing

with each iteration. One 'draw' from the predictive-posterior PDF for $y$ can be written as

$$p^{(j)}\left(y|\boldsymbol{\alpha}^{(j)},\boldsymbol{\beta}^{(j)},\boldsymbol{w}^{(j)}\right)=\sum_{k=1}^{K^{(j)}+1}w_k^{(j)}B\left(y|\alpha_k^{(j)},\beta_k^{(j)}\right). \tag{2}$$

The Bayesian non-parametric density estimate for $y$ is given by the average of (2) over all iterations,

$$\widehat{p}(y)=\frac{1}{M}\sum_{j=1}^{M}p^{(j)}\left(y|\boldsymbol{\alpha}^{(j)},\boldsymbol{\beta}^{(j)},\boldsymbol{w}^{(j)}\right). \tag{3}$$

For a given value of $y$, the spread of the values in (2) provides an indication of the reliability of (3) as an estimate of $p(y)$. The PDF in (3) is not used directly for comparing the health status of populations, but the quantities based on it are. Also, graphing (3), as we do in our empirical work, is useful for gaining an appreciation of the nature of the distribution. In the next section we describe the criteria we employ for health-status comparisons, except for dominance, which is deferred to Section 4.

## 3. Criteria for comparing distributions

Values of quantities used as criteria for comparing health status at different points in time and across different subgroups of population are computed from the MCMC output $\left(\boldsymbol{\alpha}^{(j)},\boldsymbol{\beta}^{(j)},\boldsymbol{w}^{(j)}\right)$, $j=1,2,\ldots,M$. The first of these is the mean health score; observations on it drawn from its posterior PDF are given by

$$\mu^{(j)}=\sum_{k=1}^{K^{(j)}+1}w_k^{(j)}m_k^{(j)}=\sum_{k=1}^{K^{(j)}+1}\frac{w_k^{(j)}\alpha_k^{(j)}}{\alpha_k^{(j)}+\beta_k^{(j)}},$$

where $m_k=\alpha_k/(\alpha_k+\beta_k)$ is the mean of the $k$th component of the mixture. The average of the MCMC draws, $\widehat{\mu}=\sum_{j=1}^{M}\mu^{(j)}/M$, is an estimate of the posterior mean which in turn is an estimate for $\mu$. The standard deviation of the $\mu^{(j)}$ is an estimate of the posterior standard deviation of $\mu$ and is an indication of the reliability of the estimate $\widehat{\mu}$.

In a similar way, we can use averages and standard deviations of MCMC draws to estimate posterior means and standard deviations for other quantities of interest. In line with the MCMC algorithm, infinite sums in these quantities are truncated to $K^{(j)}+1$ in the $j$th iteration of the algorithm. For comparing health status for those with acute mental illness, we consider properties of the distribution below a pre-specified threshold $z$. The proportion of the population below this threshold, known as the headcount ratio in the poverty literature, is given by

$$\begin{aligned}HC&=\int_0^z p(y|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{w})\,dy\\&=F(z|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{w})\\&=\sum_{k=1}^{\infty}w_kF(z|\alpha_k,\beta_k),\end{aligned}$$

where $F(z|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w})$ is the CDF for $y$ evaluated at $z$ and $F(y|\alpha_k, \beta_k) = \int_0^y B(t|\alpha_k, \beta_k)\, dt$ is the CDF of the $k$th component.

A family of measures that considers not just the number of individuals below a threshold, but also the severity of the mental illness for those below the threshold, is that attributable to Foster *et al.* (1984). It is given by

$$FGT_a = \int_0^z \left(\frac{z-y}{z}\right)^a p(y|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w})\, dy.$$

It can be interpreted as a weighted average of the discrepancy between each mental health score and the threshold, expressed relative to the threshold. The larger the value of $a$, the larger is the weight placed on scores which are further from the threshold. In our empirical work, we compare $FGT$ values for $a = 1$ and $a = 2$. Details of how to compute them for the beta mixture model are provided in the Supporting Information.

## 4. Assessing stochastic dominance

There are several ways that one can approach the question: has mental health of a population improved over time? Or, alternatively: is the mental health of one population subgroup better or worse than another? A simple way is to compare measures of central tendency such as the mean, median or mode, enabling one to say the 'average' or 'typical' level of mental health is better for one case than another. Using the posterior distribution for the mean mental health score, introduced in the previous section, is one example of this approach. Another more complete and more exacting way is to use a metric that compares whole distributions of mental health scores. Stochastic dominance concepts are useful for making such comparisons. Suppose we are comparing two distributions of mental health scores that we call $A$ and $B$, with corresponding beta-mixture CDFs, $F_A\left(y|\boldsymbol{\alpha}_A, \boldsymbol{\beta}_A, \boldsymbol{w}_A\right)$ and $F_B\left(y|\boldsymbol{\alpha}_B, \boldsymbol{\beta}_B, \boldsymbol{w}_B\right)$. We say that $A$ first-order stochastically dominates $B$ if and only if

$$F_A\left(y|\boldsymbol{\alpha}_A, \boldsymbol{\beta}_A, \boldsymbol{w}_A\right) \leq F_B\left(y|\boldsymbol{\alpha}_B, \boldsymbol{\beta}_B, \boldsymbol{w}_B\right) \tag{4}$$

for all mental health scores $0 \leq y \leq 1$, with strict inequality holding for some $0 < y < 1$. Using a simplified description of (4) that ignores equalities and the zero-one end points, $A_{FSD}B$ implies that the proportion of people with a health score less than some value $y$ is less in population $A$ than it is in population $B$; and this is true for all $0 < y < 1$. Except at the endpoints, the CDF for $A$ lies to the right of the CDF for $B$.

We say that $A$ second-order stochastically dominates $B$ if and only if

$$\int_0^y F_A\left(t|\boldsymbol{\alpha}_A, \boldsymbol{\beta}_A, \boldsymbol{w}_A\right) dt \leq \int_0^y F_B\left(t|\boldsymbol{\alpha}_B, \boldsymbol{\beta}_B, \boldsymbol{w}_B\right) dt \tag{5}$$

for all $0 \leq y \leq 1$, with strict inequality holding for some $0 < y < 1$. An intuitive interpretation of (5) is that the total mental health score for everybody with a score less than or equal to any mental health score $y$ is less in population $A$ than it is in population $B$. The frequencies of some mental scores that are less than $y$ can be greater in $A$ than in $B$, but the frequencies of scores that are lower in $A$ than in $B$ are always sufficient to compensate. This result holds for all $y$. This condition is less strict than FSD; $A_{FSD}B$ implies $A_{SSD}B$. It is possible for SSD to exist when the two CDFs cross. For our analysis

of SSD, it turns out that there is a more convenient expression than that in (5). Using the first-moment distribution function $F^{(1)}(y|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{w}) = (1/\mu)\int_0^y t\,p(t|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{w})dt$, we can show that $A_{SSD}B$ if and only if

$$yF_A\left(y|\boldsymbol{\alpha}_A,\boldsymbol{\beta}_A,\boldsymbol{w}_A\right)-\mu_A F_A^{(1)}\left(y|\boldsymbol{\alpha}_A,\boldsymbol{\beta}_A,\boldsymbol{w}_A\right) \leq yF_B\left(y|\boldsymbol{\alpha}_B,\boldsymbol{\beta}_B,\boldsymbol{w}_B\right) - \mu_B F_B^{(1)}\left(y|\boldsymbol{\alpha}_B,\boldsymbol{\beta}_B,\boldsymbol{w}_B\right) \tag{6}$$

for all $0 \leq y \leq 1$ and with strict inequality holding for at least some $0 < y < 1$. A proof of this result is provided in the Supporting Information. Inequalities (4) and (6) are the ones used to compare posterior probabilities of dominance, along the lines that we now discuss.

To assess whether conditions (4) and/or (6) hold for two specified populations, the distribution functions $F$, and the first-moment distribution functions $F^{(1)}$ need to be estimated using samples from each of the populations. Then, recognising the existence of sampling error, we need a way of presenting the degree to which the sample information supports the existence of dominance in either direction, or the existence of no dominance. To make this inference problem specific, let

$$D_1(y) = F_B(y) - F_A(y) \tag{7}$$

and

$$D_2(y) = \left[yF_B(y) - \mu_B F_B^{(1)}(y)\right] - \left[yF_A(y) - \mu_A F_A^{(1)}(y)\right], \tag{8}$$

where $\left[F_A(y), F_A^{(1)}(y), F_B(y), F_B^{(1)}(y)\right]$ is abbreviated notation for the functions in (4) and (6). Let $\widehat{D}_1(y)$ and $\widehat{D}_2(y)$ denote estimates of the functions $D_1(y)$ and $D_2(y)$, respectively, obtained by replacing the quantities in (7) and (8) by their estimates, $\left[\widehat{\mu}_A, \widehat{F}_A(y), \widehat{F}_A^{(1)}(y), \widehat{\mu}_B, \widehat{F}_B(y), \widehat{F}_B^{(1)}(y)\right]$. For $A$ to dominate $B$ we require $D_i(y) \geq 0$ for all $0 \leq y \leq 1$ and $D_i(y) > 0$ for some $0 < y < 1$, where $i = 1$ for FSD, and $i = 2$ for SSD. For $B$ to dominate $A$, the inequalities for $D_i(y)$ are reversed.

A vast number of sampling theory tests that typically use a non-parametric estimate $\widehat{D}_i(y)$ to test for dominance have appeared in the literature. This literature can be accessed through the extensive list of references in Lander *et al.* (2020). Special problems arise because $D_i(y)$ is a function and not a point. Some tests resolve this issue by estimating $D_i(y)$ at a number of points and using its maximum value as a test statistic. Others base a test statistic on the joint distribution of $D_i(y)$ evaluated at a number of points. Also relevant is whether dominance or no dominance is chosen as the null hypothesis; this choice has a bearing on what conclusions are possible. As described by Davidson & Duclos (2013), if dominance is specified as the null hypothesis, dominance of one distribution over another can never be established. Failure to reject a null that $A$ dominates $B$ can mean that (a) $A$ dominates $B$, (b) neither distribution is dominant, or (c) that $A$ does not dominate $B$, but the magnitude of the test statistic was not sufficiently large to conclude it was 'significant'. Reversal of the distributions so that $B$ dominates $A$ is the null hypothesis does little to resolve this dilemma. If both $A$ dominating $B$ and $B$ dominating $A$ are rejected, one can conclude with some confidence that neither distribution is dominant. However, other combinations of outcomes do not lead to a firm conclusion. Changing the null hypothesis to, say, $A$ does not dominate $B$ has the advantage that it leads to a legitimate claim that $A$ dominates $B$ if the null hypothesis is rejected. However, as pointed out by Davidson &

Duclos (2013), for continuous distributions such a null hypothesis can never be rejected unless the range of the variable being considered is restricted. The problem lies in the tails where the distributions converge. In our case, for FSD, we have $D_1(0) = D_1(1) = 0$, and, for SSD, we have $D_2(0) = 0$ and $D_2(1) = \mu_A - \mu_B$; the result $D_2(1) = \mu_A - \mu_B$ implies $\mu_A \geq \mu_B$ is a necessary condition for SSD.

In contrast to the sampling theory approach, we do not use a formal hypothesis testing framework for summarising the sample information about dominance. Instead, following our earlier work (Gunawan *et al.* 2020; Lander *et al.* 2020), we avoid the need to specify a null hypothesis by computing the posterior probabilities for each possible outcome: $A$ dominates $B$, $B$ dominates $A$, and neither distribution is dominant. Further discussion of the difficulties associated with the sampling theory approach, and a comparison of some Bayesian and sampling theory results can be found in our earlier work. To compute dominance probabilities, we begin with the MCMC draws for $\boldsymbol{\theta}^\top = \left(\boldsymbol{\alpha}_A^\top, \boldsymbol{\beta}_A^\top, \boldsymbol{w}_A^\top, \boldsymbol{\alpha}_B^\top, \boldsymbol{\beta}_B^\top, \boldsymbol{w}_B^\top\right)$ from their posterior distributions. For each value of $y$ in the interval [0,1], there will be posterior distributions for $D_1(y)$ and $D_2(y)$ implied by the posterior distributions for the elements in $\boldsymbol{\theta}$. For each MCMC draw for $\boldsymbol{\theta}$ there are corresponding draws for $D_1(y)$ and $D_2(y)$ for every value of $y$. Let $D_i^{(j)}(y)$, $i = 1, 2$, be the $j$th MCMC draw on $D_i(y)$. Then, an estimate of the probability that $D_i(y)$ is non-negative at the point $y$ is equal to the proportion of draws for which $D_i^{(j)}(y) \geq 0$. That is,

$$\Pr\left[D_i(y) \geq 0\right] = \frac{1}{M}\sum_{j=1}^{M} \mathbf{1}\left(D_i^{(j)}(y) \geq 0\right), \tag{9}$$

where $\mathbf{1}(\cdot)$ is an indicator function equal to one if its argument is true and zero otherwise. To evaluate the probability of dominance, we need the proportion of draws for which $D_i^{(j)}(y) \geq 0$ for all $y$, and with $D_i^{(j)}(y) > 0$ for at least one $y$. Because $y$ is continuous, the best we can do in terms of evaluating $D_i^{(j)}(y)$ at all values of $y$, is to compute it for a fine grid of values in the interval $0 \leq y \leq 1$, say $y_1, y_2, \ldots, y_H$. We used 1000 equally spaced values from 0.1 to 0.999; there were very few values less than 0.1. Except at, or close to, the endpoints, it is unlikely that computed values of $D_i^{(j)}(y)$ will be exactly zero, and so, whether we treat $D_i^{(j)}(y) \geq 0$ as a strict or non-strict inequality is immaterial. Given this framework, estimates of the posterior probabilities can be specified as

$$\Pr\left(A \text{ dominates } B\right) = \frac{1}{M}\sum_{j=1}^{M}\prod_{h=1}^{H} \mathbf{1}\left(D_i^{(j)}\left(y_h\right) \geq 0\right),$$

$$\Pr\left(B \text{ dominates } A\right) = \frac{1}{M}\sum_{j=1}^{M}\prod_{h=1}^{H} \mathbf{1}\left(D_i^{(j)}\left(y_h\right) \leq 0\right).$$

$$\Pr\left(\text{neither distribution dominates}\right) = 1 - \Pr\left(A \text{ dominates } B\right) - \Pr\left(B \text{ dominates } A\right).$$

As one might expect, the results can be sensitive to the endpoints selected for the grid of $y$ values. At 0 and 1 for FSD and 0 for SSD, the variances of the posterior distributions collapse to zero. Close to these points there will be only a small variation which can

impact on the dominance probabilities. This impact can be monitored by using (9) to plot $\Pr\left[D_i(y) \geq 0\right]$ against $y$, yielding curves that we call 'probability curves'. Because

$$\frac{1}{M}\sum_{j=1}^{M}\prod_{h=1}^{H}\mathbf{1}\left(D_i^{(j)}(y_h) \geq 0\right) \leq \min_{y_h}\left\{\frac{1}{M}\sum_{j=1}^{M}\mathbf{1}\left(D_i^{(j)}(y_h) \geq 0\right)\right\}$$

these curves provide a convenient device for checking the values of $y$ having the greatest effect on the probability of dominance. Excluding values of $y$ close to the endpoints is akin to Davidson & Duclos (2013) notion of restricted dominance. It may be less than ideal, but we can always document the range being considered, choose ranges that are of particular interest, and examine sensitivity to changes in the range.

Earlier studies used mixtures of gamma densities to compute posterior probabilities of dominance for income distributions, a finite mixture in Lander *et al.* (2020) and an infinite mixture in Gunawan *et al.* (2020). In these cases, to avoid subjectively specifying a maximum income for the grid of income values, they expressed the conditions for FSD and SSD in terms of quantile functions and checked for dominance over a range of population proportions that must lie between zero and one. For our health score distributions where the mental health score lies in the [0,1] interval, it is natural to compute dominance probabilities using FSD and SSD conditions expressed in terms of their distribution functions. In addition, there is an extensive literature on non-parametric Bayesian inference using Dirichlet process priors subject to a partial stochastic ordering (see Hoff 2003, Gelfand & Kottas 2001, Dunson & Peddada 2008, and Hwang & Chen 2015). These approaches, where stochastic order restrictions are built into the prior, differ from our approach where no a priori assumptions about an ordering are made.

## 5. Results

We first examine how the distribution of mental health scores has changed over time for the complete population, and then compare distributions for the sub-populations Aboriginal/non-Aboriginal and male/female. To conserve space, we summarise the results of the male/female comparison in the paper and provide a complete set of results for this comparison in the Supporting Information.

### 5.1. Comparing distributions over time

Summary statistics from the raw data, describing the characteristics of the samples in each of the years 2001, 2006, 2010, 2014 and 2017, are reported in Table 1. Estimates of the health score densities for the years 2010 and 2017, obtained by averaging the functions at each of the MCMC draws of the parameters (see equation (3)), are plotted in Figure 1. Two years were chosen as examples because it is hard to distinguish the separate densities in a figure with all 5 years. The figure with all 5 years is provided in the Supporting Information. All densities are negatively skewed and bimodal, with one mode lying between 0.8 and 0.9 and a secondary mode at the maximum value of 0.999. To examine changes over time in the mean scores, the headcounts and the FGT indices, we consider their posterior means and standard deviations reported in Table 2. The year-to-year changes in these values are not large, but they show a consistent pattern: an improvement

Table 1. Summary statistics for mental health scores.

| | 2001 | 2006 | 2010 | 2014 | 2017 |
|---|---|---|---|---|---|
| Sample mean | 0.7351 | 0.7407 | 0.7410 | 0.7373 | 0.7291 |
| Minimum | 0.0400 | 0.0400 | 0.0400 | 0.0400 | 0.0400 |
| Maximum | 0.9990 | 0.9990 | 0.9990 | 0.9990 | 0.9990 |
| Standard deviation | 0.1748 | 0.1717 | 0.1679 | 0.1739 | 0.1789 |
| Sample size | 12873 | 11545 | 11911 | 15387 | 15781 |
| Headcount | 0.1049 | 0.0978 | 0.0949 | 0.1051 | 0.1147 |
| $FGT_1$ | 0.0261 | 0.0235 | 0.0217 | 0.0262 | 0.0293 |
| $FGT_2$ | 0.0111 | 0.0098 | 0.0086 | 0.0110 | 0.0125 |
| Proportion below 0.1 | 0.0021 | 0.0017 | 0.0013 | 0.0021 | 0.0020 |

*Note*s: These values were computed from the raw data, weighted with the weights provided by HILDA. Health scores at 1.000 were set to 0.999 to avoid instability in the estimation process. The percentages of observations changed from 1.000 to 0.999 were 3.2% in 2001, 2.7% in 2006, 2.3% in 2010, 2.1% in 2014 and 2.5% in 2017. For the headcounts and *FGT* indices, poor mental health was defined as those scores less than 0.5, a threshold suggested by the SF-36 survey.

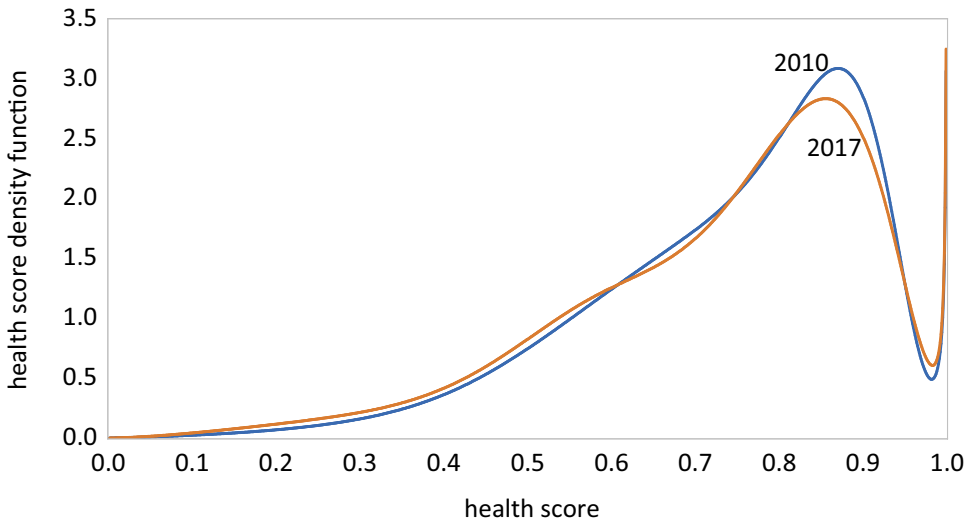


Figure 1. MCMC estimated health score density functions for the example years 2010 and 2017. The density functions for all years have similar shapes, with two modes, one between 0.8 and 0.9 and the other at 1.

in mental health from 2001 to 2010, followed by a decline. The posterior distributions for the mean scores (Figure 2) display considerable overlap, but there is a definite decline from 2010 to 2017.

A more complete picture of whether there has been an improvement in the distribution of mental health scores is obtained by comparing cumulative distribution functions and computing their dominance probabilities. The CDFs for 2010 and 2017 are plotted in Figure 3; a figure with the CDFs for all 5 years is given in the Supporting Information. From the CDFs and the FSD probabilities reported in Table 3, there is no clear ranking between all the years. Except for 2017 versus 2006 where the probability of no dominance is 0.69, all probabilities for no dominance are greater than 0.84, and most are greater

Table 2. Posterior means (standard deviations) for mental health score measures.

| | 2001 | 2006 | 2010 | 2014 | 2017 |
|---|---|---|---|---|---|
| Means | 0.7346 | 0.7406 | 0.7407 | 0.7368 | 0.7287 |
| | (0.0028) | (0.0029) | (0.0028) | (0.0028) | (0.0028) |
| Headcount | 0.1075 | 0.1002 | 0.0962 | 0.1066 | 0.1173 |
| | (0.0047) | (0.0049) | (0.0044) | (0.0042) | (0.0046) |
| $FGT_1$ | 0.0263 | 0.0236 | 0.0216 | 0.0259 | 0.0293 |
| | (0.0016) | (0.0017) | (0.0015) | (0.0015) | (0.0015) |
| $FGT_2$ | 0.0112 | 0.0099 | 0.0085 | 0.0109 | 0.0125 |
| | (0.0009) | (0.0009) | (0.0008) | (0.0008) | (0.0009) |

*Notes*: These values are calculated from the MCMC draws from the parameters of the beta mixture model. They match closely those calculated from the raw data, reported in Table 1. For the headcounts and *FGT* indices, poor mental health was defined as those scores less than 0.5, a threshold suggested by the SF-36 survey.

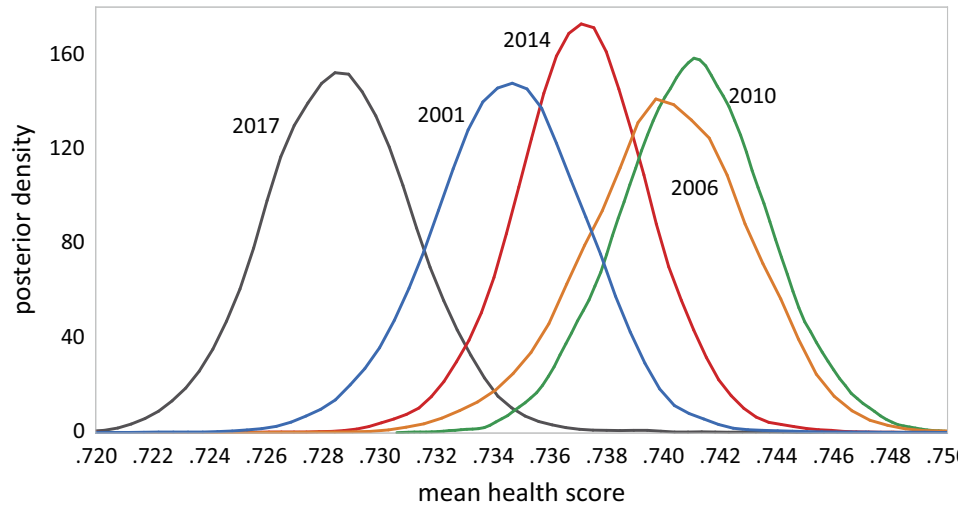


Figure 2. Posterior density functions for mean health scores computed from the MCMC draws on the beta-mixture parameters.

than 0.9. To gain more insights into these results, it is useful to examine a sample of the dominance probability curves. When just the mean scores were examined, we found 2006 and 2010 had the highest score—they are approximately the same—and 2017 has the lowest. The probabilities for the years with the highest mean scores dominating that with the lowest mean score are $\Pr(2006_{FSD}2017) = 0.308$ and $\Pr(2010_{FSD}2017) = 0.062$. The behaviour of the dominance probability curves, depicted in Figure 4, shows why these two probabilities differ and why they are relatively small. Both curves suggest that, for scores in the interval $0.3 \leq y \leq 0.7$, the probability of dominance would be close to 1. Including the tails leads to a reduction in these probabilities, with the reduction attributable to the right tail being particularly dramatic for $2010_{FSD}2017$. Also included in Figure 4, with scale on the right axis, are the posterior means for $D_1^{(a)}(y) = F_{2017}(y) - F_{2006}(y)$ and $D_1^{(b)}(y) = F_{2017}(y) - F_{2010}(y)$. As expected, the larger values for $\Pr\left[D_1(y) \geq 0\right]$ correspond to the largest values for $F_{2017}(y) - F_{2006}(y)$. In Figure 4b where the curves

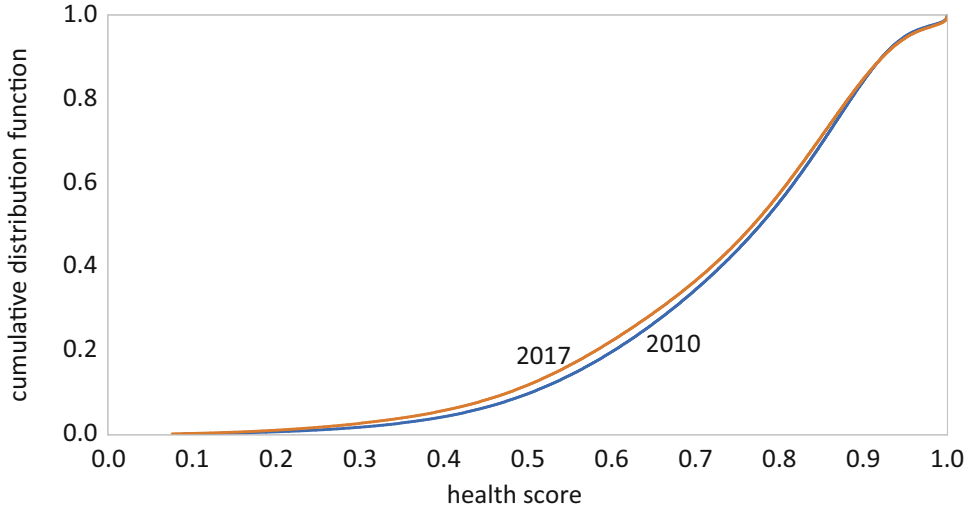


Figure 3. MCMC estimated health score cumulative distribution functions for the example years 2010 and 2017.

Table 3. First-order stochastic dominance probabilities comparing mental health score distributions for each pair of the years 2001, 2006, 2010, 2014 and 2017.

| | *A* 2017 | *B* 2014 | *A* 2014 | *B* 2010 | *A* 2010 | *B* 2006 | *A* 2006 | *B* 2001 |
|---|---|---|---|---|---|---|---|---|
| $\Pr(A_{FSD}B)$ | 0.0001 | | 0.0005 | | 0.0022 | | 0.0361 | |
| $\Pr(B_{FSD}A)$ | 0.0256 | | 0.0874 | | 0.0094 | | 0.0003 | |
| Pr (no dominance) | 0.9743 | | 0.9121 | | 0.9884 | | 0.9636 | |
| | | | 2017 | 2010 | 2014 | 2006 | 2010 | 2001 |
| $\Pr(A_{FSD}B)$ | | | 0.0000 | | 0.0004 | | 0.0011 | |
| $\Pr(B_{FSD}A)$ | | | 0.0624 | | 0.1033 | | 0.0000 | |
| Pr (no dominance) | | | 0.9376 | | 0.8963 | | 0.9989 | |
| | | | | | 2017 | 2006 | 2014 | 2001 |
| $\Pr(A_{FSD}B)$ | | | | | 0.0000 | | 0.0006 | |
| $\Pr(B_{FSD}A)$ | | | | | 0.3083 | | 0.0041 | |
| Pr (no dominance) | | | | | 0.6917 | | 0.9953 | |
| | | | | | | | 2017 | 2001 |
| $\Pr(A_{FSD}B)$ | | | | | | | 0.0001 | |
| $\Pr(B_{FSD}A)$ | | | | | | | 0.1505 | |
| Pr (no dominance) | | | | | | | 0.8494 | |

*Notes*: In each pairwise comparison *A* refers to the later year and *B* refers to the earlier year.

for $2010_{FSD}2017$ are plotted, the estimated CDFs cross leading to some negative values for $D_1(y)$ and a very low probability of dominance.

When we move to consider the less strict criterion of SSD, we find more evidence of dominance of some years over the others. From Table 4, the probabilities for 2001 being dominated by 2006 and 2010 are 0.548 and 0.809 respectively; these represent large increases from the FSD values of 0.036 and 0.001. Comparing 2006 and 2010 with 2017,

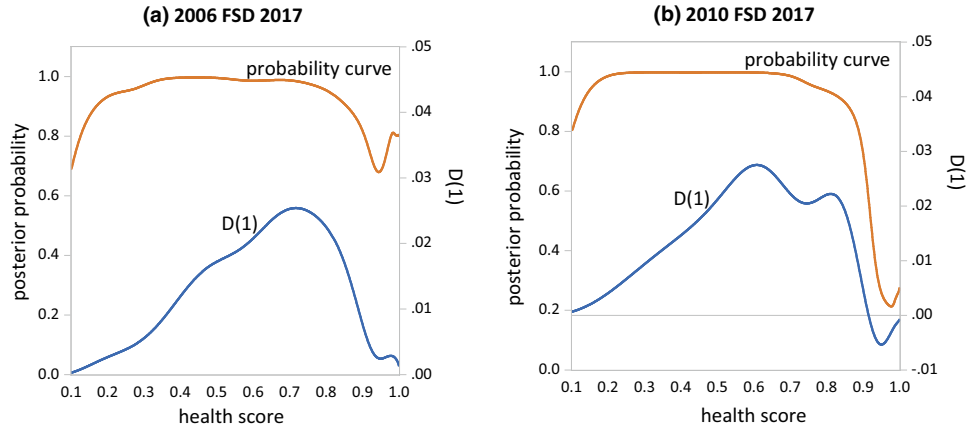


Figure 4. First-order stochastic dominance probability curves $\Pr\left[D_1(y) \geq 0\right]$ (see equation (9)) and estimated cumulative distribution function differences $\widehat{D}_1(y) = \widehat{F}_B(y) - \widehat{F}_A(y)$ for the mental health score distributions in 2006 and 2010 dominating that in 2017. The probability curves are plotted on the left axes and the CDF differences are plotted on the right axes. The probability that one distribution FSD another is less than or equal to the minimum of the probability curve. In panel (b), some negative values for $\widehat{D}_1(y)$ lead to a low FSD probability.

Table 4. Second-order stochastic dominance probabilities comparing mental health score distributions for each pair of the years 2001, 2006, 2010, 2014 and 2017.

| | $A$ 2017 | $B$ 2014 | $A$ 2014 | $B$ 2010 | $A$ 2010 | $B$ 2006 | $A$ 2006 | $B$ 2001 |
|---|---|---|---|---|---|---|---|---|
| $\Pr(A_{FSD}B)$ | 0.0024 | | 0.0029 | | 0.2805 | | 0.5479 | |
| $\Pr(B_{FSD}A)$ | 0.4797 | | 0.6730 | | 0.0501 | | 0.0087 | |
| Pr (no dominance) | 0.5179 | | 0.3241 | | 0.6694 | | 0.4434 | |
| | | | 2017 | 2010 | 2014 | 2006 | 2010 | 2001 |
| $\Pr(A_{FSD}B)$ | | | 0.0004 | | 0.0302 | | 0.8094 | |
| $\Pr(B_{FSD}A)$ | | | 0.7985 | | 0.4093 | | 0.0009 | |
| Pr (no dominance) | | | 0.2011 | | 0.5605 | | 0.1897 | |
| | | | | | 2017 | 2006 | 2014 | 2001 |
| $\Pr(A_{FSD}B)$ | | | | | 0.0003 | | 0.2971 | |
| $\Pr(B_{FSD}A)$ | | | | | 0.6744 | | 0.0559 | |
| Pr (no dominance) | | | | | 0.3253 | | 0.6470 | |
| | | | | | | | 2017 | 2001 |
| $\Pr(A_{FSD}B)$ | | | | | | | 0.0084 | |
| $\Pr(B_{FSD}A)$ | | | | | | | 0.3740 | |
| Pr (no dominance) | | | | | | | 0.6176 | |

*Notes*: In each pairwise comparison $A$ refers to the later year and $B$ refers to the earlier year.

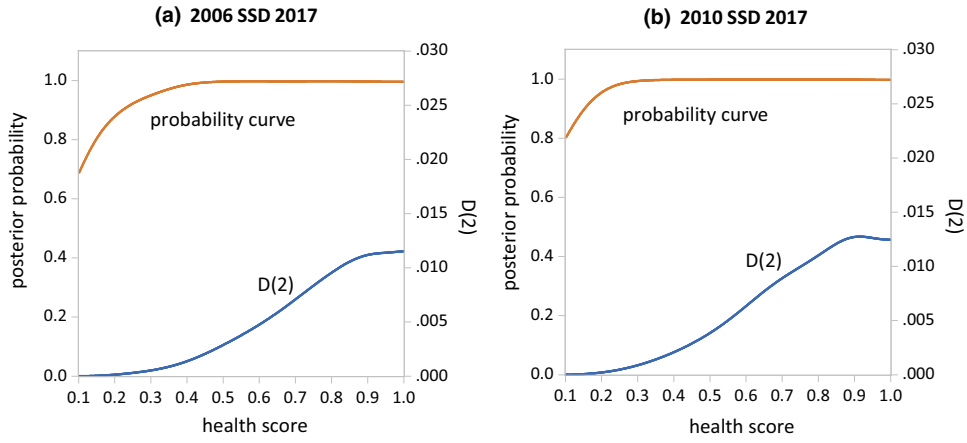


Figure 5. Second-order stochastic dominance probability curves $\Pr\left[D_2(y) \geq 0\right]$ and the differences $\widehat{D}_2(y) = \left[y\widehat{F}_B(y) - \widehat{\mu}_B\widehat{F}_B^{(1)}(y)\right] - \left[y\widehat{F}_A(y) - \widehat{\mu}_A\widehat{F}_A^{(1)}(y)\right]$ for the mental health score distributions in 2006 and 2010 dominating that in 2017. The probability curves are plotted on the left axes and $\widehat{D}_2(y)$, the differences between integrals of the cumulative distribution functions, are plotted on the right axes. The probability that one distribution SSD another is less than or equal to the minimum of the probability curve. In both cases it is the left tails of the distributions that have the greatest impact on the probability of dominance.

we find the probability that 2006 dominates 2017 has increased from 0.308 to 0.674, while that for 2010 over 2017 has increased from 0.062 to 0.799. These increases can be attributed to behaviour in the upper tails. As illustrated in Figure 5, the probability curves remain at one, in the upper tail, and the values for $D_2(y)$ remain large, unlike those for FSD.

Summarising the FSD and SSD information, we conclude that there is little evidence to suggest that complete health score distribution has shifted to the right or left over time. However, in terms of the accumulated health score for every segment of the population, there is relatively strong evidence to suggest improvement from 2001 to 2010 and a decline thereafter.

Also of interest are the dominance results for those with poor mental health. The single index measures of headcount, $FGT_1$ and $FGT_2$ reported in Tables 1 and 2 for scores less than 0.5 suggest an improvement from 2001 to 2010 and a decline thereafter, results that are in line with the SSD results for the whole population. The dominance results in Tables 5 and 6 support this conclusion. We have $\Pr(2010_{FSD}2001) = 0.805$, $\Pr(2010_{SSD}2001) = 0.838$, $\Pr(2010_{FSD}2017) = 0.795$, and $\Pr(2010_{SSD}2017) = 0.800$. There is less evidence of dominance when more adjacent years are considered.

### 5.2. Aboriginal and non-Aboriginal subgroups

Dominance criteria can also be used to track the welfare over time for subgroups of the population and to compare subgroups at a particular point in time. In this section we compare the distributions of mental health scores of Aboriginal and non-Aboriginal subgroups in each of the years 2001, 2006, 2010, 2014 and 2017, as well as examine how the distributions for each subgroup have changed over time. Summary statistics describing

Table 5. First-order stochastic dominance probabilities comparing mental health score distributions for scores below 0.5 for each pair of the years 2001, 2006, 2010, 2014 and 2017.

| | *A* 2017 | *B* 2014 | *A* 2014 | *B* 2010 | *A* 2010 | *B* 2006 | *A* 2006 | *B* 2001 |
|---|---|---|---|---|---|---|---|---|
| $\Pr(A_{FSD}B)$ | 0.0106 | | 0.0021 | | 0.3114 | | 0.4746 | |
| $\Pr(B_{FSD}A)$ | 0.4482 | | 0.7371 | | 0.0332 | | 0.0137 | |
| Pr (no dominance) | 0.5412 | | 0.2608 | | 0.6554 | | 0.5117 | |
| | | | 2017 | 2010 | 2014 | 2006 | 2010 | 2001 |
| $\Pr(A_{FSD}B)$ | | | 0.0003 | | 0.0197 | | 0.8052 | |
| $\Pr(B_{FSD}A)$ | | | 0.7954 | | 0.4214 | | 0.0011 | |
| Pr (no dominance) | | | 0.2043 | | 0.5589 | | 0.1937 | |
| | | | | | 2017 | 2006 | 2014 | 2001 |
| $\Pr(A_{FSD}B)$ | | | | | 0.0003 | | 0.2157 | |
| $\Pr(B_{FSD}A)$ | | | | | 0.6595 | | 0.1129 | |
| Pr (no dominance) | | | | | 0.3402 | | 0.6714 | |
| | | | | | | | 2017 | 2001 |
| $\Pr(A_{FSD}B)$ | | | | | | | 0.0098 | |
| $\Pr(B_{FSD}A)$ | | | | | | | 0.3609 | |
| Pr (no dominance) | | | | | | | 0.6293 | |

*Notes*: In each pairwise comparison *A* refers to the later year and *B* refers to the earlier year.

Table 6. Second-order stochastic dominance probabilities comparing mental health score distributions for scores below 0.5 for each pair of the years 2001, 2006, 2010, 2014 and 2017.

| | *A* 2017 | *B* 2014 | *A* 2014 | *B* 2010 | *A* 2010 | *B* 2006 | *A* 2006 | *B* 2001 |
|---|---|---|---|---|---|---|---|---|
| $\Pr(A_{FSD}B)$ | 0.0440 | | 0.0058 | | 0.4674 | | 0.6029 | |
| $\Pr(B_{FSD}A)$ | 0.4821 | | 0.7760 | | 0.0767 | | 0.0405 | |
| Pr (no dominance) | 0.4739 | | 0.2182 | | 0.4559 | | 0.3566 | |
| | | | 2017 | 2010 | 2014 | 2006 | 2010 | 2001 |
| $\Pr(A_{FSD}B)$ | | | 0.0009 | | 0.0578 | | 0.8382 | |
| $\Pr(B_{FSD}A)$ | | | 0.7999 | | 0.5170 | | 0.0038 | |
| Pr (no dominance) | | | 0.1992 | | 0.4252 | | 0.1580 | |
| | | | | | 2017 | 2006 | 2014 | 2001 |
| $\Pr(A_{FSD}B)$ | | | | | 0.0022 | | 0.3603 | |
| $\Pr(B_{FSD}A)$ | | | | | 0.6773 | | 0.1787 | |
| Pr (no dominance) | | | | | 0.3205 | | 0.4610 | |
| | | | | | | | 2017 | 2001 |
| $\Pr(A_{FSD}B)$ | | | | | | | 0.0466 | |
| $\Pr(B_{FSD}A)$ | | | | | | | 0.3900 | |
| Pr (no dominance) | | | | | | | 0.5634 | |

*Notes*: In each pairwise comparison *A* refers to the later year and *B* refers to the earlier year.

Table 7. Posterior means (standard deviations) for mental health score measures for Aboriginal and non-Aboriginal subgroups.

| | Mean | Headcount | $FGT_1$ | $FGT_2$ |
|---|---|---|---|---|
| 2001 | | | | |
| Aboriginal | 0.7048 | 0.1416 | 0.0348 | 0.0150 |
| | (0.0208) | (0.0352) | (0.0121) | (0.0073) |
| Non-Aboriginal | 0.7388 | 0.1016 | 0.0253 | 0.0109 |
| | (0.0032) | (0.0051) | (0.0017) | (0.0011) |
| 2006 | | | | |
| Aboriginal | 0.6989 | 0.1628 | 0.0442 | 0.0206 |
| | (0.0232) | (0.0386) | (0.0149) | (0.0094) |
| Non-Aboriginal | 0.7475 | 0.0923 | 0.0222 | 0.0094 |
| | (0.0033) | (0.0056) | (0.0019) | (0.0011) |
| 2010 | | | | |
| Aboriginal | 0.7023 | 0.1347 | 0.0324 | 0.0142 |
| | (0.0193) | (0.0320) | (0.0105) | (0.0064) |
| Non-Aboriginal | 0.7466 | 0.0945 | 0.0212 | 0.0083 |
| | (0.0032) | (0.0051) | (0.0016) | (0.0009) |
| 2014 | | | | |
| Aboriginal | 0.6644 | 0.2313 | 0.0718 | 0.0348 |
| | (0.0199) | (0.0353) | (0.0156) | (0.0098) |
| Non-Aboriginal | 0.7388 | 0.1054 | 0.0267 | 0.0114 |
| | (0.0030) | (0.0049) | (0.0016) | (0.0010) |
| 2017 | | | | |
| Aboriginal | 0.6708 | 0.1893 | 0.0484 | 0.0212 |
| | (0.0168) | (0.0307) | (0.0118) | (0.0071) |
| Non-Aboriginal | 0.7295 | 0.1190 | 0.0300 | 0.0129 |
| | (0.0031) | (0.0052) | (0.0018) | (0.0011) |

*Notes:* These values are calculated from the MCMC draws from the parameters of the beta mixture model. They match closely those calculated from the raw data, reported in Table S9 of the Supporting Information. For the headcounts and *FGT* indices, poor mental health was defined as those scores less than 0.5, a threshold suggested by the SF-36 survey.

the characteristics of the raw data for each of the subsamples are provided in Table B9 of the Supporting Information. To examine evidence on narrowing of the gap between Aboriginal and non-Aboriginal mental health, we begin by considering the posterior means and standard deviations for the mean scores, the headcounts and the FGT indices reported in Table 7. The mean mental health scores for the non-Aboriginal sample are above those for the Aboriginal sample in all years. Also, the gap between the two is relatively constant between 2001 and 2010, but increases dramatically in 2014, and then decreases slightly in 2017, leaving a gap larger in 2017 than it was in 2001. Considering the proportion of people with poor mental health (the headcount), and the severity of the poor mental health ($FGT_1$ and $FGT_2$), reveals a similar story. From 2001 to 2017, the proportion of people with poor mental health and its severity increased for both groups, and the gap between the values for the Aboriginal and non-Aboriginal samples widened.

The much smaller sample sizes for the Aboriginal population—about 2%–3% of the total sample size—mean that their population quantities are estimated less accurately than those for the non-Aboriginal population. To illustrate the difference, in Figures 6 and 7 we plot the posterior densities for the means and the headcounts for the years 2001 and 2017. The greater uncertainty associated with the Aboriginal measures is reflected in wider dispersion in their posterior distributions. Overlapping of the distributions means we must be

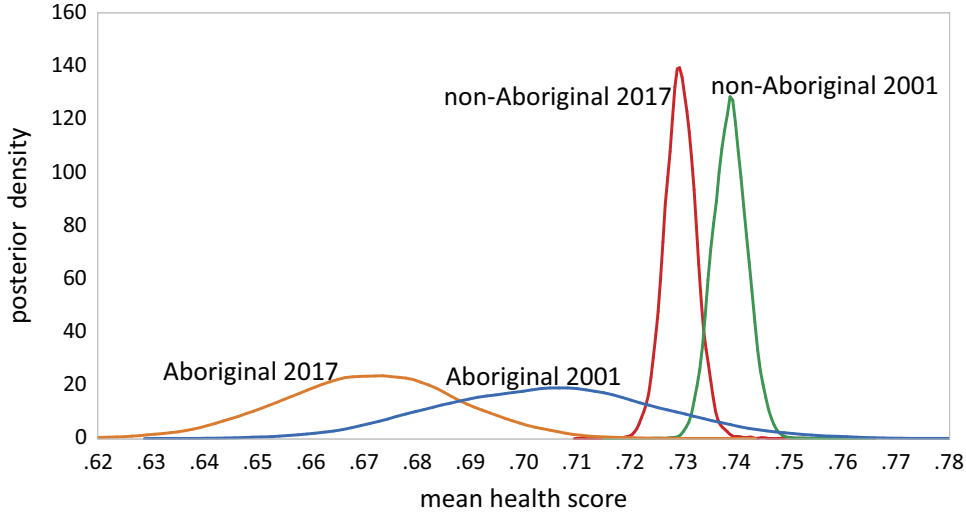


Figure 6. Posterior densities for mean health score for Aboriginal and non-Aboriginal populations in 2001 and 2017.

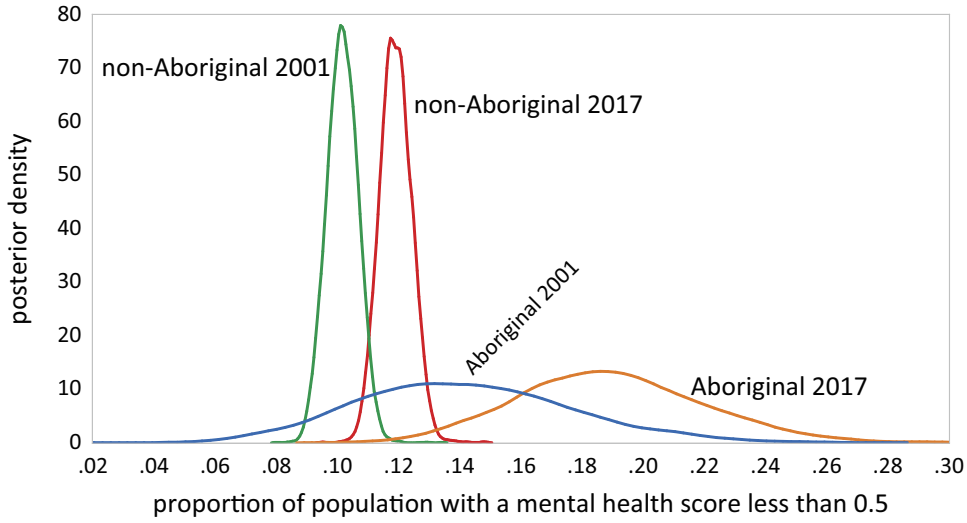


Figure 7. Posterior densities for proportions of people with poor mental health (headcounts) for Aboriginal and non-Aboriginal populations in 2001 and 2017.

cautious about drawing conclusions about differences that could be attributable to sampling error, particularly for the headcounts in Figure 7. However, the evidence points towards a deterioration in mental health and a widening of the gap.

Dominance probabilities for comparing the Aboriginal/non-Aboriginal subgroups are reported in Table 8 for their complete populations ($y < 1$) and for those with poor mental health ($y \leq 0.5$). For the complete populations and FSD, no dominance is the most likely

Table 8. First- and second-order stochastic dominance probabilities for comparing Aboriginal and non-Aboriginal mental health score distributions in each of 5 years.

| | $A_{FSD}$Non-A | Non-$A_{FSD}$A | No FSD | $A_{SSD}$Non-A | Non-$A_{SSD}$A | No SSD |
|---|---|---|---|---|---|---|
| 2001 | | | | | | |
| $y < 1$ | 0.0029 | 0.1094 | 0.8877 | 0.0235 | 0.4240 | 0.5525 |
| $y \leq 0.5$ | 0.0756 | 0.3978 | 0.5266 | 0.1725 | 0.4358 | 0.3917 |
| 2006 | | | | | | |
| $y < 1$ | 0.0004 | 0.0684 | 0.9312 | 0.0026 | 0.7121 | 0.2853 |
| $y \leq 0.5$ | 0.0103 | 0.6965 | 0.2932 | 0.0314 | 0.7212 | 0.2474 |
| 2010 | | | | | | |
| $y < 1$ | 0.0001 | 0.1112 | 0.8887 | 0.0017 | 0.6251 | 0.3732 |
| $y \leq 0.5$ | 0.0336 | 0.5727 | 0.3937 | 0.0801 | 0.6315 | 0.2884 |
| 2014 | | | | | | |
| $y < 1$ | 0.0000 | 0.3845 | 0.6155 | 0.0000 | 0.9479 | 0.0521 |
| $y \leq 0.5$ | 0.0000 | 0.9470 | 0.0530 | 0.0001 | 0.9480 | 0.0519 |
| 2017 | | | | | | |
| $y < 1$ | 0.0000 | 0.1788 | 0.8212 | 0.0001 | 0.6469 | 0.3530 |
| $y \leq 0.5$ | 0.0032 | 0.6172 | 0.3796 | 0.0306 | 0.6470 | 0.3224 |

*Notes*: 'A' refers to Aboriginal and 'non-A' refers to non-Aboriginal; $y < 1$ refers to comparisons of whole distributions, while $y \leq 0.5$ refers to comparisons of the left tails of the distributions that represent poor mental health.

outcome with the probabilities for no dominance all being greater than 0.82 in all years except in 2014 where the no-dominance probability was 0.62. This result is perhaps surprising given the differences in the means. It can be explained by the smaller sample size for the Aboriginal subgroup. This smaller sample size results in greater dispersion of $D_1(y)$ and $D_2(y)$ at each point $y$, increasing the probability that their values can be both positive and negative, and making no dominance a more likely outcome. When we consider FSD for those with poor mental health, there is more evidence of dominance of the Aboriginal group by the non-Aboriginal group. It is the most likely outcome in all years except 2001 with the highest probability being 0.95 in 2014. With SSD, the results suggest dominance of the Aboriginal group by the non-Aboriginal group for both the complete population and for those with poor mental health, in all years except 2001. The two sets of probabilities are almost identical, with the 2014 probability being 0.95 and the other probabilities in years 2006–2017 ranging between 0.63 and 0.72. For both FSD and SSD, the probability of the Aboriginal distribution dominating the non-Aboriginal distribution is close to zero. From 2014 to 2017, evidence of a gap between the Aboriginal and non-Aboriginal populations weakens, but the evidence of a gap is stronger in 2017 than it is in 2001. In Figure 8 we plot the probability curves and $D_1(y)$ and $D_2(y)$ for the non-Aboriginal distribution dominating the Aboriginal distribution in 2001 and 2017. It is clear why the FSD probabilities for non-Aboriginal over Aboriginal are so much less than those for SSD. There is a sharp decline in the FSD probability curves for $y > 0.9$, with the mean CDFs crossing at approximately $y = 0.95$. In contrast, the SSD probability curves reach their highest point when $y > 0.8$. The sharp decline in the FSD probability curves also explains why the FSD dominance probabilities are considerably higher when we consider only those with poor mental health. Finally, the SSD probability curves having their highest values for large $y$ explains why the SSD dominance probabilities for the whole populations are approximately equal to those for the population where $y \leq 0.5$.

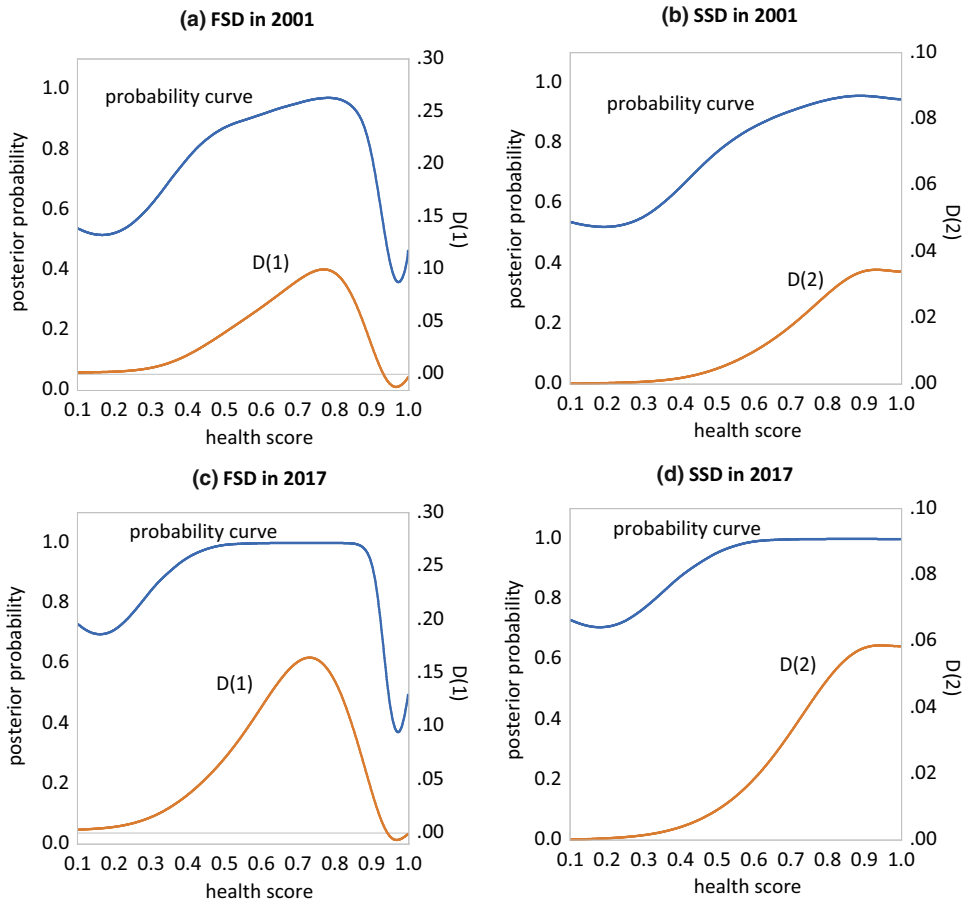


Figure 8. Probability curves $\Pr\left[D_i(y) \geq 0\right]$ (see equation (9)) and function differences $D_i(y)$ (see equations (7) and (8)) for the non-Aboriginal mental health score distributions dominating the Aboriginal mental health score distributions in 2001 and 2017. The probability that one distribution dominates another is less than or equal to the minimum of the probability curve. For FSD, it is the right tails of the distributions that have the greatest impact on the probability of dominance; for SSD, it is the left tails.

Dominance probabilities reflecting how the Aboriginal distribution of scores has changed over time are reported in Table 9. For both the whole population and the population with poor mental health, there is no evidence that the distribution of mental health scores has improved or deteriorated over time. The strongest piece of evidence is for deterioration from 2010 to 2014. In this case, posterior probabilities greater than 0.7 for 2010 dominating 2014 were obtained for SSD for the whole population and for both FSD and SSD for those with poor mental health. Comparing 2001 with 2017, the two end points of the sample, we find very small probabilities for 2017 dominating 2001; no dominance and 2001 dominating 2017 have approximately equal probabilities except for FSD using the whole population, where no dominance is more likely. A complete set of probabilities is given in Tables B10 to B13 of the Supporting Information.

Table 9. First- and second-order stochastic dominance probabilities comparing Aboriginal health score distributions in selected years.

| | Whole population | | Population with $y \leq 0.5$ | |
|---|---|---|---|---|
| | FSD | SSD | FSD | SSD |
| $\Pr(2006_{DOM}2001)$ | 0.0452 | 0.1254 | 0.1314 | 0.1901 |
| $\Pr(2001_{DOM}2006)$ | 0.0705 | 0.3788 | 0.4239 | 0.5087 |
| $\Pr(2006_{NO-DOM}2001)$ | 0.8843 | 0.4958 | 0.4447 | 0.3012 |
| $\Pr(2010_{DOM}2006)$ | 0.0549 | 0.3377 | 0.4115 | 0.4769 |
| $\Pr(2006_{DOM}2010)$ | 0.0454 | 0.1204 | 0.1144 | 0.1876 |
| $\Pr(2010_{NO-DOM}2006)$ | 0.8997 | 0.5419 | 0.4741 | 0.3355 |
| $\Pr(2014_{DOM}2010)$ | 0.0012 | 0.0097 | 0.0064 | 0.0157 |
| $\Pr(2010_{DOM}2014)$ | 0.1984 | 0.7298 | 0.7673 | 0.7814 |
| $\Pr(2014_{NO-DOM}2010)$ | 0.8004 | 0.2605 | 0.2263 | 0.2029 |
| $\Pr(2017_{DOM}2014)$ | 0.0708 | 0.4381 | 0.5774 | 0.6554 |
| $\Pr(2014_{DOM}2017)$ | 0.0092 | 0.0524 | 0.0360 | 0.0632 |
| $\Pr(2017_{NO-DOM}2014)$ | 0.9200 | 0.5096 | 0.3866 | 0.2814 |
| $\Pr(2017_{DOM}2001)$ | 0.0063 | 0.0400 | 0.0612 | 0.1305 |
| $\Pr(2001_{DOM}2017)$ | 0.2269 | 0.5248 | 0.5007 | 0.5514 |
| $\Pr(2017_{NO-DOM}2001)$ | 0.7668 | 0.4352 | 0.4381 | 0.3181 |

*Notes*: Whole population refers to the complete distributions; population with $y \leq 0.5$ refers to the left tails of the distributions that define poor mental health.

### 5.3. Male and female subgroups

In this section we summarise the results for the male/female subgroups. A detailed analysis like that provided in the previous section for the Aboriginal/non-Aboriginal groups can be found in the Supporting Information. The findings can be summarised as follows.

- Examining mean health scores suggests male mental health scores are substantially above female mental health scores for all years.
- Comparing the health score distributions, the probabilities for the female distribution dominating the male distribution are zero in all years. There is limited evidence of the male distribution dominating the female distribution in the earlier years; in 2014 and 2017 there is strong evidence of SSD of male health scores over female health scores.
- When considering how the male and female distributions have changed over time, the FSD probabilities show no evidence of dominance in either direction. Similarly, the SSD probabilities for males show no evidence of dominance in either direction. There is moderate evidence that mental health scores for females in 2010 and 2006 have SSD over 2017 (probabilities of 0.83 and 0.77 respectively); comparing 2001 with 2017, we find that 2001 dominating 2017, and neither of these years being dominant, are approximately equally likely.
- When considering a male–female comparison for those with poor health (those with a score below 0.5), the FSD and SSD probabilities for the female distribution dominating the male distribution are all approximately zero. Those for the male distribution dominating the female distribution are greater than 0.8 in 2014 and 2017, and greater

than 0.65 in 2010, pointing towards greater incidence of poor female mental health in the later years.
- There is evidence to suggest that, for those with poor health, the mental health of males improves more than that of females when mental health of the whole population is improving, and it deteriorates less than that of females when population mental health is deteriorating.

## 6. Conclusions

We have used stochastic dominance concepts to present evidence on how the distribution of Australian mental health scores has changed over time and to compare distributions for Aboriginal/non-Aboriginal and male/female subgroups of the population. Summarising the evidence leads to the following key findings:

1. In terms of SSD, there were improvements in mental health from 2001 to 2010, but a deterioration from 2010 to 2017. This conclusion holds for the whole population, for both males and females, and for those with poor mental health.
2. When considering the stricter criterion of FSD, the conclusions in (1) also hold for the subset of the population with poor mental health. However, it is difficult to establish dominance in either direction for the whole population.
3. Males experience a better level of mental health than females, and this is particularly the case at levels of mental health considered poor. There is also evidence that the gap between female and male mental health has been widening towards the later years.
4. The relatively small sample from the Aboriginal population makes conclusions about the Aboriginal/non-Aboriginal comparison less definite, but the results point towards poorer mental health for the Aboriginal population and a widening of the gap between Aboriginal and non-Aboriginal mental health.

These findings lend support to mental health clinicians who have been advocating for increased funding for mental health, and to the government of the Australian State of Victoria who has recently recognised the need for increased funding with a substantial grant.

Drawing conclusions using the criterion of stochastic dominance is a novel and more exacting approach that considers complete distributions of mental health scores rather than single statistics such as the mean. Socially undesirable outcomes such as a deterioration in mental health for some segments of the population can be hidden by single statistics but will show up as 'no dominance' when distributions are compared via stochastic dominance. Computing posterior probabilities for each of the three possible outcomes provides more information about those outcomes than accept–reject hypothesis testing decisions where failing to reject a null hypothesis does not constitute strong evidence in favour of that hypothesis. Any conclusions do have to be qualified, however, because they do depend on the validity of the scores as a representation of mental health, and on the legitimacy of comparing distributions over time and for different population subgroups. We need to assume similar responses to the questionnaire in different time periods, or from different population subgroups, are indicative of the same level of mental health.

## Supporting information

Additional supporting information may be found in the online version of this article at http://wileyonlinelibrary.com/journal/anzs.

**Appendix S1:** Supporting Information.

## References


BECHTEL, L., LORDAN, G. & RAO, D.S.P. (2012). Income inequality and mental health, empirical evidence from Australia. *Health Economics*, **21**, 4–17.

BROOM, H., SOUZA, R.M., STRAZDINS, L., BUTTERWORTH, P., PARSLOW, R. & RODGERS, B. (2006). The lesser evil: bad jobs or unemployment? A survey of mid-aged Australian. *Social Science and Medicine*, **63**, 575–586.

BUTTERWORTH, P. & CROSIER, T. (2004). The validity of SF-36 in an Australian National Household Survey: demonstrating the applicability of the household income and labour dynamics in Australia (HILDA) survey to examination of health inequalities. *BMC Public Health*, **4**, 1–11.

DAVIDSON, R. & DUCLOS, J.Y. (2013). Testing for restricted stochastic dominance. *Econometric Reviews*, **32**, 84–125.

DUNSON, D.B. & PEDDADA, S.D. (2008). Bayesian nonparametric inference on stochastic ordering. *Biometrika*, **95**, 859–874.

ESCOBAR, M.D. & WEST, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of American Statistical Association*, **90**, 577–588.

FOSTER, J., GREER, J. & THORBECKE, E. (1984). A class of decomposable poverty measures. *Econometrica*, **52**, 761–766.

GELFAND, A.E. & KOTTAS, A. (2001). Nonparametric Bayesian Modeling for stochastic order. *Annals of the Institute of Statistical Mathematics*, **53**, 865–876.

GUNAWAN, D., GRIFFITHS, W.E. & CHOTIKAPANICH, D. (2020). Posterior probabilities for Lorenz and stochastic dominance of Australian income distributions. *Economic Record*, **97**, 504–524.

HOFF, P.D. (2003). Bayesian methods for partial stochastic orderings. *Biometrika*, **90**, 303–317.

HWANG, B.S. & CHEN, Z. (2015). An integrated Bayesian nonparametric approach for stochastic and variability orders in ROC curve estimation: an application to endometriosis diagnosis. *Journal of American Statistical Association*, **110**, 923–924.

LANDER, D., GUNAWAN, D., GRIFFITHS, W.E. & CHOTIKAPANICH, D. (2020). Bayesian assessment of Lorenz and stochastic dominance. *Canadian Journal of Economics*, **53**, 767–799.

MCGORRY, P. (2005). Every me and every you: responding to the hidden challenge of mental illness in Australia. *Australian Psychiatry*, **13**, 3–15.

SCHMITS, H. (2011). Why are the unemployment in worse health? The casual effect of unemployment on health. *Labour Economics*, **18**, 71–78.

WALKER, S.G. (2007). Sampling for Dirichlet mixture model with slices. *Communications in Statistics-Simulation and Computation*, **36**, 45–54.

WATSON, N. & WOODEN, M. (2012). The HILDA survey: a case study in the design and development of a successful household panel study. *Longitudinal and Life Course Studies*, **3**, 369–381.

WARE, J.E., SNOW, K.K., KOLINSKI, M. & GANDEK, B. (1993). *SF-36 Health Survey Manual and Interpretation Guide*. Boston, MA: The Health Institute New England Medical Centre.

WARE, J.E. & GANDEK, B. (1998). Overview of the SF-36 health survey and the international quality of life assessment (IQOLA) project. *Journal of Clinical Epidemiology*, **11**, 903–912.