Psychometric properties of the Perceived Stress Scale (PSS), Social Support Scale (SSS) and Sense of Personal Control Scale (SPCS) in Aboriginal Australian populations

by

PEDRO HENRIQUE RIBEIRO SANTIAGO

Thesis submitted for the degree of Doctor of Philosophy (PhD)
Adelaide Dental School

Supervised by:

Prof. Lisa Jamieson
Adelaide Dental School
The University of Adelaide

A/Prof. Lisa Gaye Smithers
School of Public Health
The University of Adelaide

A/Prof. Rachel Roberts
School of Psychology
The University of Adelaide

January 2020

Table of Contents

Research Outcomes

The following research outcomes were produced during the PhD candidature.

Publications

The current thesis included four research articles, out of which one is published and three have been submitted for publication and are currently under review:

1. Santiago, P. H. R., Roberts, R., Smithers, L. G., & Jamieson, L. (2019). Stress beyond coping? A Rasch analysis of the Perceived Stress Scale (PSS-14) in an Aboriginal population. PloS one, 14(5), e0216333.
2. Santiago, P. H. R., Nielsen, T., Smithers, L. G., Roberts, R., & Jamieson, L. (2019). Measuring Stress in Australia: Validation of the Perceived Stress Scale (PSS-14) in a Nationally Representative Sample. Submitted to the Behavioral Medicine journal.
3. Santiago, P. H. R., Smithers, L. G., Roberts, R., & Jamieson, L. (2019). Networks of support: psychometric properties of the Social Support Scale (SSS) in two Aboriginal samples. Submitted to the Australian and New Zealand Journal of Psychiatry journal.
4. Santiago, P. H. R., Nielsen, T., Roberts, R., Smithers, L. G., & Jamieson, L. (2019). Sense of Personal Control: can it be assessed fairly across Aboriginal and non-Aboriginal Australians?. Submitted to the PloS one journal.

In the appendix, this thesis also includes two additional research articles. Although these articles were not part of the original research question, they were included since: (1) I contributed during the period of my candidature as a joint first author or second author by conducting the psychometric analysis and providing advice on methodological issues; and (2) the research topics were relevant to the current project, since they consisted of the validation of an ethnic-racial identity instrument for Aboriginal children and the development and validation of an instrument to measure racial-related attitudes and multiculturalism in Australia. In general, despite the psychometric analysis being conducted with a different methodology (i.e. traditional factor analysis rather than Rasch modelling) aiming at answering other research questions, these two additional papers were included in the Appendix to provide further evidence of my learning process and research output during the PhD candidature.

5. Manzini, D.M., Santiago, P. H. R., Roberts, R., Smithers, L. G., Paradies, Y., & Jamieson, L. (2019). Ethnic-racial identity affirmation: Validation in Aboriginal Australian children. PloS one, 14(11), e 0224736.
6. Santiago, P. H. R., Haag, D., Manzini, D.M., Bastos, J.L.D., Paradies, Y., & Jamieson, L. (2019). Development and initial psychometric assessment of the Race-related Attitudes and Multiculturalism Scale in Australia. Submitted to the PloS one journal.

Conference presentations

1. 11th Florey International Postgraduate Research Conference 2017 - with the poster "Psychometric properties of the Perceived Stress Scale (PSS), Social Support Scale (SSS) and Sense of Personal Control Scale (SPCS) in an Aboriginal Population".

2. South Australian State Population Health Conference 2017 - with the poster "Psychometric properties of the Perceived Stress Scale (PSS), Social Support Scale (SSS) and Sense of Personal Control Scale (SPCS) in an Aboriginal Population".

3. 7th International Conference on Probabilistic Models for Measurement 2018 - oral presentation with the title "Psychometric Properties of the Perceived Stress Scale in an Aboriginal Population".

4. 25th Biennial Meeting of the International Society for the Study of Behavioural Development – with the poster "Cross-cultural validity and psychometric properties of the Perceived Stress Scale (PSS) in an Aboriginal population".

5. 2018 Florey International Postgraduate Research Conference - with the poster "Measuring stress in Australia: cross-cultural validity of the PSS-14 between Aboriginal and non-Aboriginal respondents".

6. 2019 International Meeting of the Psychometric Society – with the poster "Psychometric properties of the SSS in an Aboriginal population".

7. Social Medicine & Population Health and International Epidemiology Association European Congress Joint 63rd Annual Scientific Meeting – with the oral (rapid fire) presentation "Support that comes from culture: a Rasch analysis of the Social Support Scale (SSS) in an Aboriginal population".

Grants and awards

1. Carlsberg Foundation travel grant for the project "Unbiased measurement of psychological and emotional well-being in Australian Aboriginal and Torres Strait Islander populations – is it possible?" (CF18-0384) – Co-Investigator.

2. Northern Communities Health Foundation Prize - for the poster presentation "Psychometric properties of the Perceived Stress Scale (PSS), Social Support Scale (SSS) and Sense of Personal Control Scale (SPCS) in an Aboriginal Population" at 11th Florey International Postgraduate Research Conference.

3. Eustace Travelling Scholarship 2018 for the presentation "Psychometric Properties of the Perceived Stress Scale in an Aboriginal Population" in the Seventh International Conference on Probabilistic Models for Measurement.

4. Eustace Travelling Scholarship 2019 for the presentation " Support that comes from culture: a Rasch analysis of the Social Support Scale (SSS) in an Aboriginal population" in the Social Medicine & Population Health and International Epidemiology Association European Congress Joint 63rd Annual Scientific Meeting.

5. Adelaide Dental School Prize - for the poster presentation "Drop-the-*p*: Bayesian CFA of the Multidimensional Scale of Perceived Social Support" at 13th Florey International Postgraduate Research Conference.

Invited presentations

1. “Psychometric properties of the Perceived Stress Scale in Australia (PSS-14) and Denmark (PSS-10), a Graphical Loglinear Rasch analysis” at PsychMeasure/Department of Psychology at The University of Copenhagen (14th December 2018).

2. “Psychometric properties of the Perceived Stress Scale in Australia (PSS-14) and Denmark (PSS-10), a Graphical Loglinear Rasch analysis” at Methods and statistics workgroup/Faculty of Social and Behavioural Sciences at The University of Amsterdam (30th November 2018).

Abstract

Background: The history of colonization contributed to Aboriginal and Torres Strait Islanders becoming one of the most disadvantaged groups in Australia. The experienced inequalities in virtually all areas, including employment, income and educational attainment, generate chronic stress, low sense of personal control and lack of social support in the Aboriginal population. Despite these psychosocial variables (perceived stress, sense of personal control and social support) being suggested as important to Aboriginal health, the only measurement instruments available were originally developed in Western countries, with no instruments validated specifically for Aboriginal Australians. The aim of this PhD project was to evaluate the validity and reliability of the Perceived Stress Scale (PSS-14), Social Support Scale (SSS) and Sense of Personal Control Scale (SPCS) in an Aboriginal population.

Methods: The main sample was composed of 367 pregnant Aboriginal women who participated in the Baby Teeth Talk Study, an oral-health randomized controlled trial (RCT) conducted in South Australia. Secondary samples comprised: (1) 317 Aboriginal participants from the Teeth Talk Study, an RCT designed to improve oral-health literacy; and (2) 3,857 non-Aboriginal Australians in the population-based cross-sectional study Australia's National Survey of Adult Oral Health 2004-2006. The psychometric properties of the three scales were analyzed with the Rasch model and Graphical Log-linear Rasch models. The properties evaluated were: (a) dimensionality, (b) model fit, (c) item fit, (d) local dependence, (e) differential item functioning (DIF), (f) reliability, (g) targeting and (h) criterion validity.

Conclusions: The findings indicated initial evidence of validity from a revised PSS, after the exclusion of one misfitting item, and a revised SPCS, after the exclusion of five misfitting items. In the case of the SPCS, the development of new culturally specific items is recommended. There was robust evidence that the original 4-item version of the SSS is valid for Aboriginal Australians considering that the good psychometric properties were replicated

in two independent samples. The overall conclusion was that, while certain instruments required more modifications than others (e.g. SPCS compared to the SSS), adapted versions of the 3 instruments are available for future research with Aboriginal Australians.

Thesis Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree. I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works. I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signature:

Date: 13th of January 2020.

"On the great road of buddha ancestors there is always unsurpassable practice, continuous and sustained. It forms the circle of the way and is never cut off. Between aspiration, practice, enlightenment, and nirvana, there is not a moment's gap; continuous practice is the circle of the way" – Shobogenzo, 1240 (Eihei Dogen).

Acknowledgments

I acknowledge the Kaurna people as the traditional custodians of the Adelaide region, where this PhD research was conducted, and that their cultural and heritage beliefs are still as important to the living Kaurna people today.

I would like to acknowledge, first and foremost, my family. My father Diógenes Gori Santiago, my mother Maria Thereza Teixeira Soares Ribeiro and my sister Nicole Santiago. I remember growing up in a house full of books and reading Robinson Crusoe as a child, I remember when my father bought me VHSs with English lessons and going to my first private English classes. You have always been supportive of my life choices and endured staying (physically) away from me during these years so I could pursue and further my studies. Diógenes, Tetê and Nicole, you are the most important people for me, I'll love you forever.

Secondly, I'd like to acknowledge my supervisors Prof. Lisa Jamieson, A/Prof. Lisa Smither and A/Prof. Rachel Roberts. Since day one of my candidature I had the perception that I had the best supervisory group, a perception that remains unchanged up to this day. Over these 3 years, you were able to create an extremely supportive environment that facilitated my growth and learning throughout my candidate. Your unremitting dedication to helping me and positive influence impacted much more than my academic trajectory but will stay with me for life. I once heard that the best lessons are not the ones directly taught but the ones given by example. For this reason, more than anything, I'd like to acknowledge my principal supervisor, Lisa Jamieson, which gave me daily examples of competence, wisdom, and kindness. In every conversation we had over the past years I felt mentored and every suggestion she gave put me on a whole new trajectory. The output of this PhD is a direct consequence of her work and life-long dedication to Indigenous research. I will never be able to thank you enough Lisa but will be forever grateful.

Thirdly, I'd like to acknowledge my friends Dr. Heloisa Godoi and Dr. Greiciely Oliveira for the friendship during our first months in Adelaide, and Dr. Benjamin Attcliffe and Dr. Prashanth Rajan for the friendship over the years. I'd also like to acknowledge PhD student Jasmine Lau for the coffees every Monday morning. I'd like to acknowledge my personal friend Bernardo Torres for the intellectually stimulating conversations. I want to acknowledge all other PhD students from ARCPOH, informally known as the "*Yellow Dice Gang*", namely Davi Manzini, Sathvika Justine, Dr. Helena Schuch, Dr. Dandara Haag, Dr. Ankur Singh, Mi Du, Dr. Rahul Nair, YoungHa Song, Sneha Sethi, Anna Ali, Dr. Kostas Kapellas, Dr. Ninuk Hariyani, Dr. Saima Islam, Dr. Kammal Hanna, Arash Ghanbarzadegan, Mehrsa Zakershahrak, and honorary members Dr. Sharukh Khan and Gustavo Hermes Soares. Furthermore, I want to acknowledge the friendship and amazing work conducted by all members of the Indigenous Oral Health Unit, including Jacqueline Aldis, Joanne Hedges, Helen Mills, and others. Finally, I'd also like to acknowledge Prof. John Lynch, Dr. Angela Gialamas and all members of the Better Start Research Group for the learning over the years.

Chapter 1 – Introduction

Background

From Porteus to Ewert: 20th-century psychological assessment of Aboriginal Australians

The history of psychological assessment of Aboriginal Australians started during the first decades of the 20th century. Initially, psychological assessment was conducted using instruments originally developed for Western populations and without consideration of cultural differences. This practice brought devastating consequences to the Aboriginal people. One example was the work conducted by the Australian psychologist Stanley Porteus. In the early 1900s, Porteus applied the maze test, a paper-and-pencil test in which the respondent needs to trace a line to exit a schematically printed maze, to measure the general intelligence of Aboriginal Australians (Bin-Sallik, 1990; Goldstein, Princiotta, & Naglieri, 2015; Porteus, 1950). The results showed that Aboriginals underscored in comparison to white non-Aboriginal Australians, leading Porteus to conclude that: “The available evidence with regard to Aboriginal mentality seems to point indubitably towards a somewhat general mental inferiority as regards abstract intelligence or the capacity to deal with abstract symbols of thought” (Ranzijn, McConnochie, & Nolan, 2009, p. 188). These results, extensively reported by the press at the time as “scientific evidence”, supported the implementation of assimilation policies in Australia and culminated in the removal of Aboriginal children from their parents; these children are called the Stolen Generations (Dudgeon, Rickwood, Garvey, & Gridley, 2014). The forced removal from the parents resulted in trauma and life-long consequences for the children. For example, many of them experienced sexual abuse after being placed in state care were adopted, were told that their parents were dead or had abandoned them, and received low levels of education, among other forms of severe human rights violations (Behrendt, 2012).

In the early 1900s, when Porteus applied the maze test, it was not common in Aboriginal Australian society to draw using paper and pencils. Aboriginal Australians usually drew when they were telling stories to each other, for example in ceremonial functions, and this was done with their fingers or by using twigs to make forms in the sand (Behrendt, 2012; Bin-Sallik, 1990). The maze test was applied and the results were interpreted without any consideration of cultural bias (Goldstein et al., 2015).

Many decades later, in 1967, Porteus retested Aboriginal Australians with the maze test after performing cultural adaptations. The Aboriginals were told that the maze was a monkey house (an Aboriginal term for prison) and that at the center was a hungry kangaroo that needed to escape without hitting the walls. This time Aboriginal Australians scored *the same* as their non-Aboriginal counterparts (Bin-Sallik, 1990). Unfortunately, these results were 38 years too late and the forcible removal of Aboriginal Australian children from their families as a governmental policy lasted for two more years (Young, 2009).

The example of Stanley Porteus's work shows how psychological assessment in Aboriginal Australians without cultural appropriateness led to incorrect evidence which provided a ("scientific") rationale to assimilationist and eugenics policies. Since the application of the maze test by Porteus in the early 1900s, the development of modern statistical methods gave researchers new tools to evaluate test validity in specific cultures and the importance of psychometric research has been increasingly recognized. For example, recently, a seminal decision by the Supreme Court of Canada brought worldwide attention to the debate about culturally-appropriate psychological instruments for Indigenous people.

In 2015, an Aboriginal offender named Jeffrey G. Ewert legally challenged the Correctional Services Canada pertaining to the use of Western-developed risk assessment tools to decide parole eligibility for Indigenous adults (McCuish, Mathesius, Lussier, & Corrado, 2018). Over his three decades of incarceration, Jeffrey G. Ewert was assessed with several risk assessment tools, such as the Psychopathy Checklist-Revised (Hare, 2003), and the results indicated he had "too great risk of reoffending" (Haag, Boyes, Cheng, MacNeil, &

Wirove, 2016, p. 70). However, the plaintiff (Jeffrey G. Ewert) argued that these instruments, previously validated in Western cultures, were applied without consideration of validity for Indigenous culture and results could potentially be biased. The case was judged favourably to the plaintiff by the Supreme Court of Canada (2018, p. 168), which declared that: “The Correctional Services Canada had long been aware of concerns regarding the possibility of these tools exhibiting cultural bias yet took no action to confirm their validity and continued to use them in respect of Indigenous offenders, despite the fact that research would have been feasible”. Hence, the Supreme Court of Canada (2018, p. 179) ruled that the CORRECTIONAL SERVICES CANADA should conduct psychometric validation specific for Canadian Indigenous and “take all reasonable steps to ensure that any information about an offender that it uses is as accurate as possible”. The case Ewert. vs Canada brought the attention of the scientific community (and civil society) to questions such as: “Are there universal psychological concepts that operate between people within a culture, people between cultures, different genders, different periods in history, and so forth?” (Haag et al., 2016, p. 71).

In the following sections of this chapter (Chapter 1), I aim to provide an answer to this question through a literature review that establishes the scientific rationale behind the research conducted in this PhD. I’ll focus the discussion on categorical fallacy, a concept proposed by Harvard professor of cross-cultural psychiatry and medical anthropology Arthur Kleinman (1987) to explain why psychological constructs are culturally-bound and should not be assumed to be equivalent across cultures. I’ll discuss how colonialism impacted the creation of Indigenous identity (Section 1), the historical trauma experienced by Indigenous people (Section 2), culture-bound psychological constructs (Section 3), the notion of validity for psychological instruments (Section 4) and procedures for developing culturally-appropriate instruments (Section 5). In Chapter 2, “General aim and specific objectives”, I will present the research questions/objectives and describe the main sample and measures used in this research project. In Chapters 3 to 6, I will present the four Research Papers.

In Chapter 7, I discuss the R function Item Characteristic Curves developed during the PhD candidature. The Item Characteristic Curve is the graphical representation of the regression model that calculates expected item responses (outcome) as a function of the latent trait (exposure) according to the Rasch model (link function). Thus, the visual inspection of the Item Characteristic Curve is used for the evaluation of model fit since it is possible to examine whether the observed item responses diverged from the expected item responses. During my candidature, I developed a R function to plot Item Characteristic Curves since the available functions from mainstream Rasch analysis R packages had limitations. For example, the Item Characteristic Curve function from R package *eRm* functioned only for dichotomous items (e.g. 0=Disagree, 1=Agee), while the three scales validated in this PhD were composed of polytomous items (1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly Agree). Furthermore, the R package *ltm* plots Item Characteristic Curves for polytomous items but does not display observed item responses, so visual investigation of model fit was not possible. The most complete Item Characteristic Curves can be found in the private software RUMM2030 (Rasch Unimensional Measurement Models); however, it was impossible to integrate this function with the broader, free and open-source statistical framework R software. Hence, I developed the function initially for plotting graphs in the papers, but the project received attention from researchers in the area, evolved and now the function will be soon available on CRAN as part of the R package *iarm*.

Finally, in Chapter 8, I discuss the Final Considerations. In the Appendix section, I included two additional research papers that were produced as a result of the learning process occurred during this candidature.

Literature review

Colonialism and the creation of Indigenous identity

The definition of the term Indigenous people is of a group that were early inhabitants of a country and which were affected by colonization (C. E. Burnette & Figley, 2016). The word Indigenous refers to something that grows, lives or occurs naturally in an environment, being used not only to refer to Indigenous people but also to Indigenous plants, Indigenous cultures, among others. The word Indigenous derives from the Latin word *indigena*, and from old Latin *indu*, which are nouns that mean native (Budyartati, 2015). It is important to notice that the definition of Indigenous people is made in relation to the process of colonization, and the designation of any people as "Indigenous" acquires substance only when there are other populations in the region that can be described as settlers or aliens (Béteille, 1998).

To designate a population as Indigenous, it is necessary to have a large temporal gap between the onset of early populations and a subsequent invasion by other groups. Throughout history, populations used to mix constantly by being conquered or simply moving around. The frequent change in demographics experienced by certain populations makes the definition of Indigenous not always clear (Bowen, 2000). For example, in Africa, the term Indigenous has been adopted by certain groups but the fact that there was not a dominant colonizer makes it harder to establish who is Indigenous and who is not (Hodgson, 2002). On the other hand, in places such as Australia and North America, there is no doubt about the differences in identities of initial inhabitants and later invaders (Béteille, 1998). The clear differences in modes of life and physical appearance between native populations and late settlers made the concept of "Indigenous people" in these countries broadly accepted (Bowen, 2000). Thus, the definition of Indigenous people cannot be untangled from the invasion and settlement by a foreign population, a process which is called *colonialism*.

The word colonialism in The Dictionary of Human Geography (1981) is described as "an enduring relationship of domination and mode of dispossession, usually (at least initially) between an Indigenous (or enslaved) majority and a minority of interlopers (colonizers), who are conviced of their own superiority, pursue their interest and exercise power through a mixture of coercion, persuasion, conflict and collaboration". Colonialism is the conquest and control of a population, lands and goods, called the colony, by another country, called a mother country, metropolis or empire (Johnston, 1981).

The process of colonialism did not start, as it is mostly associated, with the expansion of various European countries through Africa, Asia, and America in the 16th century; it dates back to the conquests of the Roman Empire. Nevertheless, modern capitalist colonialism was initiated during the 16th century by introducing new forms of colony subjugation. Notably, in addition to stealing resources and enslaving the population, the goal of modern capitalist colonialism was to transform the dominated country into a compulsory market for products of the metropolis. For example, India raw cotton was taken to America, where African slaves produced clothes from it and the final product, clothes, was moved back to be sold to Indians (Loomba, 1988). This unequal economic relationship between the countries, mother country, and colony, was essential to the growth of European capitalism and the financial uprise of European nations. There were two forms of modern colonialism: administrative and settler colonialism. Administrative colonialism did not involve the migration of a large population of the colonizing country to the colony, only a sufficient amount of people to perform administrative tasks (Loomba, 1988). Examples of administrative colonialism are Papua New Guinea, under Australian administration, and South Africa by the Dutch, initially (Hawksley, 2001; Ndlovu, 2013). Conversely, settler colonialism consisted of wholesale land expropriation, where the settlers come to stay, as happened in Australia under British invasion.

The colonizers had an innate sense of superiority when dealing with the native people and when their records were examined, like diaries, memoirs, and reports, it is possible to ascertain their racist attitude. However, it would be naïve to assume that racism, like

Negrophobia or Islamophobia, was the driving force behind the subjugation of certain populations (e.g. Indigenous people, black people). The identity of a distinct group - "Indigenous people" or "black people" - was *created* by the colonizers precisely because these groups resided on the land they aspired to possess. To create an "other", an opposition, was crucial for Europeans to justify their behavior: if colonised people are black, Europeans are white; if colonised people are irrational, Europeans are rational; if the Indigenous are barbaric and lazy, Europeans will benefit them by bringing civilization. What was created was a discourse about Indigenous people, ways of thinking about them that helped to perpetuate and legitimize domination. In this way, the primary motivation for elimination of Indigenous groups was not race, but access to territory; territoriality is settler colonialism's irreducible element (Hawksley, 2001; Loomba, 1988; Wolfe, 2006).

A characteristic of settler colonialism is the aim to disassemble the Indigenous society and replace it with "modern society", a process named *modernization*. Achille Mbembe explains that: "Like Islam and Christianism, colonization is a universalizing project. Its ultimate aim is to inscribe the colonized in the space of modernity" (Cooper, 2005, p.143). Modernization reproduces the discourse of an evolutionary transformation of societies: from a primitive society, like the Indigenous ones, to modern capitalist society, leading the colonized to growth, progress, and liberation. This reasoning was used to justify the imposition of Western social, economic, and political norms into Indigenous groups (Cooper, 2005; Garuba, 2013; Johnston, 1981), enforced through resocialization mechanisms such as missions, encouraged miscegenation, religious conversion and boarding schools (Wolfe, 2006).

One fundamental aspect of modernity is understanding the world through *rationality*, with science and evidence-based knowledge the center of this enterprise (Johnston, 1981). This belief in rationality was used by the colonizers to generate skepticism about the value and utility of the knowledge held by Indigenous people. Westerners started to disregard Indigenous knowledge and wisdom, claiming they were rooted in belief rather than reason and, consequently, should be discarded. This example shows how the modernization process brought upon Indigenous people was insidious. While modern society preached to the Indigenous that the Western way of life (e.g. consumerism, rationality, higher education, whiteness) was the standard to aspire to; it also stated that this is a standard, which due to their Indigenous history and cultural heritage, they would never be able to fully achieve (Cooper, 2005; Garuba, 2013; Johnston, 1981).

To increase this ambiguity, once Indigenous groups are labelled "Indigenous", there are certain "Indigenous" behaviors that are expected from them and deviating from those can make them subject to harassment (Bowen, 2000; Dove, 2006). For example, a common-sense belief is that an Indigenous person with white skin and blue eyes are "less" Indigenous than those with black skin and brown eyes. Furthermore, if Indigenous people present themselves as *too primitive*, they risk abuse from a more powerful society; if they present themselves as *not primitive enough*, they can appear opportunistic and inauthentic. Ultimately, Indigenous people need to conform to the *settler*'s characterization of what "Indigenousness" is (e.g. black skin, practicing traditional rituals, living in remote areas) to be heard and politically recognized in contemporary world (Lindroth & Sinevaara-Niskanen, 2013)

In summary, the creation of the Indigenous identity, what it "means" to be Indigenous in contemporary Australia, can only be understood in the context of colonization and the subsequent marginalization of Indigenous people. In the next section, I will show that to understand the well-being and suffering of Aboriginal people, it is necessary to investigate the *consequences* of colonization, the decades of discrimination and disempowerment that created more than individual trauma in Aboriginal people, but a historical trauma experienced by Aboriginal Australians as a community.

The impact of historical trauma on Indigenous well-being

The term historical oppression designates the intergenerational and pervasive experiences of oppression by Indigenous people that were imposed and later normalized into their daily lives. The historical oppression started with colonization but it is perpetuated nowadays due to ongoing social inequities such as poverty, health inequities, exclusion, and discrimination. The discrimination includes, for example, microaggressions, which are everyday injustices and demeaning messages that become part of the Indigenous experience (C. Burnette, 2015). Microaggresions contrast with more *overt* manifestations of racism, such as physical aggression and name-calling, and include subtle snubs, slights, and insults that *implicitly* communicate hostility. For example, a professor might display “happiness” and incredulity that an Aboriginal student achieved a perfect test score, displaying racism by implying that this is an unexpected event due to the student Indigenous status. Nonetheless, authors such as Lilienfeld (2017) recommend caution since the concept of microagression “have yet to be subjected to adequate scientific scrutinity” and is “little different from other nascent psychological constructs that await refinement in light of additional scientific knowledge” (p. 158). Despite the different manifestations of discrimination (e.g. physical abuse, microagressions), it is established that the historical oppression suffered by Indigenous population produce pervasive effects on their well-being; for example, Indigenous people present disproportional rates of pathological distress compared to any other non-Indigenous group (Gone, 2013). Hence, the concept of *historical trauma* encompasses the suffering experienced by Indigenous groups resulting from colonization and marginalization (Brave Heart, Chase, Elkins, & Altschul, 2011).

The historical trauma can be divided into four categories: the *colonial injury* perpetrated by the European invaders which includes murder, subjugation and dispossession; the *collective experience* of these injuries by the Indigenous population, whose identity, culture, ways of life and interactions were radically altered; the *cumulative effects* from injuries, which are the harmful consequences accumulated due to the ongoing adverse practices and policies against Indigenous people; and the *cross-generational impacts*, which are the legacies of risk and vulnerabilities that are passed from ancestors to descendants in unremitting fashion (Kirmayer, Gone, & Moses, 2014). Over recent decades, Indigenous health researchers started to realize that Western diagnostic categories such as post-traumatic stress disorder (PTSD) were not enough to encompass the complexity of the trauma experienced by Indigenous people (Gone, 2013). Thus, historical trauma (sometimes referred to as “complex PTSD”) encompasses more than the trauma of an Indigenous *individual*, it is a *collective* phenomenon shared by the members of an Indigeneous group (Kirmayer et al., 2014, p. 310).

In Western societies, psychological disorders such as PTSD and depression are usually understood through a biomedical approach, which considers causes to be found in disturbances or abnormalities within the brain, hence the term “*mental* disorder” and “*mental* health” (Haslam, 2000; McNally, 2012). For example, the cause of depression is sometimes attributed to dysfunction in the brain’s serotonin levels (i.e. the chemical imbalance theory) (Whitaker, 2005). This biomedical perspective to psychological problems, which attributes causes to biological reasons (and ultimately to the *individual*), delegitimizes individual suffering as political and moral commentary (Arthur Kleinman & Kleinman, 1991). The concept of historical trauma emphasizes that, rather than broken brain or faulty genes, colonization and marginalization is the main cause behind the high level of distress and suffering experienced by Indigenous groups (Gone, 2013).

In summary, the suffering experienced by Aboriginal Australians can be described as a historical trauma: a trauma experienced as a community due to the history of colonization and exclusion in Australian society. In the next section, I’ll discuss how psychological instruments are designed to evaluate unobservable constructs and why these constructs can not be assumed to be equivalent across cultures. This discussion will raise implications for the psychological assessment of Aboriginal Australians.

Culture-bound concepts: hypothetical constructs and categorical fallacy

One form of assessment commonly employed within health sciences is psychological tests. A psychological test is defined as a systematic procedure to compare the behaviors of two or more people through responses to a set of items (Furr & Bacharach, 2013). The use of psychological tests to understand and predict human *behavior* has expanded beyond its original field of psychology and psychiatry. For example, in dentistry, the concept of oral health has been recently defined by the World Dental Federation as "multi-faceted and includes the ability to speak, smile, smell, taste, touch, chew, swallow and convey a range of emotions through facial expressions with confidence and without pain, discomfort and disease of the craniofacial complex" (Sessle, 2017, p.5). This new definition, accepted by more than 200 national institutes, includes several *behaviors* (such as speaking, smiling, chewing, among others) as fundamental for oral health and within the scope of dentistry research (Sessle, 2017). Examples of questionnaires measuring oral-health relevant behaviors are the Health Literacy in Dentistry (HeLD) (Jones, Parker, Mills, Brennan, & Jamieson, 2014) or the Oral Health Impact Profile (OHIP) (Slade & Spencer, 1994). Hence, psychological tests are now adopted in all health sciences (Ginty, 2013)

Psychological tests are predominantly used (and useful) to measure observable behaviour *as a way to infer* unobserved attributes such as depression, extroversion, intelligence, oral-health literacy, quality of life, among others. When a psychological characteristic, process or state cannot be directly observed, it is defined as a *hypothetical construct*. For example, mental disorders are hypothetical constructs (Sanislow et al., 2010) since their investigation is done not by *directly observing the disorder* but rather by examination of a cluster of symptoms (Cuthbert & Kozak, 2013). Thus, an established psychological test such as the Beck Depression Inventory will make *inferences* about depression through 21 questions that assess behaviors of crying (e.g. "I cry all the time now") and suicidal thinking (e.g. "I don't have any thoughts of killing myself") (Beck, Steer, & Carbin, 1988; Furr & Bacharach, 2013). The behaviors that constitute depression, such as suicidal thinking, hopelessness, avoidance of social contact, are *functionally related behaviors*; they are behaviors that mutually influence each other. As an example, if a person avoids social contact, she/he will feel more lonely, might start having hopeless thoughts (e.g. “No one likes me”), which leads to more social avoidance and so forth. These sets of functionally related behaviors are then labelled as “depression” (Furr & Bacharach, 2013; Mooi & Sarstedt, 2010).

Since behaviors occur only within a given environment, mental disorders can not be promoted as universally valid. The *creation* of hypothetical constructs such as “depression” or “social anxiety” is not value-free; they contain notions of what a “disorder is” and what constitutes it. These notions are bound to cultural contexts and the theoretical understandings of a certain time (Follette & Houts, 1996; Summerfield, 2008). For example, in the history of the Diagnostic and Statistical Manual of Mental Disorders (DSM), homosexuality was considered to be a “mental disorder” (and later a “disturbance”) until 1973. Nowadays it is widely agreed that homosexuality is not a “disorder” (or “disturbance”) and scientists have demanded the removal of all sexual related categories from the DSM (Downing, 2015; Drescher, 2015; Russo & Venâncio, 2006).

To explain that psychological constructs, such as mental disorders, are culturally-bound, Kleinman (1987) introduced the concept of category fallacy. Category fallacy is the use of a diagnosis constructed for one cultural context in another where it lacks coherence and validity. The fallacy is that, because it is possible to identify similar signs and symptoms in different cultures, they would have the same meaning (Follette & Houts, 1996; Summerfield, 2008). As an example, the symptoms that Western psychiatry would classify as a psychotic depression (e.g. hearing voices, hallucinations) are recognized by the Baganda Indigenous as eByekika. The eByekika is caused when the living do not perform their obligations with the dead, such as rituals and other ceremonies. Thus, although the symptoms of eByekika seemingly correspond to psychotic depression, the conceptualization, causes, and treatment are totally different. The Baganda Indigenous people understand, for example, that going to a hospital to receive medical interventions have nothing to do with the successful treatment of eByekika (Okello & Musisi, 2006). Another piece of evidence that corroborates Kleinman’s (1987) ideas is the existence of culture-bound syndromes. In Japan, there is a culture-bound form of distress called Taijin Kyofusho. Taijin Kyofusho refers to a set of behaviors centered on anxiety of presenting oneself, such as fear of offending or displeasing others, having an offensive odor, making excessive eye contact, among others. Although Taijin Kyofusho has similarities with Western "social anxiety", the full spectrum of Taijin Kyofusho's behaviors occurs only in Japan due to specific concerns of the Japanese culture on self-presentation (Kirmayer, 1991).

To avoid misconceptions due to concepts derived in foreign cultures (such as "mental health"), Aboriginal and Torres Strait Islanders started creating their own *discursive space* about the meaning of their Indigeneity (Paradies, 2006) and advocating the term Social and Emotional Well-Being (SEWB) to conceptualize their experiences of happiness and suffering. For Westerners, mental health describes the health of the mind, *inside* an individual; while for Aboriginal Australians, individual well-being cannot be *disentangled* from the social and spiritual well-being of the whole community (Garvey, 2008).

In summary, psychological constructs (e.g. well-being, stress, personal control) created in the West should not be assumed to be equivalent to Aboriginal Australians. Therefore, psychological instruments developed to measure these constructs that were validated in Western countries do not necessarily retain their validity for Aboriginal Australians. However, before discussing the procedures to conduct an adequate validation for Aboriginal people, it is necessary first to scrutinize to concept of validity itself.

The validity of psychological instruments

A central concern to researchers when developing an instrument is establishing its validity and reliability. The validity refers to whether the instrument is measuring what it is supposed to measure, while reliability refers to whether the instrument can *consistently* measure the attribute of interest (Cook & Beckman, 2006; Furr & Bacharach, 2013; LoBiondo-Wood & Haber, 2010). The most accepted definition of validity is "the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses" (American Educational Research Association et al., 1999, p.9). The most essential concept is construct validity. Construct validity refers to the degree to which test scores can be interpreted as reflecting a particular psychological construct (Furr & Bacharach, 2013).

The cultural nature of psychological constructs requires the development of culturally-appropriate measurement instruments (Miller et al., 2006). Therefore, when instruments are being adapted to a different culture, it is necessary to demontrate *conceptual equivalence* for the instrument to maintain construct validity. Conceptual equivalence refers to when a construct possesses the same meaning in another culture (Mishra, 2013). One example where conceptual equivalence is lacking is the construct of "brand loyalty". Brand loyalty refers to how loyal costumers are regarding buying the same brand and several psychological instruments have been developed to measure this construct. The concept of brand loyalty, however, only makes sense in a capitalist society. In a socialist society, such as the Soviet Union, individuals had only one brand to choose (the governmental one) and the idea of "brand loyalty" immediately loses its significance (Mishra, 2013). Therefore, although it was possible to apply an instrument to measure "brand loyalty" in the Soviet Union and obtain test scores, there was no conceptual equivalence and the results would have no meaning. The example of "brand loyalty" highlights another important feature of validity. Although expressions such as "the test is valid" or the "validity of a test" are commonly used, validity is not a property of any test. A test is neither valid nor invalid. Valid refers to the interpretation of the test scores in a particular context (Furr & Bacharach, 2013). In the next section, after being acquainted with the concept of construct validity, I'll discuss the recommended procedures for the development of culturally-appropriate instruments.

Procedures for developing culturally-appropriate instruments

A well-established set of methods and procedures has been proposed to developed instruments for a new culture or to perform adaptations of existing ones (Prince, 2008). These procedures comprise four stages: (1) investigation of the construct in the new culture; (2) translation/adaptation of the instrument; (3) pilot testing in a sample of the population; and (4) testing in a complete sample followed by psychometric analysis.

The first stage is to investigate the relevance and equivalence of the construct in the new culture (Prince, 2008). Although certain psychological constructs, such as stress and well-being, are present in different cultures, the dimensions that compose these constructs are not necessarily similar (Ingersoll-Dayton, 2011). Hypothetical constructs possess an inner structure, being divided into distinct *dimensions* (Streiner, 2003). The research on construct dimensionality initiated in the early 1900s, when psychologists such as Charles Spearman investigated the hypothesis that the different domains of human intelligence (e.g. quantitative reasoning, visual-spatial processing) were influenced by a general attribute *g*. The "*g* theory of intelligence" proposed that human intelligence was *unidimensional*, since a *unique* general intelligence attribute would influence performance in distinct intelligence areas (e.g. solving an equation, playing music). However, the theory was met with disagreement within the scientific community, leading to the development of statistical techniques to uncover the number of dimensions constituting intelligence – techniques such as *factor analysis* which were later applied to many other psychological constructs and fields of research (Thompson, 2004).

Furthermore, the most famous example of dimensionality in all psychological research is the "Big Five" dimensions of personality. The "Big Five" dimensions of personality were empirically found to be "Openness to experience", "Conscientiousness", "Extraversion", "Agreeableness" and "Neuroticism" (Borsboom, Mellenbergh, & Van Heerden, 2003). Despite being extensively replicated among various cultures (including Indigenous ones), certain cultures displayed personality dimensions that do not correspond to the "Big Five", reinforcing thus the recommendation that dimensionality needs to be investigated specifically to a certain culture (McCrae & Allik, 2002). Several methods were then developed to identify dimensions in a given culture (e.g. an Indigenous culture). For example, narratives of Indigenous people can be evaluated to identify the number of dimensions through content analysis (Miller et al., 2006) or grounded theory (Osborne, Batterham, Elsworth, Hawkins, & Buchbinder, 2013). These approaches have already been used with Aboriginal Australians. One study showed that a construct named *Kurunpa*, which means a weakened, displaced or misaligned spirit, was to be preferred rather than "depression" in Aboriginal people (A. Brown et al., 2012).

In other cases, researchers aimed to adapt an existing instrument for a new culture rather than develop a new one. The second stage is then the process of translation/adaptation. The focus of translation/adaptation is not linguistic equivalence, but rather *conceptual equivalence*. Instead of a word-to-word translation, translators consider the concept that the original term meant and try to find a similar one in the targeted culture (Prince, 2008). The process of translation/adaptation starts with a *forward translation* from a bilingual translator, whose native language is the language of the target culture and who is acquainted with the cultural background. The next step is the creation of a *bilingual expert panel*, which will evaluate discrepancies between the translated items and the original ones, making suggestions of modifications when necessary. The third step is the *back-translation* by a new translator who has not taken part in the research yet, whose native language is the original language of the instrument and who does not have prior knowledge about the measurement instrument (Prince, 2008). The rationale of back-translations is that, if there are discrepancies between the back-translated version and the original instrument, this can indicate failures of the initial translation (Guillemin, Bombardier, & Beaton, 1993).

After the instrument has been adapted to a new culture, the third stage is *pre-testing* with a sample of the population to search for difficulties in comprehension and to ensure face validity (Guillemin et al., 1993). Although these three steps are considered "state of the art" recommendations, most studies omit these steps and proceed directly to the fourth step, statistical evaluation to identify psychometric properties (Prince, 2008).

Finally, during the fourth step, when the instruments have already been applied and the psychometric properties have been evaluated, one important concept is *measurement equivalence* (i.e measurement invariance) (Milfont & Fischer, 2015). Measurement equivalence can be defined as whether under different conditions, such as different groups of respondents belonging to distinct cultures, the *measurement operations* (i.e. participants responding to a test) yield the same *measures* of the construct (Steenkamp & Baumgartner, 1998). That is, measurement equivalence refers to the extent in different groups (e.g. Aboriginal or non-Aboriginal Australians) the construct is being measured *in the same way*. A more operational definition of measurement equivalence is whether respondents with *identical levels of the construct* from different cultures will *on average* have *identical scores* when evaluated by the instrument (Raju, Laffitte, & Byrne, 2002). The equivalence of measures is important in cross-cultural research precisely because, if scores obtained with the same instrument from two or more distinct groups will be compared, the instrument must display measurement equivalence (Milfont & Fischer, 2015).

In summary, considering that psychological constructs differ from one culture to another, these four procedures are required to ensure that psychological instruments from one culture can be adapted to another and that they will measure what they are supposed to measure (i.e. they are valid). In this PhD research, the recommended steps 2 to 4 were conducted (step 1 was skipped since the instruments were adapted rather than developed). That is, we adapted the wording of the instruments to better reflect the Aboriginal culture, the instruments were piloted in an Aboriginal Reference Group to discuss face and content validity, and later applied to a large sample from which psychometric analysis was conducted. Particularly, the discussions with the Aboriginal Reference Group were the reason why the Perceived Stress Scale, Social Support Scale and Sense of Personal Control Scale were chosen to be adapted for Aboriginal Australians. Compared to other available measures, the Aboriginal Reference Group believed that these three instruments could potentially reflect their experiences of stress, social support and sense of personal control. In the next chapters, we will detail the adaption of the Perceived Stress Scale, Social Support Scale and Sense of Personal Control Scale and discuss in-depth why these instruments can capture the Aboriginal experience of stress, social support and personal control in contemporary Australia.

Chapter 2 – General aims and specific objectives

Research questions

Is the Perceived Stress Scale (PSS) a valid and reliable measure for an Aboriginal Australian population?

Is the Social Support Scale (SSS) a valid and reliable measure for an Aboriginal Australian population?

Is the Sense of Personal Control Scale (SPCS) a valid and reliable measure for an Aboriginal Australian population?

General Aim

The aim of this thesis was to assess the validity and reliability of the Perceived Stress Scale (PSS), Social Support Scale (SSS) and Sense of Personal Control Scale (SPCS) among an Aboriginal Australian population.

Specific objectives

Objectives: Evaluate the validity and reliability of the Perceived Stress Scale (PSS) in an Aboriginal Australian culture.

Evaluate the validity and reliability of the Social Support Scale (SSS) in Aboriginal Australian culture.

Evaluate the validity and reliability of the Sense of Personal Control Scale (SPCS) in an Aboriginal Australian culture.

Main sample and measures

The Perceived Stress Scale (PSS), Social Support Scale (SSS) and Sense of Personal Control Scale (SPCS) were applied in the baseline questionnaire of the Baby Teeth Talk Study, a randomized controlled trial aimed at reducing early childhood caries among an Aboriginal population in South Australia. Participants were women who identified as being pregnant with an Aboriginal child during the study's recruitment phase (Feb 2011 to May 2012). The study received approval from the University of Adelaide Human Research Ethics Committee, the Aboriginal Health Council of South Australia, the Government of South Australia and the Human Research Ethics Committees of three participating South Australian hospitals. Participants were recruited through referrals from a variety of sources including community services, Indigenous groups, and hospitals. Potential participants were provided with information about the study from health services providers and the study staff. For those interested in participating, the researchers explained the project in detail and answered questions. All participants were informed that participation was voluntary and that they could refuse or withdrawn to participate at any stage without having to provide a reason. The individuals who decided to participate were then asked to complete and sign a form indicating consent. All participants provided signed informed consent.

The PSS is an instrument widely used to measure perceived stress, which evaluates if a person's life is perceived as 'unpredictable, uncontrollable, overloading,' being comprised of 14-items in its original version. The SSS evaluates through 4 items the emotional, appraisal, instrumental and informational domains of social support. The SPCS is a 12-item measure that assesses one's sense of personal control through "mastery," individual's beliefs regarding the ability to influence outcomes, and “perceived constraints,” how outcomes are believed to be determined by external factors. All instruments were responded to on a five-point rating scale (1 = Not at all, 2 = Rarely, 3 = Sometimes, 4 = Fairly often, 5 = Very often).

Chapter 3 – Stress beyond coping? A Rasch analysis of the Perceived Stress Scale (PSS-14) in an Aboriginal population

Statement of Authorship

Title of Paper	Stress beyond coping? A Rasch analysis of the Perceived Stress Scale (PSS-14) in an Aboriginal population
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Santiago, P. H. R., Roberts, R., Smithers, L. G., & Jamieson, L. (2019). Stress beyond coping? A Rasch analysis of the Perceived Stress Scale (PSS-14) in an Aboriginal population. PloS one, 14(5), e0216333.

Principal Author

Name of Principal Author (Candidate)	Pedro Henrique Ribeiro Santiago		
Contribution to the Paper	Conceptualized the idea, performed the analysis, wrote the primary draft and acted as corresponding author.		
Overall percentage (%)	80%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	21/11/2019

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);
ii. permission is granted for the candidate in include the publication in the thesis; and
iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Rachel Roberts		
Contribution to the Paper	Conceptualized the idea, supervised development of work, provided intellectual contribution, and critically reviewed the draft manuscript.		
Signature		Date	21/11/2019

Name of Co-Author	Lisa Smithers
Contribution to the Paper	Conceptualized the idea, supervised development of work, provided intellectual contribution, and critically reviewed the draft manuscript.
Signature	Date: 21/11/2019

Name of Co-Author	Lisa Jamieson
Contribution to the Paper	Conceptualized the idea, supervised development of work, provided intellectual contribution, and critically reviewed the draft manuscript.
Signature	Date: 21/11/2019

Linkage to the body of work

In Australia, the risk of being exposed to stressful life events is two to five times greater for Aboriginal compared with non-Aboriginal individuals (Kelly, Dudgeon, Gee, & Glaskin, 2009). For example, one in five Aboriginal youth reported living in a family confronted with at least seven major stressful life events over the last year, such as death, arrest, and alcohol abuse (Amarasena et al., 2015; Blair, Zubrick, & Cox, 2005). One of the most affected groups is pregnant Aboriginal women (Prandl, 2017), who have a two to three-fold increase in the odds of experiencing stressful events in pregnancy relative to non-Indigenous pregnant women (Weetra et al., 2016). In a recent study, the prevalence of Aboriginal women that experienced stressful life events during pregnancy ranged from 36% to 58% (Mah et al., 2017).

Considering the high prevalence of stress in the Aboriginal population, it is necessary to measure perceived stress with instruments that are culturally-appropriate for Aboriginal people. This study evaluated the psychometric properties of the Perceived Stress Scale (PSS-14), the most widely used instrument to measure perceived stress, in an Aboriginal population. The PSS-14 was initially examined by a 15-member Aboriginal Reference Group, comprising Aboriginal community members and Aboriginal Infant Care workers, who evaluated item wording and indicated that the instrument had content and face validity for Aboriginal Australians. The result of this study was the aPSS-13, the adapted Perceived Stress Scale for Aboriginal Australians. Implications for research and practice are provided.

Highlights

- The findings indicated initial evidence of the 13-item adapted version of the Perceived Stress Scale (aPSS-13) as a valid and reliable instrument to measure perceived stress in Aboriginal Australians.

- The Perceived Stress Scale (aPSS-13) was composed of two subscales, Perceived Stress and Perceived Coping. Therefore, total scores need to be computed for each subscale independently instead of a total score across all items.
- In contrast with previous PSS validations in non-Indigenous groups, the latent correlation between Perceived Stress and Perceived Coping in Aboriginal Australians was weak. One possible explanation is that social inequalities experienced by the Aboriginal population are so pronounced that even Aboriginal pregnant women that perceived themselves as coping well with life challenges ended up endorsing items regarding high levels of stress.
- Considering the limited nature of the sample, comprised of pregnant Aboriginal women, future studies need to replicate these results in samples including Aboriginal men.

Research and Policy Implications

- The adapted Perceived Stress Scale (aPSS-13) should be used to measure perceived stress in Aboriginal Australians instead of other versions of the Perceived Stress Scale (i.e. PSS-14, PSS-10). Total scores should be computed independently for the Perceived Stress and Perceived Coping subscales.

PLOS ONE

RESEARCH ARTICLE

Stress beyond coping? A Rasch analysis of the Perceived Stress Scale (PSS-14) in an Aboriginal population

Pedro Henrique Ribeiro Santiago[1]*, Rachel Roberts[2], Lisa Gaye Smithers[3], Lisa Jamieson[1]

1 Australian Research Centre for Population Oral Health (ARCPOH), Adelaide Dental School, The University of Adelaide, Adelaide, South Australia, Australia, **2** School of Psychology, The University of Adelaide, Adelaide, South Australia, Australia, **3** School of Public Health, The University of Adelaide, Adelaide, South Australia, Australia

* pedro.ribeirosantiago@adelaide.edu.au

OPEN ACCESS

Citation: Santiago PHR, Roberts R, Smithers LG, Jamieson L (2019) Stress beyond coping? A Rasch analysis of the Perceived Stress Scale (PSS-14) in an Aboriginal population. PLoS ONE 14(5): e0216333. https://doi.org/10.1371/journal.pone.0216333

Editor: Karl Bang Christensen, University of Copenhagen, DENMARK

Received: August 31, 2018

Accepted: April 18, 2019

Published: May 3, 2019

Data Availability Statement: Data cannot be shared publicly because of its sensitive nature. The study participants constituted a significant proportion of the Aboriginal and Torres Strait Islander community in South Australia and the release of data could lead to the participants' identification. Data are available from the Aboriginal Research Advisory Committee of the Indigenous Oral Health Unit (Email: iohu@adelaide.edu.au. Phone: +61 8 8313 4611) for researchers who meet the criteria for access to confidential data.

Abstract

The history of colonization contributed to Aboriginal and Torres Strait Islanders becoming one of the most disadvantaged groups in Australia. The multiple social inequalities, and therefore the constant insecurities for many about low income, poor living conditions, unemployment, and discrimination, generate chronic stress in this population. In the Baby Teeth Talk Study, an oral-health randomized controlled trial, the Perceived Stress Scale (PSS-14) was administered to 367 pregnant Aboriginal women at baseline. The aim of the present study was to evaluate the validity and reliability of the PSS-14 in an Aboriginal population. The study analysed: (a) model fit; (b) dimensionality; (c) local dependence; (d) differential item functioning; (e) threshold ordering and item fit; (f) targeting; (g) reliability; and (h) criterion validity. The dimensionality analysis indicated a two-factor structure, with negatively and positively worded items clustering together and 21.7% (95% Agresti-Coull C.I. [17.8%, 26.2%]) statistically significant t-tests between the persons' estimates. After the creation of composite items, the revised Perceived Distress (χ^2 (21) = 11.74, p = 0.946) and Perceived Coping (χ^2 (28) = 17.63, p = 0.935) subscales fitted the Rasch model. Reliability was modest ($PersonSeparationIndex_{distress}$ = 0.72; $PersonSeparationIndex_{coping}$ = 0.76). The latent correlation between the Perceived Distress and Perceived Coping subscales was r = 0.14. It is hypothesized that the social inequalities experienced by the Aboriginal population are so pronounced that even Aboriginal pregnant women that perceived themselves as coping well with life challenges ended up endorsing items regarding high levels of stress. The present research showed that a revised PSS-14 is a culturally valid and modestly reliable psychological instrument to measure stress in a population of pregnant Aboriginal women in Australia.

Introduction

A history of colonization and genocide contributed to Aboriginal and Torres Strait Islanders becoming one of the most disadvantaged groups in Australia [1]. The multiple social

Funding: The Baby Teeth Talk RCT was funded by an International Collaborative Indigenous Health Research Partnership grant from the National Health and Medical Research Council of Australia (NHMRC, Project Grant 627350).

Competing interests: The authors declare that there are no conflicts of interest.

inequalities, and therefore the constant insecurities for many about low income, poor living conditions, unemployment, and discrimination, generate chronic stress and impact on both physical and mental health [2]. The risk of being exposed to stressful life events is two to five times greater for Aboriginal Australians [3]. For example, one in five Aboriginal youth reported living in a family confronted with at least seven major stressful life events in the last year, such as death, arrest, incarceration, alcohol abuse, among others [4, 5]. When the stress is experienced by Aboriginal persons in childhood, it has been associated with an increase in the odds of depression, anxiety, dementia, and Alzheimer disease during the life-course [6].

One of the most affected groups is pregnant Aboriginal women [7], who have a two to three-fold increase in the odds of high psychological distress associated with experiencing stressful episodes (e.g. family violence) during pregnancy [8]. In a recent study, the prevalence of Aboriginal women that experienced stressful life events while pregnant ranged from 36% to 58%. For example, 49% of the Aboriginal women reported witnessing the death of at least one family member or close friend during their gestation [9]. Rates of teenage pregnancy were also higher among Aboriginal woman, with the low socioeconomic position of many teenage Aboriginal mothers making them particularly vulnerable to stressful situations [10]. Prenatal stress has been associated with negative child outcomes such as low birth weight and reduced length of pregnancy [11], alongside with increased risk of impaired cognitive development [12]. In-utero exposure to the stress hormone cortisol has also been associated with lower adult educational attainment and verbal cognition [13].

The stress experienced by Aboriginal Australians needs to be contextualized in relation to the long-term effects of colonization on their sense of personal control and the social support of their communities. The theoretical associations between stress, sense of personal control and social support have been extensively investigated in the general population. For example, when individuals have a high sense of personal control (i.e. the belief about being able to influence outcomes in life) and encounter stressful events, they are more prone to engage in active efforts to reduce the demands, instead of more passive or avoidant emotion-focused coping strategies [14]. Furthermore, when a person is faced with multiple problems that might exceed her/his individual coping capability, having social support leads to a perception that others will provide the necessary resources and attenuates the stress reaction [15]. These associations are especially relevant for understanding the stress experienced by Aboriginal Australians in the contemporary world. The intended dismantling of Aboriginal culture and the undermining of self-determination with respect to political and social decisions led individuals to lose the feeling of personal control over their lives. Previous research among Aboriginals living in Arnhem land, an isolated Aboriginal reserve in the northeast of Australia, has shown that perceived stress was negatively associated with mastery (i.e. the ability to influence outcomes, one important aspect of personal control) [1]. Additionally, the forced separations caused by decades of assimilation policies have undermined their societal cohesion. For Aboriginal Australians, individual emotional well-being cannot be disentangled from their connection with the well-being of the whole community. The social support derived from a complex kinship system reinforces Aboriginal persons' cultural identity and protects against discrimination and negative interactions with non-Indigenous persons [16, 17].

Notwithstanding the high exposure to stressful life events, the investigation of its impact on Aboriginal and Torres Strait Islander well-being has been faced with conceptual and methodological challenges [18]. One fundamental concern is the development of psychological instruments that are valid for this group [19]. This concern was initially raised by Aboriginal leaders, who brought awareness to the fact that their experiences of happiness, suffering, and well-being should not be assumed as equivalent to the western conceptualization of "mental health" and needed to be understood through a culturally-sensitive framework [16, 19].

The usual practice in research has been the application of mainstream Western psychological instruments to Aboriginal Australians to investigate associations between the test results and health outcomes without a comprehensive evaluation of test psychometric functioning [19]. The assumption is that the psychological instrument applied in the Aboriginal population will retain its psychometric properties and be equally valid; however, if the construct being measured has meaningful differences in the Aboriginal culture, the interpretation of item responses is subject to construct bias [20, 21].

The Perceived Stress Scale (PSS)

One psychological instrument that has been applied to measure stress in Indigenous populations is the Perceived Stress Scale (PSS). The PSS is the most widely used instrument to measure perceived stress [22], comprising 14 items in its original version (PSS-14) [23] and was developed based on the theoretical perspective of Lazarus [24]. The notion of perceived stress, which differs from earlier views that focused on the biological aspects of stress (e.g. secretion of corticoids and other hormones as a reaction to stressors) [25], postulates that stressor effects occur only when a situation is both *appraised as threatening* and there is a *perception of insufficient coping resources* [24, 26]. The PSS has been previously administered to Indigenous groups in Canada [27, 28] and the United States [29, 30]. However, in all cases, the PSS was applied without an evaluation of construct validity for the Indigenous culture and in one study it was simply assumed due to being "extensively validated" elsewhere [29].

With respect to non-Indigenous populations, the construct validity and psychometric properties of the PSS have been well-established (e.g.: USA, Japan, Mexico, France, Brazil, China, Qatar, among others) [31]. The cumulative evidence from psychometric studies indicates that the PSS (PSS-14 and PSS-10) has a two-factor structure composed of the positively-worded and negatively-worded items [22]. The two-factor dimensionality is consistent with Lazarus's [24] theory of stress as appraisal and coping [32], and has been interpreted as "Perceived Distress" and "Perceived Coping" [33], although other terminologies such as "Perceived Helplessness" and "Perceived Self-efficacy" have been used [22]. Notwithstanding that Cohen [34], in the first factor analysis of the PSS-14, stated that for "purposes of measuring perceptions of stress, the distinction between the two factors was considered irrelevant", the majority of subsequent empirical studies found evidence of two distinct psychological factors [22]. For example, while both factors predicted depression in women, only the "Perceived Helplessness" factor had predicted depression in men [33]. The PSS scores have been consistently correlated with depression and anxiety [31], and recent studies added to the evidence that the magnitude of these associations differs according to each dimension [35, 36].

Despite the large body of evidence, the two-factor dimensionality of the PSS is not a consensus among researchers [22] and has been challenged in recent studies [32, 37, 38]. In relation to the PSS-14, for example, the main concern is that the two-factor solution consistently accounted for less than 50% of the total variance [31]. Two studies found that a bi-factor structure with a specific factor for the "Perceived Coping" items better explained the items' covariance than the two-factor structure [32, 37]. In addition, Medvedev, Krägeloh [38] showed that, after resolving local dependence by creating composite items, the PSS-10 exhibited a unidimensional structure.

Regarding internal consistency, the Cronbach's alpha of the PSS-14 was >.70 in 11 of 12 examined studies and test-retest reliability was >.70 in 2 of 3 studies[31]. Six studies have also investigated Differential Item Functioning (DIF). The evidence of DIF related to gender is mixed and some studies indicated no gender-related DIF [22, 39]. Cole [40] reported that the PSS-10 items 3, 6, 7, 8 and 10 had DIF by gender and this result was partially confirmed by

Gitchel, Roessler [41] who found DIF for the same items 3 and 4, in addition to items 1 and 6. Regarding education, evidence of DIF was found for PSS-10 items 3, 4, 8 and 9 [40], and the PSS-10 items 3 and 4 showed evidence of DIF by age [39]. Other sources of DIF investigated by previous studies were ethnicity and literacy [40, 42].

The present research

In South Australia, the Perceived Stress Scale (PSS-14) was completed by 364 pregnant Aboriginal women in the Baby Teeth Talk Study (BTT). The Baby Teeth Talk Study, whose name was given by the study's Aboriginal Reference Group, is a randomized controlled trial that aimed to investigate if implementing a culturally-sensitive intervention would reduce early childhood caries of Aboriginal children [43]. The study received approval from the University of Adelaide Human Research Ethics Committee, the Aboriginal Health Council of South Australia, the Government of South Australia and the Human Research Ethics Committees of three participating South Australian hospitals.

Considering the disproportionate rates of stressful life events experienced by Aboriginal Australians [3] and Aboriginal women [7, 8], it is necessary to develop or adapt psychological instruments that measure stress and are culturally valid for this group. In the context of the BTT randomized controlled trial, the results of the PSS-14 were not used to modify the intervention and had no direct consequences for the Aboriginal women (i.e. it was not a high-stakes scenario). The intention was to use the PSS-14 results to measure stress on a group rather than individual level and inform the effects of stress (i.e. the exposure) on future outcomes (e.g. children caries at age 2, nutrition, among others). Nonetheless, the aim of validating the PSS-14 is that future applications can measure stress in Aboriginal Australians in a variety of different contexts. To the best of our knowledge, there are no studies that have evaluated the psychometric properties of the PSS in any Indigenous culture, including an Aboriginal and Torres Strait Islander population. The aim of the present study is to evaluate if the PSS-14 is a valid and reliable measure of perceived stress in an Aboriginal population, using the Rasch Measurement Model [44]. The hypothesis that the PSS-14 is a valid and reliable measure will be evaluated through the analysis of the: (a) model fit; (b) dimensionality; (c) local dependence; (d) differential item functioning; (e) threshold ordering and item fit; (f) targeting; (g) reliability; and (h) criterion validity. The criterion validity will be evaluated by inspecting convergent and divergent validity of the PSS-14 with respect to the complementary measures of Sense of Personal Control Scale (SPCS) and the Social Support Scale (SSS) and concurrent validity with health behaviors. These complementary measures were chosen due to the role of these two theoretical constructs (i.e. sense of personal control and social support) in the stress experienced by Aboriginal Australians in the contemporary world.

Methods

Participants and procedures

The sample comprised of 367 pregnant Aboriginal women living in South Australia. Participants were recruited through referrals from a variety of sources including community services, Indigenous groups, and hospitals. Potential participants were provided with information about the study from health services providers and the study staff. For those interested in participating, the researchers explained the project in detail and answered questions. All participants were informed that participation was voluntary and that they could refuse or withdrawn to participate at any stage of the study without justification. The individuals who decided to participate were then asked to complete and sign a form expressing consent [45]. All participant provided signed informed consent.

The mothers' average age was 24.9 years (SD = 5.9, range = 14–43). Regarding education, approximately 3% of mothers had not attended school or had only attended primary school, 70% of mothers had completed secondary school, 20% had completed Technical and Further Education (TAFE) or Trade qualifications, and 7% had completed university or post-graduate degrees. TAFE is the biggest provider of post-secondary education in Australia. Unlike universities, which are composed mostly of full-time students, TAFE institutions allow students to combine study and work, and encourage programs of apprenticeships and traineeships [46]. In respect to socio-economic position, approximately 53% of mothers were on the lowest quintile of the Index of Relative Socio-Economic Advantage and Disadvantage [47] and only 6% were on the top two quintiles, indicating a population that was largely socio-economically disadvantaged. Among the mothers, approximately 51% were current tobacco smokers, 27% former smokers and 22% had never smoke. Additionally, 10% reported currently drinking alcohol, 82% said they used to drink alcohol and 8% have never drunk alcohol. The psychological instruments were administered as part of a broader questionnaire to the Aboriginal mothers by four research staff (three Indigenous and one non-Indigenous) at baseline, and the data was later entered into password-protected databases.

Measures

Initial pilot testing. As a recommended procedure in cultural adaptation [48], the PSS-14, SPCS and SSS were initially discussed and administered to a 15-member Aboriginal Reference Group, comprising Aboriginal community members and Aboriginal Infant Care workers. The reference group evaluated the items' wording to ensure content and face validity with respect to an Aboriginal culture and suggested minor modifications. The modifications made, and the subsequent new wording of all PSS-14 items can be found in Table 1.

Subsequent to the pilot testing, the measures were then applied at the study baseline.

Primary measure. Perceived Stress Scale (PSS): PSS-14 items were responded to on a five-point Likert scale (0 = Not at all, 1 = Rarely, 2 = Sometimes, 3 = Fairly often, 4 = Very often). The PSS-14 was responded by 364 out of the 367 pregnant women. The positively worded items were reverse-scored prior to the analysis.

Table 1. New wording of the PSS-14 items after adaptation for an Aboriginal culture.

Item number	Item content*
1	. . . felt upset because of something that happened?
2	. . . felt like you couldn't control the important things in your life?
3	. . . felt nervous or stressed?
4	. . .dealt well with life hassles?
5	. . . coped well with important changes in your life?
6	. . . felt able to handle your personal problems?
7	. . . felt things were going your way?
8	. . . felt unable to cope with all the things that you had to do?
9	. . . felt able to control irritations in your life?
10	. . . felt you were on top of things?
11	. . . felt angered because of things that happened outside of your control?
12	. . . found yourself thinking about all the things that you have to do?
13	. . . felt able to control how you spend your time?
14	. . . felt troubles were piling up so high that you could not deal with them?

Note.
*Every item started with the sentence "How often during the LAST YEAR have you. . .."

https://doi.org/10.1371/journal.pone.0216333.t001

Complementary measures. The Sense of Personal Control Scale (SPCS): The Sense of Personal Control Scale is a 12-item scale developed by Lachman [49] to measure an individual's sense of personal control. The original factorial analysis identified the two dimensions of Mastery (MS), individual's beliefs regarding the ability to influence outcomes, and Perceived Constraints (PC), how outcomes are believed to be determined by external factors [49]. The items were responded to on a five-point rating scale (1 = Not at all, 2 = Rarely, 3 = Sometimes, 4 = Fairly often, 5 = Very often). The SPCS has been validated for an Aboriginal population [50].

The Social Support Scale (SSS): The Social Support Scale is composed of 4 items, each one designed to evaluate the emotional, appraisal, instrumental and informational domains of social support as theorized by House [51]. The items were responded to on a five-point rating scale (1 = Not at all, 2 = Rarely, 3 = Sometimes, 4 = Fairly often, 5 = Very often). The SSS has also been previously validated for an Aboriginal population [50].

Alcohol drinking and smoking status: Alcohol drinking status was measured with a single question "Alcohol drinking status" with three response categories (1 = Currently drink alcohol, 2 = Used to drink alcohol, 3 = Have never drunk alcohol). Similarly, smoking status was measured with the question "Smoking status" (1 = Currently smoke, 2 = Used to smoke, 3 = Never smoked). In the analysis of criterion validity, the categories of "Used to. . ." and "Never. . ." were collapsed since they represented the participants who weren't currently smoking or drinking.

Statistical methods

The Rasch measurement model. The Rasch Model is part of the family of Item Response Theory (IRT) psychometric models [52]. The Rasch model for ordered response categories [53] is a generalization of the original model for dichotomous items [44] and is displayed in Eq 1.

$$P(X_{vi} = x|\Theta = \theta) = \frac{e^{\theta_v x - \sum_{z=1}^{x} \beta_{iz}}}{\sum_{h=0}^{mi} e^{\theta_v h - \sum_{z=1}^{h} \beta_{iz}}} \tag{1}$$

Eq 1 indicates that the probability of the person v endorsing category x (e.g. "Strongly Agree") of item i with a total of m categories is a function of the person's latent trait θ and the item thresholds β_{iz}. The *item thresholds* parameters β_{iz} are conceived in terms of the latent trait θ and represent in the latent trait scale the point of equal probability of choosing between two adjacent categories (e.g. "Agree" and "Strongly Agree"). Therefore, in the polytomous Rasch Model, the probability of endorsing the category of an item is a function of characteristics of the person (the latent trait θ) and characteristics of the item (the item thresholds β_{iz}) [54].

The development of the Rasch Model aimed to establish in social sciences the fundamental properties of measurement found in the natural sciences [55]. One of the fundamental properties of measurement is invariance. In natural sciences, the measurement of processes that can be observed (e.g. height) or cannot be directly observed (e.g. temperature) is independent of the measurement instrument (e.g. ruler/thermometer) being used [56]. The Rasch Model is a psychometric model that has an analogous property (named objectivity): the comparison of two individuals regarding a latent trait, such as the *difference of the amount* of stress experienced by two persons, is independent of the psychological instrument being used [57]. The objectivity is derived from a unique characteristic of the Rasch Model in comparison to other IRT models, that it has a statistically sufficient total score [58]. A statistic S(X) (e.g. the total score) of a vector of observable variables X (e.g. the item responses) is considered sufficient for a parameter θ (e.g. the latent trait) if X and θ are conditionally independent given S(X) [59]. This means that the total score, which is the summation of the item responses, encompasses all the available information for statistical inference regarding the respondent's latent trait; and, for example, no information would be added by knowing the respondent's characteristics (e.g. male or female, smoker or non-smoker) [60]. Since the parameters of the Rasch model are mathematically symmetrical [61], the functioning of the items (i.e. their difficulty and hierarchical order) is also independent of the population [57].

The invariance of persons' and items' parameters is an *intrinsic mathematical property* of the Rasch model and one of its main requirements. Therefore, what is investigated is if the measurement of stress by the PSS-14 approximates the fundamental properties of measurement by satisfying the *a priori* measurement requirements of the Rasch Model. Therefore, when the measurement requirements encompassed in the Rasch Model are not achieved, it is the discrepancies that will inform problems with the items and the psychological instrument [62]. In summary, the items fitting a Rasch model exhibit the measurement properties of (1) unidimensionality: the items measure a single psychological trait (i.e. perceived stress); (2) monotonicity: endorsement of "higher" categories are an increasing function of the latent trait; (3) homogeneity: the item difficulty order is the same for all respondents; (4) local independence: items are conditionally independent given the latent trait; and (5) absence of differential item functioning: items are conditionally independent of exogenous variables given the latent trait [52, 58].

Statistical analysis. The Rasch analysis was conducted with the RUMM2030 software [63]. The person estimates scatterplot was adapted from Winsteps software [64]. Descriptive statistics and Agresti-Coull binomial 95% Confidence Intervals (C.I.) were computed with R software [65].

The first stage in the analysis was to determine model parametrization and investigate model fit. The estimation method used was Pairwise Conditional Maximum Likelihood Estimation. The advantage of Pairwise Conditional Maximum Likelihood Estimation over Conditional Maximum Likelihood Estimation is that the estimates of each item parameter is a function of frequencies in all categories rather than just a function of frequencies of the adjacent categories. Therefore, the estimates of item parameters are less affected by categories with a small frequency of responses [66]. The model was initially estimated with an unrestricted parametrization (i.e. Partial Credit model). However, considering that the distances between item thresholds might be equal across items (i.e. Rating Scale model) in certain cases, a likelihood-ratio test was conducted to determine whether an unrestricted parametrization (Partial Credit model) would provide a better description of the item responses in comparison with a restricted parametrization (Rating Scale model) [67, 68]. Missing data for individual items ranged from 0.0% to 1.1%, and missing was 1.2% when considering all items. The impact of missing values was unsubstantial [69] and a Rasch analysis with missing values has been shown to outperform multiple imputation under most conditions [70, 71]. Therefore, the Rasch analysis with missing values was conducted. Additionally, in the conditional framework, a sample size around 390 participants is sufficient to detect a model deviation in item difficulty of 1 logit with a significance level of 5% and power of 80%. However, since these samples requirements are influenced by several factors including targeting [72], other authors have argued that samples with more than 250 participants provide enough power for most practical purposes [73]. The overall model fit was evaluated with a summary item-trait χ^2 statistic, which is calculated through the summation of the χ^2 of all individual items [63].

The second stage was the analysis of dimensionality. The dimensionality was evaluated through a Principal Component Analysis (PCA) of the residuals. In this procedure, two item sets were identified according to their opposite loadings on the first residual factor. The persons' estimates of these two item sets were compared through independent t-tests. If the data is unidimensional it is expected less than 5% of these differences to be statistically significant [74]. In addition, an Agresti-Coull binomial 95% CI was computed for the proportion of statistically significant differences [75, 76]. To investigate the strength of the association between the persons' estimates, the disattenuated correlation coefficient was calculated [77]. When multidimensionality was detected, the Rasch model was applied to each subscale independently to assess validity and reliability. In addition, a subtest analysis was conducted to evaluate the magnitude of the multidimensionality through the estimation of the latent correlation between subscales. The latent correlation is a single effective summary about how the distinct latent traits are associated and is estimated based on the change in the reliability index when subscales are analysed separately compared to when they are analysed combined [63].

The third stage was the evaluation of response dependence [78]. Response dependence was evaluated through the correlation matrix of standardized residuals [79]. The magnitude of the observed residual correlations is considered *relative to the overall set of correlations*, and higher values *compared to the average residual correlation* indicate that the items responses were not completely accounted by the psychological trait. Therefore, it is also presented the adjusted residual correlations, which are the differences between the observed residual correlations and the average residual correlations. In addition, a cut-off point of 0.15 for the adjusted residual correlations from simulation studies that evaluated polytomous items in a sample size of 350 participants was used [80, 81]. Response dependence was resolved through the creation of composite items (i.e. "super items"). To create composite items, the individual items were summated [82, 83].

The following stage was the analysis of Differential Item Functioning [84]. In this study, the investigation of uniform and non-uniform DIF was conducted statistically, through a two-way ANOVA of the residuals [85] with the calculation of η^2 and partial η^2 as measures of effect size [86]; and graphically, by diving the sample based on the trait level into adjacent class interval (CI) and plotting the average observed item responses against the model theoretical expectations indicated by the Item Characteristic Curves (ICCs). Uniform DIF refers to when the magnitude of the conditional dependence between item responses and exogenous variable given the latent trait is constant across the trait level. When the magnitude of the conditional dependence between item responses and exogenous variable given the latent trait varies across the trait level, it is said that the item has non-uniform DIF [87]. Ideally, the sample would be divided into a unique group for each possible total score (and, therefore, person estimate), but since it is unlikely to have the necessary amount of respondents with the same total score for each possible score, class intervals (CI) were created [63, 88]. The characteristics analysed for DIF were age group (14 to 20 years old; 21 to 30 years old; 31 years old or more); education (education level up to High School; TRADE, TAFE, or University); socioeconomic position (1^{st}, 2^{nd}, 3^{rd}, 4^{th} and 5^{th} quintiles of the IRSAD); and smoking status (Never smoked tobacco; Used to smoke tobacco; and Currently smoke tobacco). The choice of age and education was due to DIF of PSS items been previously reported [39, 40]. It was also hypothesized that socioeconomic position and smoking status could be additional sources of DIF. In case DIF was present, it was solved by splitting the item into group specific items [85]. Considering that in the analysis of DIF multiple null hypothesis significance testing (NHST) were performed, to avoid the increase in the probability of performing a Type 1 error, a Bonferroni adjustment with $\alpha = 0.01$ was applied [89].

The fifth stage was the analysis of item threshold ordering [90]. It was investigated if the thresholds positions were ordered and if every category became the most probable for a definitive range of the latent trait. In case disordering was found, adjacent categories were considered to be collapsed [68]. The only exception was composite items since the disordering is a consequence of local dependence and collapsing categories is not necessary [58].

Item misfit was analysed only after evidence of unidimensionality, local independence, and absence of DIF was established. If these three conditions are violated, the idea that separate items should fit a Rasch model loses its meaning [61]. The assessment of fit in the Rasch Model was conducted using a "family approach" [91] through statistical and graphical evaluation of the Item Characteristic Curves. Statistically, the *magnitude of* item misfit was evaluated with the Fit Residual statistic, with the value of 0 meaning fit of the data to the Rasch model and acceptable values between -2.5 and 2.5 [63, 88]. Additionally, the χ^2 statistic was used to evaluate item-trait interaction and the *probability* of the misfit occurring due to sampling variation [63, 85]. To consider an item as misfitting, the misfit needed to be flagged by both statistics [92] and a Bonferroni adjustment with $\alpha = 0.01$ was applied [89]. Similarly to the DIF analysis, the graphical analysis was conducted by evaluating the average observed item responses in each CI compared to the expected item responses.

The subsequent stage of the analysis was the evaluation of targeting. The analysis investigated the mean of the persons' parameters [76], and a value between ±0.5 logits indicates optimal targeting [93]. The internal consistency reliability was analysed with the Person Separation Index (PSI), an index analogous in construction to the Cronbach's α [94] but calculated using the persons' parameters rather than the raw scores.

The last stage of the analysis was the investigation of criterion validity. Criterion validity was evaluated by: a) inspecting convergent and divergent validity of the latent trait with theoretically relevant constructs of its nomological network [95]; and b) concurrent validity with respect to health behaviours. For the analysis of convergent and divergent validity, the selected constructs were sense of personal control and social support. It was expected a negative correlation of perceived stress with MS and SSS, and a positive correlation with PC. Since the PSS, SPCS and SSS scores are ordinal, the non-parametric Kendall's τ was used [96]. A Bonferroni adjustment was applied for the calculation of 95% C.I. [89].

For the analysis of concurrent validity, the mean of the total scores for each subscale was used to dichotomize the participants into lower perceived distress/higher perceived distress and lower perceived coping/higher perceived coping. Log-binomial models were then used to evaluate if there was an effect of perceived distress (or perceived coping) on smoking status and alcohol consumption after controlling for the confounders general health and education. It was expected that participants with higher perceived distress would have a higher risk of currently smoking and drinking alcohol than participants with lower perceived distress, and participants with higher perceived coping would have a lower risk.

Results

Model parametrization and model fit: The Likelihood Ratio test (LRT) showed that the model with an unrestricted parametrization was a significantly better fit to the data compared to the model with a restricted parametrization (χ^2 (38) = 60.89, p = 0.011). Therefore, an unrestricted parametrization was chosen. After the model was estimated, the summary test-of-fit (χ^2 (98) = 386.90, p < .001) indicated no overall fit to the Rasch Model.

Dimensionality: The analysis of dimensionality conducted through a PCA of the residuals (S1 Table) showed the positively worded and negatively worded items loading on the first residual factor with an opposite factor loading valence (S2 Table).

The items were then divided into positively and negatively worded subsets and two estimates for each participant were calculated based on these two subsets of items. The estimates were then compared, which led to 21.7% statistically significant t-tests (95% Agresti-Coull C.I. [16.6%, 24.1%]) and a disattenuated correlation coefficient of 0.06. To illustrate the effects of multidimensionality, the items were also divided into subsets of even and odd items. For even

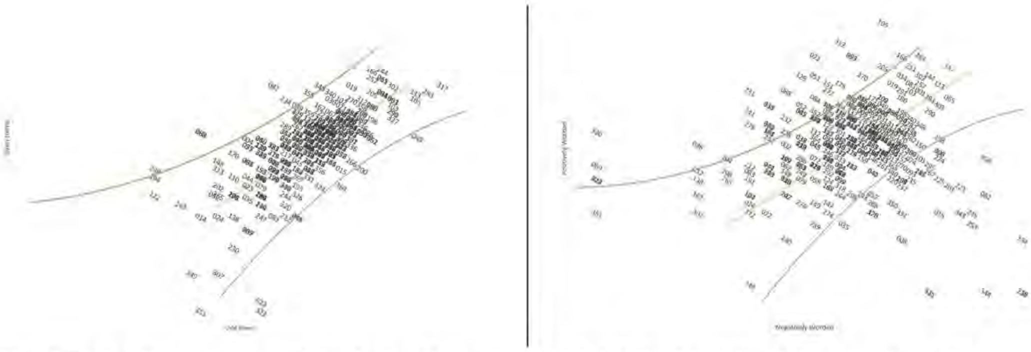


Fig 1. Scatterplot of the persons estimates after the division of the items into even and odd (left) and positively worded and negatively worded items (right). The de-identified numbers indicate the study participants. The x-axis and y-axis indicate the participants' latent trait estimated from a different subset of items (e.g. in the picture on the left, the x-axis displays the participants' latent trait according to their responses to odd-numbered items and the y-axis displays the participants' latent trait according to their responses to even-numbered items). The yellow line is the identity line and the confidence bands are the 95% C.I.

https://doi.org/10.1371/journal.pone.0216333.g001

versus odd items, there were 2.2% statistically significant t-tests between the person estimates (95% Agresti-Coull C.I. [1.0%, 4.3%]) and a disattenuated correlation coefficient of 1.0 (Fig 1).

Therefore, the 21.7% statistically significant t-tests between person estimates from each subset (95% Agresti-Coull C.I. [16.6%, 24.1%]) and the disattenuated correlation coefficient of 0.06 indicated two dimensions composed of negatively worded ("Perceived Distress") and positively worded ("Perceived Coping") items. The items were separated into the two subscales and analysed separately. The latent correlation between the Perceived Distress and Perceived Coping subscales was r = 0.14 and the score correlation between the Perceived Distress and Perceived Coping subscales was r = 0.13.

Negatively worded items ("Perceived distress")

Model parametrization and response dependence: The Likelihood Ratio test (LRT) showed that the model with an unrestricted parametrization was not a significantly better fit to the data compared to the model with a restricted parametrization (χ^2 (17) = 20.85, p = 0.233). For this reason, the model with a restricted parametrization was estimated. The summary test-of-fit (χ^2 (42) = 80.31, p<0.001) indicated no overall fit to the Rasch Model.

Table 2 shows the residual correlations matrix where there was response dependence between items 1 ("...felt upset because of something that happened?"), 2 ("...felt like you couldn't control the important things in your life?") and 3 ("...felt nervous or stressed?"), and between items 11 ("...felt angered because of things that happened outside of your control?"), 12 ("...found yourself thinking about all the things that you have to do?") and 14 ("...felt troubles were piling up so high that you could not deal with them?"). The average residual correlation was -0.16. The adjusted residual correlations surpassed the bootstrapped cut-off point of 0.15. These items were combined into two composite items and the Rasch model with an unrestricted parametrization was applied. The summary test-of-fit (χ^2 (21) = 11.74, p = 0.946) indicated overall fit to the Rasch Model.

Dimensionality and DIF: The PCA of the residuals indicated the first residual factor loadings for Composite Item 1, composed of Items 1, 2 and 3 (Λ = 0.988), Composite Item 2, composed of Items 11, 12 and 14 (Λ = -0.877) and Item 8 (Λ = -0.292). Composite Item 2 and Item

Table 2. Residual correlations of the negatively worded items.

		Item 1	Item 2	Item 3	Item 8	Item 11	Item 12	Item 14
Item 1		1						
Item 2	Obs	**0.095**	1					
	Adj	**0.255**						
Item 3	Obs	**0.043**	**0.046**	1				
	Adj	**0.203**	**0.206**					
Item 8	Obs	-0.273	-0.238	-0.296	1			
	Adj	-0.113	-0.078	-0.136				
Item 11	Obs	-0.256	-0.307	-0.221	-0.156	1		
	Adj	-0.096	-0.147	-0.061	0.004			
Item 12	Obs	-0.253	-0.344	-0.309	-0.102	**0.003**	1	
	Adj	-0.093	-0.184	-0.149	0.058	**0.163**		
Item 14	Obs	-0.288	-0.199	-0.143	-0.134	**0.039**	-0.148	1
	Adj	-0.128	-0.039	0.017	0.026	**0.199**	0.012	

Note. The residual correlations matrix displays the observed correlation between item responses after the influence of the latent trait ("Perceived Distress") was accounted by the model. It is also displayed the adjusted residual correlations, which are the differences between the observed residual correlations and the average residual correlation.

https://doi.org/10.1371/journal.pone.0216333.t002

8 were combined into a subset, and it was found 5.85% statistically significant t-tests between the person estimates (95% Agresti-Coull C.I. [3.81%, 8.82%]) and a disattenuated correlation coefficient of 1.00, indicating insufficient evidence to support multidimensionality. No uniform and non-uniform DIF were found regarding age, socioeconomic position, education and smoking status (Fig 2) (S3 Table).

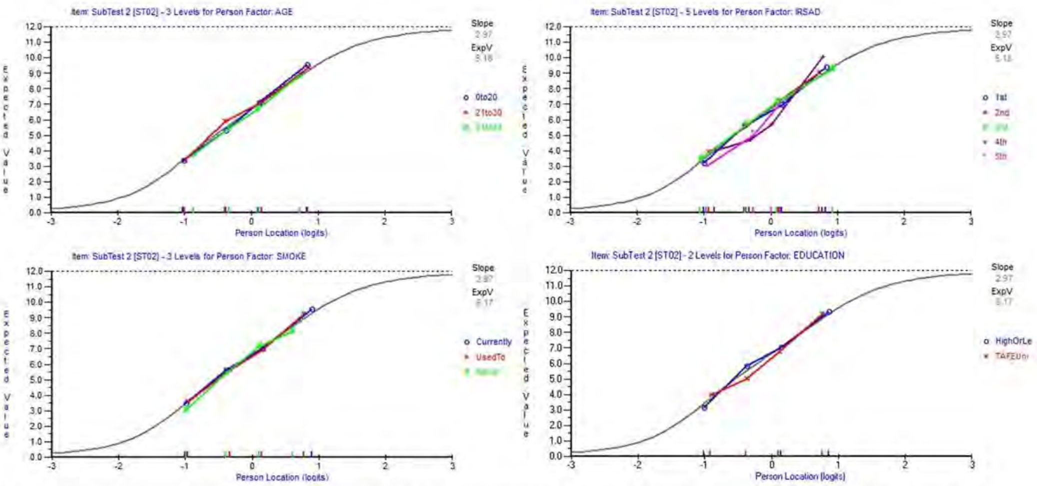


Fig 2. Analysis of Composite Item 2 DIF by Age (upper-left), Socioeconomic position (upper-right), Smoking status (bottom-left) and Education (bottom-right). Note. The colored points indicate the *average* observed item scores of each subgroup defined by an exogenous variable (e.g. the blue points in the bottom-left graph represent the average item scores of participants who currently smoked, the red points represent the average item scores of participants who used to smoke, and the green points represent the average item scores of participants who never smoked). The grey logistic curve indicates the expected item responses. The slope of the Item Characteristic Curve indicates the rate of change of the expected value with respect to the latent trait at the mid-point between the minimum and maximum scores.

https://doi.org/10.1371/journal.pone.0216333.g002

Table 3. The fit of the revised negatively worded items ("Perceived distress") to the Rasch Model.

Item[i]	Location	SE	Fit Residual	*df*	$\chi 2$	*df*	Prob
Composite Item 1	-0.030	0.030	-0.958	230.58	4.514	7	0.299
Composite Item 2	-0.159	0.030	-2.118	233.19	3.92	7	0.828
8. felt unable to cope with all the things that you had to do?	0.189	0.063	**3.104**	231.23	3.304	7	0.378

i. Every item started with the sentence "How often during the LAST YEAR have you. . ..". Note. The second column displays the items' location on the latent trait scale (i.e. the item difficulty). Values of the Fit Residual statistic indicating item misfit (i.e. lower than -2.5 or higher than 2.5), as well as statistically significant $\chi 2$ indicating misfit due to item-trait interaction, were highlighted in bold.

https://doi.org/10.1371/journal.pone.0216333.t003

Threshold ordering: Composite Item 1 and 2 had disordered thresholds and Item 8 thresholds were ordered (S4 Table). Since only composite items had threshold disordering, collapsing categories was not necessary.

Item fit: The individual item fit statistics is displayed below (Table 3). A Bonferroni adjustment of 0.0033 for the 3 items was applied.

Item 8 (Fit Residual = 3.104; $\chi 2$ (7) = 3.304, p = 0.828) had a higher Fit Residuals but was not flagged as misfitting by the χ^2-statistic. The graphical evaluation of the ICC (Fig 3) indicated that deviations of the average observed item responses in comparison to the expected item responses were unsubstantial. For these reasons, Item 8 was not considered as misfitting and was retained.

Since overall model fit was achieved and there was no more evidence of DIF, local dependence, thresholds disordering or item misfit, the measurement model for the negatively worded items ("Perceived Distress") was established.

Targeting and reliability: The Perceived Distress subscale is well targeted for the Aboriginal pregnant women. The mean of perceived distress was -0.179 logits and item thresholds covered almost all of the persons' distribution (S1 Fig). The PSI was 0.72, exhibiting a decrease when compared to the value of 0.82 before the creation of composite items.

Positively worded items ("Perceived coping")

Model parametrization: The Likelihood Ratio test (LRT) indicated that the model with an unrestricted parametrization was not a significantly better fit to the data compared to the model with a restricted parametrization (χ^2 (17) = 11.14, p = 0.849). For this reason, the model with a restricted parametrization was chosen. The summary test-of-fit (χ^2 (49) = 86.20, p = < .001) showed no overall fit to the Rasch Model.

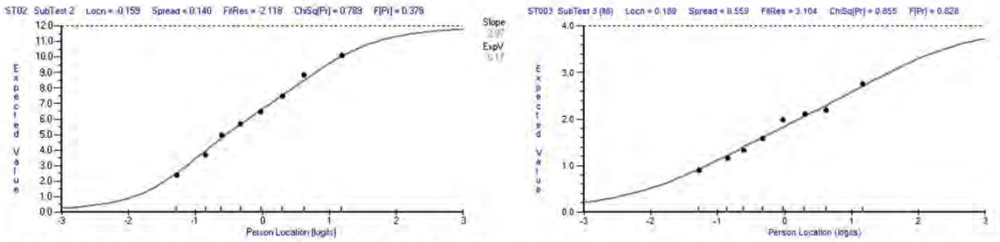


Fig 3. Item characteristic curve for Composite Item 2 (left) and Item 8 (right). Note. The x-axis indicates the latent trait and the y-axis indicates the item score. The black points represent the *average* observed item responses in each class interval. The grey logistic curve indicates the expected item responses. The slope of the Item Characteristic Curve indicates the rate of change of the expected value with respect to the latent trait at the mid-point between the minimum and maximum scores.

https://doi.org/10.1371/journal.pone.0216333.g003

Dimensionality: The PCA of the residuals showed that Item 4 (Λ = 0.725) and Item 5 (Λ = 0.537) had factor loadings with an opposite valence compared to the other items (S5 Table). After combining them into a subset and comparing with a subset composed of the other items, there were 7.71% statistically significant t-tests between the person estimates (95% Agresti-Coull C.I. [5.35%, 10.95%]) and a disattenuated correlation coefficient of 0.84. The PCA/t-test procedure indicated that item 4 and/or item 5 might have been measuring a different psychological trait, and the higher factor loading of item 4 suggested this item to be the most affected by multidimensionality.

Response dependence: The standardized residual correlation matrix (S6 Table) showed response dependence between items 5 ("...coped well with important changes in your life?) and 6 ("...felt able to handle your personal problems?), and between items 7 ("...felt things were going your way?) and 10 ("...felt you were on top of things?). The average residual correlation of positively worded items was also -0.16 and the adjusted residual correlations surpassed the bootstrapped cut-off point of 0.15.

DIF and Item fit: The analysis of DIF showed no uniform and non-uniform DIF by age, socioeconomic position, education and smoking status (S7 Table). However, misfit by item-trait interaction was consistently flagged for Item 4, when DIF was analysed by age ($\eta^2 = 0.111$, $\eta_p^{\ 2} = 0.112$, $p<0.001$), socioeconomic position ($\eta^2 = 0.106$, $\eta_p^{\ 2} = 0.110$, $p<0.001$), education ($\eta^2 = 0.106$, $\eta_p^{\ 2} = 0.107$, $p<0.001$), and smoking status ($\eta^2 = 0.099$, $\eta_p^{\ 2} = 0.100$, $p<0.001$). Finally, the analysis of item fit (S8 Table) also indicated misfit for item 4 (Fit Residual = 6.083; χ^2 (7) = 52.751, $p<0.001$). Although the $\eta_p^{\ 2}$ found according to Cohen's [97] benchmarks for $\eta_p^{\ 2}$ is a medium effect size, interpretation of what is a "medium" effect size differs according to the field, and the Fit Residual more clearly illustrates the magnitude of Item 4 misfit. The Item Characteristic Curve displayed a clear pattern of under discrimination (Fig 4).

The evidence that item 4 ("How often during the LAST YEAR have you dealt well with life hassles?") was affected by multidimensionality, and the misfit could not be explained by response dependence or DIF, indicated that this item was not a good measure of perceived coping and lead to its deletion.

Analysis of the revised positively worded items ("Perceived Coping")

Model fit and dimensionality: After the exclusion of item 4, items 5 and 6, and item 7 and 10 were combined into two composite items. The unrestricted parametrization was applied and the new model achieved fit to the Rasch model (χ^2 (28) = 17.63, p = 0.935). The PCA of the residuals showed the factor loadings of Composite Item 3, composed of items 5 and 6 (Λ =

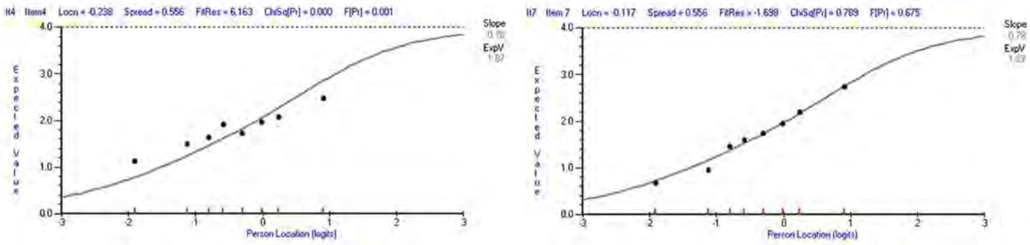


Fig 4. Item characteristic curve for Item 4 (left) and Item 7 (right). Note. The x-axis indicates the latent trait and the y-axis indicates the item score. The black points represent the *average* observed item responses in each class interval. The grey logistic curve indicates the expected item responses. The slope of the Item Characteristic Curve indicates the rate of change of the expected value with respect to the latent trait at the mid-point between the minimum and maximum scores.

https://doi.org/10.1371/journal.pone.0216333.g004

Table 4. Threshold ordering of the positively worded items.

	Composite Item 3	SE	Composite Item 4	SE	Item 9	SE	Item 13	SE
Threshold 1	-1.911	0.164	-2.194	0.201	-2.350	0.124	-1.579	0.124
Threshold 2	-1.525	0.140	-1.580	0.154	-0.315	0.113	-0.635	0.114
Threshold 3	-0.939	0.133	-0.905	0.134	1.076	0.146	**1.234**	0.167
Threshold 4	-0.249	0.146	-0.213	0.136	1.590	0.209	**0.981**	0.212
Threshold 5	0.449	0.187	0.453	0.157				
Threshold 6	1.059	0.272	1.049	0.205				
Threshold 7	1.485	0.414	1.532	0.288				
Threshold 8	1.630	0.581	1.858	0.407				

Note. The item threshold parameter indicates on the latent trait scale the point in which there is an equal probability of response to adjacent categories. Items disordered thresholds are highlighted in bold. The last two thresholds of Item 13 are disordered since the equal probability of selecting the categories "Very often" or "Fairly often" occur for participants with lower Perceived Coping (0.981 logits) than the equal probability of selecting the categories "Sometimes" or "Fairly often" (1.234 logits).

https://doi.org/10.1371/journal.pone.0216333.t004

0.894), Composite Item 4, composed of items 7 and 10 (Λ = -0.827), item 9 (Λ = 0.015) and item 13 (Λ = -0.112) to the first residual factor. Composite Item 3 and item 9 were combined into a subset, while Composite Item 4 and item 13 were combined into another subset. It was found 4.13% statistically significant t-tests between the person estimates (95% Agresti-Coull C. I. [2.47%, 6.76%]) and a disattenuated correlation coefficient of 1.00. In summary, after the deletion of Item 4, there was no more evidence of multidimensionality.

Threshold ordering: The new analysis indicated also that the upper thresholds for Item 13 ("...felt able to control how you spend your time?) were disordered (Table 4) (S2 Fig).

Collapsing the categories of "Fairly often" with "Sometimes" (χ^2 (49) = 540.46, p = < .001), or "Fairly often" with "Very often" (χ^2 (49) = 282.99, p = < .001) disturbed the measurement model. Since collapsing was not possible, the deletion of item 13 was considered. However, the benefits of retaining item 13 in terms of content validity were considered greater than the benefits of excluding it. Notwithstanding, the ordering of item 13 thresholds should be further investigated in future studies with an Aboriginal population.

Item fit: For the analysis of item fit, a Bonferroni adjustment of 0.0025 for the 4 items was applied and no misfit was found (S9 Table). The overall fit to the Rasch model, combined with no further evidence of DIF, local dependence, and item misfit, lead to the establishment of the measurement model for the positively worded items.

Targeting and reliability: The Perceived Coping subscale is well targeted for the Aboriginal pregnant women, with the mean of perceived coping of -0.363 logits (S3 Fig). The PSI was 0.76, a smaller value when compared to 0.80 before the creation of the composite items.

Dimensionality of the revised scales

After the establishment of the two measurement models, a final dimensionality analysis was conducted to test if, after response dependence was accounted with composite items, the revised PSS-14 would exhibit a unidimensional structure. The items 8, 9 and item 13, and composite items 1, 2, 3 and 4 were modeled together and the PCA/t-test procedure applied. The negatively worded (Composite Item 1, Composite Item 2 and Item 8) had positive loadings, while the positively worded (Composite Item 3, Composite Item 4, Item 9 and Item 13) had negative loadings on the first residual factor (S10 Table). It was found 15.4% statistically significant t-tests between the person estimates (95% Agresti-Coull C.I. [12.0%, 19.5%]) and a disattenuated correlation coefficient of 0.25. After the creation of composite items, the revised PSS-14 still exhibits a two-factor structure.

Table 5. Correlations between Perceived Distress and Perceived Coping subscales with Mastery, Perceived Constraints and Social Support.

	Mastery	95% C.I	Perceived Constraints	95% C.I	Social Support Scale	95% C.I
PSS subscales						
Perceived Distress	**-0.13**	**[-0.28, 0.02]**	0.29	[0.14, 0.42]	-0.21	[-0.35, -0.06]
Perceived Coping	0.33	[0.19, 0.45]	-0.32	[-0.45, -0.18]	0.28	[0.13, 0.41]

Note. The table displays the Kendall's τ rank correlations and 95% CI between total (i.e. summated) scores of the Perceived Distress and Perceived Coping subscales and other psychological instruments that evaluate constructs pertaining to the Perceived Stress' nomological network.

https://doi.org/10.1371/journal.pone.0216333.t005

Criterion validity

The subscales of Perceived Distress and Perceived Coping displayed the expected patterns of convergence and divergence regarding Mastery, Perceived Constraints and Social Support (Table 5). The only exception was the relationship between Perceived Distress and Mastery, which exhibited a weak non-significant correlation (r = -0.13 95% C.I. [-0.29, 0.05]).

The analysis of concurrent validity indicated that participants with higher perceived coping had a 10% lower risk of smoking (RR = 0.90–95% C.I. [0.73, 1.10]) and a 31% lower risk of drinking alcohol (RR = 0.69–95% C.I. [0.35, 1.36]) compared to participants with lower perceived coping. In addition, participants with higher perceived stress had a 24% higher risk of smoking (RR = 1.24–95% C.I. [1.01, 1.52]) but no effect was found on alcohol consumption (RR = 0.99–95% C.I. [0.52, 1.90]).

The score distribution of the revised PSS-14 is displayed in S11 Table.

Discussion

The aim of this study was to evaluate if the PSS-14 constitutes a valid and reliable instrument to measure perceived stress in an Aboriginal population. The results indicated that revised versions of the negatively worded ("Perceived Distress") and positively worded ("Perceived Coping") subscales are valid and modestly reliable for an Aboriginal population. However, the relationship between these two psychological traits is not meaningful and a total score of all items should not be created.

The present study started by investigating the main source of debate regarding the PSS, the dimensionality of the scale. The results are consistent with the literature about the PSS-14 indicating a two-factor structure [31]. When the two subscales were analysed separately, there was evidence of local dependence among items. For example, the response dependence between items 7 ("...felt things were going your way?) and 10 ("...felt you were on top of things?) is a form of redundancy-dependency. The overlap of item content works as if the same question was asked twice, but with a slightly different wording [80]. In this case, although these two items are similar enough to measure the latent trait of perceived coping, they are not different enough to produce one item worth of new information [98]. Since reliability indices such as Cronbach's α and PSI are computed under the assumption that items are conditionally independent, these indices can be inflated in the presence of locally dependent items [99]. This phenomenon was observed in the current study since after the local dependence was accounted through the creation of composite items, there was a reduction of the PSI for both subscales.

The internal consistency reliability of the revised PSS-14 subscales ($PSI_{distress}$ = 0.72; PSI_{coping} = 0.76) for a population of pregnant Aboriginal women was only modest [100]. Future applications of the revised PSS-14 in an Aboriginal population should consider that, although reliability between .70 and .80 is usually deemed adequate for research purposes

[101], it does not provide enough measurement precision to be applied in high-stake scenarios [102], such as the use of test results to inform interventions (e.g. which participants should receive an intervention to reduce stress or stop receiving an intervention due to stress). Therefore, the inclusion of new locally independent to increase reliability is recommended. For example, Larson et al. [103] discuss the impact of racism in Aboriginal Australians and suggests the item "Within the past for weeks, have you experienced any physical stress or symptoms as a result of how you were treated because of your race?".

A possible explanation for the modest reliability is that, despite both subscales being well-targeted, the levels of perceived stress (M = -0.18, SD = 0.75) and perceived coping (M = -0.36, SD = 1.10) (S1 Fig) (S3 Fig) were homogenous across the respondents. Aboriginal and Torres Strait Islander comprise a heterogeneous and culturally diverse population [104] but the social inequalities resulting from the process of colonization and subsequent marginalization were experienced *as a group* [105]. Therefore, it is unsurprising that levels of perceived stress (and perceived coping to a lesser extent) did not have a high variation (i.e. high standard deviation) in this group. Future studies should further investigate the reliability of the revised PSS-14 in other Aboriginal samples (i.e. more heterogeneous, containing men, wider age range, etc) and the previous recommendations based on reliability should be taken only as guidelines.

One finding of this study is that the variance of item responses of the PSS-14 could not be explained by the two psychological traits *or* by response dependence among items alone. The variance of the item responses was explained by both the two psychological traits *and* response dependence. The results contrast with findings from Medvedev, Krägeloh [38] about the PSS-10, in which after composite items were created to account for response dependence, the scale exhibited a unidimensional structure (5.2%–95% C.I. [3.2%, n.a]). In our study, the evidence of multidimensionality was clear (15.4%–95% Agresti-Coull C.I. [12.0%, 19.5%]) even after the inclusion of composite items.

The main finding of this study was the relationship between Perceived Distress and Perceived Coping subscales (r = 0.14). The low latent correlation contrasts with the overall literature about the PSS, in which latent factor correlations usually range from r = 0.50 to r = 0.70 [106], sometimes reaching values (>0.80) that pose a threat to discriminant validity [107, 108]. The plausible explanation for the low latent correlation between subscales (r = 0.14) is that the social inequalities experienced by the Aboriginal population are so pronounced that even Aboriginal pregnant women that perceived themselves as coping well with life challenges ended up endorsing items regarding high levels of stress. This means that, irrespectively of their behavioral coping abilities, the Aboriginal pregnant women could not escape experiencing stressful life events. This plausible explanation was further supported by the weak non-significant relationship between Perceived Distress and Mastery (r = -0.13 95% C.I. [-0.29, 0.05]). In contrast to Brown et al. [1], the perception of stress by the Aboriginal mothers were largely unrelated to Mastery, their beliefs regarding the ability to influence external outcomes.

The weak association between Perceived Distress and Perceived Coping/Mastery in Aboriginal Australians is congruent with Lazarus' [26] theory of stress and coping. The negative association between Perceived Stress and Perceived Coping/Mastery in the general population exists because, even if *an event is appraised as threatening*, in case the individual has *a perception of sufficient coping resources* to *influence the event's outcomes*, the stress reaction will be attenuated. However, the hypothesis is that in the case of Aboriginal Australians the frequency of stressful events is so disproportionately high that individual coping capabilities (perceived coping) and the ability to influence outcomes (mastery) are not enough to deal with the increasing amount of problems. Therefore, the relationship between coping, mastery and stress is weakened. It should be noticed that the relationship between Perceived Coping and Mastery (r = 0.33 95% C.I. [0.17, 0.47]) *was not affected* and worked as expected. The weak association between Perceived Stress and Perceived Coping/Mastery indicates that perceived stress nomological network might have shifted [95]. For this reason, a composite score for Perceived Stress should not be created and both Perceived Distress and Perceived Coping scales must be analysed separately.

When the associations of the Perceived Distress and Perceived Coping subscales with health behaviors were investigated, the expected relationships were also found. Higher perceived coping was found to reduce the risk of smoking and drinking alcohol, while higher perceived distress was found to increase the risk of smoking. Despite the exception that no effect was found of perceived distress on alcohol consumption, these findings provide further support that the revised PSS-14 is construct valid for this population.

Finally, the sample had a mean score of 13.35 (SD = 5.24) on the Perceived Distress subscale and a mean score of 9.82 (SD = 4.40) on the Perceived Coping subscale. Although the score distribution (S11 Table) can be used to distinguish stress levels between individuals, the interpretation of mean scores in terms of how much stress is experienced by this Aboriginal population is not simple. Firstly, there is an almost absence of cut-offs in the literature. A few recent studies have proposed cut-off points for the PSS-14 to distinguish between low-stress and high-stress respondents. For example, a total score equal or higher to 28 has been used to define "stressed" as opposed to "non-stressed" in India [109, 110] and Pakistan [111]. However, this proposed cut-off is for total scores summed across all items (PSS-14), whether the findings of our study indicated that for Aboriginal Australians the total scores of each subscale should be used independently. In addition, Cohen and the original authors argued that the PSS is not a diagnostic instrument, so there should be no cut-offs for "high", "medium" or "low" stress and comparisons must be conducted within the sample [112], a position which became a general consensus in the PSS literature [113, 114]. Secondly, it is necessary first to ensure measurement invariance across culture, otherwise, the comparison of scores (or the use of cut-off points derived in another culture) is subject to cultural bias. Future studies should evaluate whether the revised PSS-14 is *cross-culturally valid* to inform the amount of perceived stress in an Aboriginal population by comparing it to a distinct cultural group.

There is a clear limitation of the present study. The study sample was composed of Aboriginal pregnant woman and, although no DIF was found regarding age, socioeconomic position, education, and smoking status, it was not possible to evaluated DIF by gender. Future studies should be conducted to investigate if the functioning of the revised PSS-14 is equivalent for men and non-pregnant Aboriginal women. In addition, when evaluating DIF, some subgroups had small sample sizes and it is not clear whether there was sufficient power to detect true differences. It is recommended future studies of the PSS-14 in a larger Aboriginal sample, which will increase the power to detect not only DIF but other deviations from the model. Finally, only concurrent validity was evaluated, so exposure and outcomes were measured simultaneously. Future studies should also examine predictive validity, with health behaviors (such as smoking status and alcohol consumption) measured at a second-time point, to allow inference regarding causality.

Conclusions

One of the groups most at risk of stress in Australia is pregnant Aboriginal women. The present research showed that a revised PSS-14 is a culturally valid and modestly reliable psychological instrument to measure stress in an Aboriginal population. The relationship between the Perceived Distress and Perceived Coping subscales was weak and the two psychological traits need to be evaluated separately (i.e. a total score across all items should not be created). The use of the revised PSS-14 can provide culturally valid measurement of stress and inform future health research to better understand the role of stress in Aboriginal and Torres Strait Islanders well-being.

Supporting information

S1 Table. Principal Component analysis of the residuals. Note. The table displays the principal components of the item responses' residuals, the eigenvalues of each component, the percentage of total explained variance and the standard errors.
(DOCX)

S2 Table. Factor loadings on the first Principal Component. i. The items 4, 5, 6, 7, 9, 10 and 13 constituted the positively worded items. Note. The table displays the factor loadings of the items responses' residuals on the first principal component (i.e. the first residual component).
(DOCX)

S3 Table. Analysis of DIF of the negatively worded items. i. For each of the 9 NSHT, a Bonferroni adjustment of 0.0055 was applied. The table displays the results of the two-way ANOVA of the residuals according to class intervals (i.e. item-trait interaction) on the first four columns; according to subgroups defined by the exogenous variables (i.e. uniform DIF) on the next four columns; and according to the interaction between exogenous variables and class intervals (i.e. non-uniform DIF) on the last four columns. Statistically significant p-values are highlight in bold.
(DOCX)

S4 Table. Threshold ordering of the negatively worded items. Note. The item threshold parameter indicates on the latent trait scale the point in which there is equal probability of response to adjacent categories. Items disordered thresholds are highlighted in bold.
(DOCX)

S5 Table. Factor loadings on the first Principal Component of the positively worded items. Note. The table displays the factor loadings of the items responses' residuals on the first principal component (i.e. the first residual component).
(DOCX)

S6 Table. Residual correlations of the positively worded items. Note. The residual correlations matrix displays the observed correlation between item responses after the influence of the latent trait ("Perceived Coping") was accounted by the model. It is also displayed the adjusted residual correlations, which are the differences between the observed residual correlations and the average residual correlation.
(DOCX)

S7 Table. Analysis of DIF of the positively worded items. i. For each of the 21 NSHT, a Bonferroni adjustment of 0.0024 was applied. Note. The table displays the results of the two-way ANOVA of the residuals according to class intervals (i.e. item-trait interaction) on the first four columns; according to subgroups defined by the exogenous variables (i.e. uniform DIF) on the next four columns; and according to the interaction between exogenous variables and class intervals (i.e. non-uniform DIF) on the last four columns. Statistically significant p-values are highlight in bold.
(DOCX)

S8 Table. The fit of the positively worded items ("Perceived coping") to the Rasch Model. i. Every item started with the sentence "How often during the LAST YEAR have you. . ..". Note. The second column displays the items' location on the latent trait scale (i.e. the item difficulty). Values of the Fit Residual statistic indicating item misfit (i.e. lower than -2.5 or higher than 2.5), as well as statistically significant $\chi 2$ indicating misfit due to item-trait interaction, were highlighted in bold.
(DOCX)

S9 Table. The fit of the revised positively worded items ("Perceived coping") to the Rasch Model. i. Every item started with the sentence "How often during the LAST YEAR have you. . ..". Note. The second column displays the items' location on the latent trait scale (i.e. the item difficulty). Values of the Fit Residual statistic indicating item misfit (i.e. lower than -2.5 or higher than 2.5), as well as statistically significant $\chi2$ indicating misfit due to item-trait interaction, were highlighted in bold.
(DOCX)

S10 Table. Factor loadings on the first Principal Component of the two revised scales. Note. The table displays the factor loadings of the items responses' residuals on the first principal component (i.e. the first residual component).
(DOCX)

S11 Table. Score distribution of the revised PSS-14. Note. The table displays the score distribution of the revised Perceived Distress and Perceived Coping subscales. The items were responded on a five-point Likert scale (0 = Not at all, 1 = Rarely, 2 = Sometimes, 3 = Fairly often, 4 = Very often).
(DOCX)

S1 Fig. Person-Item Threshold Map of the revised negatively worded items with the information function plotted in the background. Note. The top bars indicate the distribution of the persons parameters (i.e. "Perceived Distress") and the bottom bars indicate the distribution of the item thresholds. The Fisher Information function is plotted on the background (i.e. the green line). It should be noticed that the distribution of item thresholds matches the distribution of person parameters throughout the latent trait indicating good targeting of the revised Perceived Distress subscale for this population.
(TIFF)

S2 Fig. Categories probability curves of Item 9 (left) and Item 13 (right). Note. The graph indicates the probability of endorsing a category according to the latent trait. It can be noticed that for Item 13 category 3 ("Fairly often") never became the most probable for any range of the latent trait scale.
(TIFF)

S3 Fig. Person-Item Threshold Map of the revised positively worded items with the information function plotted in the background. Note. The top bars indicate the distribution of the persons parameters (i.e. "Perceived Coping") and the bottom bars indicate the distribution of the item thresholds. The Fisher Information function is plotted on the background (i.e. the green line). It can be notice that the distribution of person parameters throughout the latent trait is matched with the distribution of item thresholds indicating good targeting for this population. It should be noticed that the distribution of item thresholds matches the distribution of person parameters throughout the latent trait indicating good targeting of the revised Perceived Coping subscale for this population.
(TIFF)

Acknowledgments

The authors gratefully acknowledge the Baby Teeth Talk team including: all members of the Aboriginal Reference Group; intervention implementation staff (Joanne Hedges and Jessica Merrick); data collection staff (Helen Mills, Jacqueline Aldis, Kostas Kapellas, Reshika Chand and Megan Rebuli) and data management support from Kostas Kapellas.

Author Contributions

Conceptualization: Lisa Jamieson.

Formal analysis: Pedro Henrique Ribeiro Santiago.

Funding acquisition: Lisa Jamieson.

Methodology: Pedro Henrique Ribeiro Santiago, Rachel Roberts, Lisa Gaye Smithers, Lisa Jamieson.

Project administration: Lisa Jamieson.

Resources: Rachel Roberts, Lisa Gaye Smithers, Lisa Jamieson.

Software: Pedro Henrique Ribeiro Santiago.

Supervision: Rachel Roberts, Lisa Gaye Smithers, Lisa Jamieson.

Validation: Pedro Henrique Ribeiro Santiago.

Writing – original draft: Pedro Henrique Ribeiro Santiago.

Writing – review & editing: Pedro Henrique Ribeiro Santiago, Rachel Roberts, Lisa Gaye Smithers, Lisa Jamieson.

References

1. Daniel M, Brown A, Dhurrkay J, Cargo MD, O'Dea K. Mastery, perceived stress and health-related behaviour in northeast Arnhem Land: a cross-sectional study. International Journal for Equity in Health. 2006; 5(1):10.
2. Cass A, Brown A, Zhang A, Hughes JT, O'Dea K, Maple-Brown LJ, et al. Resting heart rate, physiological stress and disadvantage in Aboriginal and Torres Strait Islander Australians: analysis from a cross-sectional study. BMC Cardiovasc Disord. 2016; 16(1):36.
3. Kelly K, Dudgeon P, Gee G, Glaskin B. Living on the edge: Social and emotional wellbeing and risk and protective factors for serious psychological distress among Aboriginal and Torres Strait Islander people. Darwin: Cooperative Research Centre for Aboriginal Health. 2009.
4. Blair EM, Zubrick SR, Cox AH. The Western Australian Aboriginal Child Health Survey: findings to date on adolescents. Med J Aust. 2005; 183(8):433–5. PMID: 16225453
5. Amarasena N, Kapellas K, Brown A, Skilton MR, Maple-Brown LJ, Bartold MP, et al. Psychological distress and self-rated oral health among a convenience sample of Indigenous Australians. J Public Health Dent. 2015; 75(2):126–33. https://doi.org/10.1111/jphd.12080 PMID: 25496507
6. Radford K, Delbaere K, Draper B, Mack HA, Daylight G, Cumming R, et al. Childhood stress and adversity is associated with late-life dementia in Aboriginal Australians. The American Journal of Geriatric Psychiatry. 2017; 25(10):1097–106. https://doi.org/10.1016/j.jagp.2017.05.008 PMID: 28689644
7. Prandl KJ. Stress and Pregnancy in Western Australian Aboriginal Women: Curtin University; 2017.
8. Weetra D, Glover K, Buckskin M, Kit JA, Leane C, Mitchell A, et al. Stressful events, social health issues and psychological distress in Aboriginal women having a baby in South Australia: implications for antenatal care. BMC Pregnancy Childbirth. 2016; 16(1):88.
9. Mah B, Weatherall L, Burrows J, Blackwell CC, Gwynn J, Wadhwa P, et al. Post-traumatic stress disorder symptoms in pregnant Australian Indigenous women residing in rural and remote New South Wales: A cross-sectional descriptive study. Aust N Z J Obstet Gynaecol. 2017; 57(5):520–5. https://doi.org/10.1111/ajo.12618 PMID: 28386930
10. Hamdullahpur K, Jacobs KwJ, Gill KJ. A comparison of socioeconomic status and mental health among inner-city Aboriginal and non-Aboriginal women. Int J Circumpolar Health. 2017; 76 (1):1340693. https://doi.org/10.1080/22423982.2017.1340693 PMID: 28747094
11. Wadhwa PD, Sandman CA, Porto M, Dunkel-Schetter C, Garite TJ. The association between prenatal stress and infant birth weight and gestational age at birth: a prospective investigation. Am J Obstet Gynecol. 1993; 169(4):858–65. PMID: 8238139
12. Glover V. Maternal depression, anxiety and stress during pregnancy and child outcome; what needs to be done. Best practice & research Clinical obstetrics & gynaecology. 2014; 28(1):25–35.

13. Aizer A, Stroud L, Buka S. Maternal stress and child outcomes: Evidence from siblings. J Hum Resour. 2015.
14. Thoits PA. Stress, coping, and social support processes: Where are we? What next? J Health Soc Behav. 1995:53–79. PMID: 7560850
15. Cohen Wills TA. Stress, social support, and the buffering hypothesis. Psychol Bull. 1985; 98(2):310. PMID: 3901065
16. Garvey D. Review of the social and emotional wellbeing of Indigenous Australian peoples.; 2008.
17. Waterworth P, Rosenberg M, Braham R, Pescud M, Dimmock J. The effect of social support on the health of Indigenous Australians in a metropolitan community. Soc Sci Med. 2014; 119:139–46. https://doi.org/10.1016/j.socscimed.2014.08.035 PMID: 25181473
18. Brown A, Mentha R, Howard M, Rowley K, Reilly R, Paquet C, et al. Men, hearts and minds: developing and piloting culturally specific psychometric tools assessing psychosocial stress and depression in central Australian Aboriginal men. Soc Psychiatry Psychiatr Epidemiol. 2016; 51(2):211–23. https://doi.org/10.1007/s00127-015-1100-8 PMID: 26233468
19. Kowal E, Gunthorpe W, Bailie RS. Measuring emotional and social wellbeing in Aboriginal and Torres Strait Islander populations: an analysis of a Negative Life Events Scale. International journal for equity in health. 2007; 6(1):18.
20. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine (Phila Pa 1976). 2000; 25(24):3186–91.
21. Byrne BM, Watkins D. The issue of measurement invariance revisited. J Cross Cult Psychol. 2003; 34 (2):155–75.
22. Taylor JM. Psychometric analysis of the Ten-Item Perceived Stress Scale. Psychol Assess. 2015; 27 (1):90. https://doi.org/10.1037/a0038100 PMID: 25346996
23. Cohen S, Kamarck T, Mermelstein R. A Global Measure of Perceived Stress. J Health Soc Behav. 1983; 24(4):385. PMID: 6668417
24. Lazarus RS. Psychological stress and the coping process. New York, NY: McGraw-Hill; 1966.
25. Selye H. Forty years of stress research: principal remaining problems and misconceptions. Can Med Assoc J. 1976; 115(1):53. PMID: 1277062
26. Lazarus RS. Psychological stress and coping in adaptation and illness. The International journal of psychiatry in medicine. 1974; 5(4):321–33. https://doi.org/10.2190/T43T-84P3-QDUR-7RTP PMID: 4618837
27. Heaman MI, Chalmers K. Prevalence and correlates of smoking during pregnancy: A comparison of aboriginal and non-aboriginal women in Manitoba. Birth. 2005; 32(4):299–305. https://doi.org/10.1111/j.0730-7659.2005.00387.x PMID: 16336371
28. Spence ND, Wells S, Graham K, George J. Racial discrimination, cultural resilience, and stress. The Canadian Journal of Psychiatry. 2016; 61(5):298–307. https://doi.org/10.1177/0706743716638653 PMID: 27254805
29. Bersamin A, Wolsko C, Luick BR, Boyer BB, Lardon C, Hopkins SE, et al. Enculturation, perceived stress, and physical activity: implications for metabolic risk among the Yup'ik–The Center for Alaska Native Health Research Study. Ethn Health. 2014; 19(3):255–69. https://doi.org/10.1080/13557858.2012.758691 PMID: 23297688
30. Korn L, Logsdon RG, Polissar NL, Gomez-Beloz A, Waters T, Rÿser R. A randomized trial of a CAM therapy for stress reduction in American Indian and Alaskan Native family caregivers. The Gerontologist. 2009; 49(3):368–77. https://doi.org/10.1093/geront/gnp032 PMID: 19377083
31. Lee E-H. Review of the psychometric evidence of the perceived stress scale. Asian Nurs Res (Korean Soc Nurs Sci). 2012; 6(4):121–7.
32. Michaelides MP, Christodoulou A, Kkeli N, Karekla M, Panayiotou G. Factorial structure of the perceived stress scale and implications for scoring. Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology. 2016; 66(6):309–16.
33. Hewitt PL, Flett GL, Mosher SW. The Perceived Stress Scale: Factor structure and relation to depression symptoms in a psychiatric sample. Journal of Psychopathology and Behavioral Assessment. 1992; 14(3):247–57.
34. Cohen S, Williamson G. Psychological stress in a probability sample of the United States. In: Spacapan S, Oskamp S, editors. The social psychology of health: Claremont symposium on applied social psychology. Newbury Park, CA: Sage Publications; 1988. p. 31–7.
35. Jiang JM, Seng EK, Zimmerman ME, Sliwinski M, Kim M, Lipton RB. Evaluation of the reliability, validity, and predictive validity of the subscales of the perceived stress scale in older adults. J Alzheimers Dis. 2017; 59(3):987–96. https://doi.org/10.3233/JAD-170289 PMID: 28671128

36. Mills S, Azizoddin D, Racaza G, Wallace D, Weisman M, Nicassio P. The psychometric properties of the Perceived Stress Scale-10 among patients with systemic lupus erythematosus. Lupus. 2017; 26 (11):1218–23. https://doi.org/10.1177/0961203317701844 PMID: 28406052

37. Perera MJ, Brintz CE, Birnbaum-Weitzman O, Penedo FJ, Gallo LC, Gonzalez P, et al. Factor structure of the Perceived Stress Scale-10 (PSS) across English and Spanish language responders in the HCHS/SOL Sociocultural Ancillary Study. Psychol Assess. 2017; 29(3):320. https://doi.org/10.1037/pas0000336 PMID: 27280744

38. Medvedev ON, Krägeloh CU, Hill EM, Billington R, Siegert RJ, Webster CS, et al. Rasch analysis of the Perceived Stress Scale: Transformation from an ordinal to a linear measure. J Health Psychol. 2017:1359105316689603. https://doi.org/10.1177/1359105316689603 PMID: 28810395

39. Dougherty BE, Cooley S-SL, Davidorf FH. Measurement of Perceived Stress in Age Related Macular Degeneration. Optometry and vision science: official publication of the American Academy of Optometry. 2017; 94(3):290.

40. Cole SR. Assessment of differential item functioning in the Perceived Stress Scale-10. J Epidemiol Community Health. 1999; 53(5):319. PMID: 10396541

41. Gitchel WD, Roessler RT, Turner RC. Gender effect according to item directionality on the perceived stress scale for adults with multiple sclerosis. Rehabilitation Counseling Bulletin. 2011; 55(1):20–8.

42. Sharp LK, Kimmel LG, Kee R, Saltoun C, Chang C-H. Assessing the Perceived Stress Scale for African American adults with asthma and low literacy. J Asthma. 2007; 44(4):311–6. https://doi.org/10.1080/02770900701344165 PMID: 17530531

43. Jamieson LM, Parker EJ, Roberts-Thomson KF, Lawrence HP, Broughton J. Self-efficacy and self-rated oral health among pregnant aboriginal Australian women. BMC Oral Health. 2014; 14(1):29.

44. Rasch G, editor On general laws and the meaning of measurement in psychology. Proceedings of the fourth Berkeley symposium on mathematical statistics and probability; 1961.

45. Merrick J, Chong A, Parker E, Roberts-Thomson K, Misan G, Spencer J, et al. Reducing disease burden and health inequalities arising from chronic disease among Indigenous children: an early childhood caries intervention. BMC Public Health. 2012; 12(1):323.

46. Goozee G. The development of TAFE in Australia. Adelaide: National Centre for Vocational Education Research; 2001.

47. Pink B. Information paper: an introduction to socio-economic indexes for areas (SEIFA), 2006. Canberra: Australian Bureau of Statistics (ABS). 2008.

48. Geisinger KF. Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychol Assess. 1994; 6(4):304.

49. Lachman ME, Weaver SL. The sense of control as a moderator of social class differences in health and well-being. J Pers Soc Psychol. 1998; 74(3):763. PMID: 9523418

50. Santiago PHR, Roberts R, Smithers LG, Jamieson L. Psychometric properties of the Sense of Personal Control Scale (SPCS), Perceived Stress Scale (PSS) and Social Support Scale (SSS) in an Aboriginal Population. 2017 PHAA SA State Population Health Conference; Adelaide, South Australia2017.

51. House JS. Work stress and social support. Reading, PA: Addison Wesley; 1981.

52. Nielsen T, Dammeyer J, Vang ML, Makransky G. Gender fairness in self-efficacy? A Rasch-based validity study of the General Academic Self-efficacy scale (GASE). Scandinavian Journal of Educational Research. 2017:1–18.

53. Masters GN. A Rasch model for partial credit scoring. Psychometrika. 1982; 47(2):149–74.

54. Mesbah M, Kreiner S. Rasch models for ordered polytomous items. Rasch models in health. 2012:27–42.

55. Olsen LW. Essays on Georg Rasch and his contributions to statistics: Citeseer; 2003.

56. Stenner A. Specific objectivity-local and general. Rasch Measurement Transactions. 1994; 8(3):374.

57. Rasch G. On specific objectivity: An attempt of formalizing the generality and validity of scientific statements. 1977.

58. Poulsen I, Kreiner S, Stat C, Engberg AW. Validation of the Early Functional Abilities scale: an assessment of four dimensions in early recovery after traumatic brain injury. J Rehabil Med. 2018; 50(2):165–72. https://doi.org/10.2340/16501977-2300 PMID: 29313872

59. Arnold SF. Sufficient statistics. Encyclopedia of statistical sciences. 2004;13.

60. Nielsen T, Makransky G, Vang ML, Dammeyer J. How specific is specific self-efficacy? A construct validity study using Rasch measurement models. Studies in Educational Evaluation. 2017; 53:87–97.

61. Christensen KB, Kreiner S, Mesbah M. Rasch models in health. London: John Wiley & Sons; 2013.

62. Hagquist C, Andrich D. Is the sense of coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. Pers Individ Dif. 2004; 36(4):955–68.

63. Andrich D, Sheridan B, Luo G. Manual for the Rasch Unidimensional Measurement Model (RUMM2030). Perth, Western Australia: RUMM Laboratory. 2010.

64. Linacre J. WINSTEPS Rasch measurement computer program. Chicago: WINSTEPS com. 2006.

65. R Core Team. R: A language and environment for statistical computing. 2013.

66. Andrich D, Luo G. Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. Journal of applied measurement. 2003; 4(3):205–21. PMID: 12904672

67. Shea TL, Tennant A, Pallant JF. Rasch model analysis of the Depression, Anxiety and Stress Scales (DASS). BMC Psychiatry. 2009; 9(1):21.

68. Andrich D, De Jong J, Sheridan BE. Diagnostic opportunities with the Rasch model for ordered response categories. Applications of latent trait and latent class models in the social sciences. 1997:59–70.

69. Graham JW. Missing data analysis: Making it work in the real world. Annu Rev Psychol. 2009; 60:549–76. https://doi.org/10.1146/annurev.psych.58.110405.085530 PMID: 18652544

70. Stafford RE, Runyon CR, Casabianca JM, Dodd BG. Comparing Imputation Methods for Trait Estimation Using the Rating Scale Model. J Appl Meas. 2017; 18(1):12–27. PMID: 28453496

71. Fellinghauer CS, Prodinger B, Tennant A. The Impact of Missing Values and Single Imputation upon Rasch Analysis Outcomes: A Simulation Study. J Appl Meas. 2018; 19(1):1–25. PMID: 29561739

72. Draxler C, Alexandrowicz RW. Sample size determination within the scope of conditional maximum likelihood estimation with special focus on testing the Rasch model. Psychometrika. 2015; 80(4):897–919. https://doi.org/10.1007/s11336-015-9472-y PMID: 26155756

73. Linacre J. Sample size and item calibration stability. Rasch Mes Trans. 1994; 7:328.

74. Smith EV Jr. Understanding Rasch measurement: Detecting and evaluating the impact of multidimenstionality using item fit statistics and principal component analysis of residuals. Journal of applied measurement. 2002.

75. Hagell P. Testing rating scale unidimensionality using the principal component analysis (PCA)/t-test protocol with the Rasch model: the primacy of theory over statistics. Open Journal of Statistics. 2014; 4(6):456–65.

76. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Care Res (Hoboken). 2007; 57(8):1358–62.

77. Schumacker R, Muchinsky P. Disattenuating correlation coefficients. Rasch Measurement Transactions. 1996; 10(1):479.

78. Marais I, Andrich D. Formalizing dimension and response violations of local independence in the unidimensional Rasch model. J Appl Meas. 2008; 9(3):200–15. PMID: 18753691

79. Yen WM. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Appl Psychol Meas. 1984; 8(2):125–45.

80. Christensen KB, Makransky G, Horton M. Critical values for Yen's Q 3: Identification of local dependence in the Rasch model using residual correlations. Appl Psychol Meas. 2017; 41(3):178–94. https://doi.org/10.1177/0146621616677520 PMID: 29881087

81. Marais I. Local Dependence. In: Christensen SK K. B., & Mesbah M., editor. Rasch models in health. London, England: Wiley; 2013. p. 111–30.

82. Wainer H, Kiely GL. Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational measurement. 1987; 24(3):185–201.

83. Yen WM. Scaling performance assessments: Strategies for managing local item dependence. Journal of educational measurement. 1993; 30(3):187–213.

84. Hambleton RK, Rogers HJ. Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education. 1989; 2(4):313–34.

85. Hagquist C, Andrich D. Recent advances in analysis of differential item functioning in health research using the Rasch model. Health and quality of life outcomes. 2017; 15(1):181. https://doi.org/10.1186/s12955-017-0755-0 PMID: 28927468

86. Richardson JT. Eta squared and partial eta squared as measures of effect size in educational research. Educational Research Review. 2011; 6(2):135–47.

87. Mellenbergh GJ. Contingency table models for assessing item bias. Journal of educational statistics. 1982; 7(2):105–18.

88. Eklund M, Erlandsson L-K, Persson D, Hagell P. Rasch analysis of an instrument for measuring occupational value: Implications for theory and practice. Scand J Occup Ther. 2009; 16(2):118–28. https://doi.org/10.1080/11038120802596253 PMID: 19085211

89. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ. 1995; 310(6973):170. PMID: 7833759

90. Linacre J. Investigating rating scale category utility. J Outcome Meas. 1999; 3:103–22. PMID: 10204322

91. Smith RM, Plackner C. The family approach to assessing fit in Rasch measurement. Journal of Applied Measurement. 2009; 10(4):424–37. PMID: 19934529

92. Goh HE, Marais I, Ireland MJ. A Rasch model analysis of the mindful attention awareness scale. Assessment. 2017; 24(3):387–98. https://doi.org/10.1177/1073191115607043 PMID: 26467086

93. Gothwal VK, Bagga DK, Sumalini R. Rasch validation of the PHQ-9 in people with visual impairment in South India. J Affect Disord. 2014; 167:171–7. https://doi.org/10.1016/j.jad.2014.06.019 PMID: 24973769

94. Andrich D. Rasch models for measurement. Beverly Hills: Sage; 1988.

95. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955; 52(4):281. PMID: 13245896

96. Kendall SM. Rank correlation. Van Nostrand's Scientific Encyclopedia. 1948.

97. Cohen J. Statistical power analysis for the behavioral sciences. 2nd. Hillsdale, NJ: erlbaum; 1988.

98. Linacre J. Redundant items, overfit and measure bias. Rasch Measurement Transactions. 2000; 14 (3):755.

99. Kreiner S, Christensen KB. Person parameter estimation and measurement in Rasch models. Rasch models in health. 2012:63–78.

100. Nunnally JC, Bernstein IH. Psychometric theory: McGraw-Hill New York; 1967.

101. Furr RM, Bacharach VR. Psychometrics: an introduction. Thousand Oaks, CA: Sage; 2013.

102. Wells CS, Wollack JA. An instructor's guide to understanding test reliability. Testing & Evaluation Services University of Wisconsin. 2003.

103. Larson A, Gillies M, Howard PJ, Coffin J. It's enough to make you sick: the impact of racism on the health of Aboriginal Australians. Aust N Z J Public Health. 2007; 31(4):322–9. PMID: 17725009

104. Neumayer H. Changing the Conversation: Strengthening a Rights-based Holistic Approach to Aboriginal and Torres Strait Islander Health and Wellbeing. Indigenous Allied Health Australia, Canberra; 2013.

105. Carson B, Dunbar T, Chenhall RD, Bailie R. Social determinants of Indigenous health. Crows Nest, NSW: Allen & Unwin; 2007.

106. Andreou E, Alexopoulos EC, Lionis C, Varvogli L, Gnardellis C, Chrousos GP, et al. Perceived Stress Scale: reliability and validity study in Greece. Int J Environ Res Public Health. 2011; 8(8):3287–98. https://doi.org/10.3390/ijerph8083287 PMID: 21909307

107. Reis RS, Hino AA, Anez CR. Perceived stress scale: reliability and validity study in Brazil. J Health Psychol. 2010; 15(1):107–14. https://doi.org/10.1177/1359105309346343 PMID: 20064889

108. Maroufizadeh S, Zareiyan A, Sigari N. Reliability and validity of Persian version of perceived stress scale (PSS-10) in adults with asthma. Arch Iran Med. 2014; 17(5):361–5. PMID: 24784866

109. Walvekar SS, Ambekar JG, Devaranavadagi BB. Study on serum cortisol and perceived stress scale in the police constables. Journal of clinical and diagnostic research: JCDR. 2015; 9(2):BC10. https://doi.org/10.7860/JCDR/2015/12015.5576 PMID: 25859444

110. Brahmbhatt KR, Nadeera V, Prasanna K, Jayram S. Perceived stress and sources of stress among medical undergraduates in a private medical college in Mangalore, India. Int J Biomed Adv Res. 2013; 4(2):133–5.

111. Shah M, Hasan S, Malik S, Sreeramareddy CT. Perceived stress, sources and severity of stress among medical undergraduates in a Pakistani medical school. BMC Med Educ. 2010; 10(1):2.

112. Remor E. Psychometric properties of a European Spanish version of the Perceived Stress Scale (PSS). The Spanish journal of psychology. 2006; 9(1):86–93. PMID: 16673626

113. Morais A, Maia P, Azevedo A, Amaral C, Tavares J. Stress and burnout among Portuguese anaesthesiologists. Eur J Anaesthesiol. 2006; 23(5):433–9. https://doi.org/10.1017/S0265021505001882 PMID: 16469205

114. Cuadros JL, Fernández-Alonso AM, Cuadros-Celorrio ÁM, Fernández-Luzón N, Guadix-Peinado MJ, del Cid-Martín N, et al. Perceived stress, insomnia and related factors in women around the menopause. Maturitas. 2012; 72(4):367–72. https://doi.org/10.1016/j.maturitas.2012.05.012 PMID: 22721806

Supplementary Materials

S1 Table. Principal Component Analysis of the residuals.

Principal Component	Eigenvalue	Percentage of explained variance	SE
PC1	4.018	28.70%	0.563
PC2	1.522	10.87%	0.204
PC3	1.290	9.21%	0.171
PC4	1.086	7.76%	0.139
PC5	1.001	7.15%	0.130
PC6	0.783	5.59%	0.098
PC7	0.759	5.42%	0.105
PC8	0.693	4.95%	0.097
PC9	0.688	4.91%	0.094
PC10	0.594	4.24%	0.088
PC11	0.574	4.10%	0.087
PC12	0.511	3.65%	0.080
PC13	0.479	3.42%	0.075
PC14	0.002	0.02%	0.053

Note. The table displays the principal components of the item responses' residuals, the eigenvalues of each component, the percentage of total explained variance and the standard errors.

S2 Table. Factor loadings on the first Principal Component.

Item	Item4 [i]	Item5	Item6	Item7	Item9	Item10	Item13	Item1	Item2	Item3	Item8	Item11	Item12	Item14
Factor Loading	-0.481	-0.539	-0.593	-0.437	-0.557	-0.488	-0.590	0.565	0.476	0.573	0.361	0.666	0.602	0.508

i. The items 4, 5, 6, 7, 9, 10 and 13 constituted the positively worded items. Note. The table displays the factor loadings of the items responses' residuals on the first principal component (i.e. the first residual component).

S3 Table. Analysis of DIF of the negatively worded items.

	Perceived Stress (Class Interval)				Age[i]				Perceived Stress-by-Age			
	η2	η_p2	*df*	Prob	η2	η_p2	*df*	Prob	η2	η_p2	*df*	Prob
Composite Item1	0.007	0.007	3	0.519	0.005	0.005	2	0.460	0.014	0.014	6	0.595
Composite Item2	0.013	0.014	3	0.209	0.007	0.007	2	0.309	0.014	0.014	6	0.571
Item 8	0.008	0.009	3	0.420	0.001	0.001	2	0.845	0.022	0.022	6	0.292
	Perceived Stress (Class Interval)				Socioeconomic position				Perceived Stress-by-Socioeconomic position			
	η2	η_p2	*df*	Prob	η2	η_p2	*df*	Prob	η2	η_p2	*df*	Prob
Composite Item1	0.005	0.005	3	0.611	0.021	0.021	4	0.130	0.013	0.014	11	0.946
Composite Item2	0.015	0.015	3	0.152	0.013	0.014	4	0.317	0.025	0.026	11	0.624
Item 8	0.005	0.005	3	0.623	0.006	0.006	4	0.703	0.050	0.051	11	0.091
	Perceived Stress (Class Interval)				Education				Perceived Stress-by-Education			
	η2	η_p2	*df*	Prob	η2	η_p2	*df*	Prob	η2	η_p2	*df*	Prob
Composite Item1	0.005	0.005	3	0.626	0.004	0.004	1	0.229	0.011	0.011	3	0.264
Composite Item2	0.014	0.015	3	0.154	0.002	0.002	1	0.447	0.030	0.030	3	0.013
Item 8	0.005	0.005	3	0.656	0.001	0.001	1	0.511	0.006	0.006	3	0.524
	Perceived Stress (Class Interval)				Smoking Status				Perceived Stress-by-Smoking Status			
	η2	η_p2	*df*	Prob	η2	η_p2	*df*	Prob	η2	η_p2	*df*	Prob
Composite Item1	0.005	0.006	3	0.594	0.004	0.004	2	0.539	0.012	0.012	6	0.645
Composite Item2	0.014	0.014	3	0.170	0.006	0.007	2	0.319	0.010	0.010	6	0.726
Item 8	0.004	0.005	3	0.672	0.017	0.018	2	0.049	0.008	0.008	6	0.846

i. For each of the 9 NSHT, a Bonferroni adjustment of 0.0055 was applied. The table displays the results of the two-way ANOVA of the residuals according to class intervals (i.e. item-trait interaction) on the first four columns; according to subgroups defined by the exogenous variables (i.e. uniform DIF) on the next four columns; and according to the interaction between exogenous variables and class intervals (i.e. non-uniform DIF) on the last four columns. Statistically significant p-values are highlighted in bold.

S4 Table. Threshold ordering of the negatively worded items.

	Composite Item 1	SE	Composite Item 2	SE	Item 8	SE
Threshold 1	-0.703	0.281	-1.196	0.323	-1.581	0.124
Threshold 2	-1.050	0.254	-1.223	0.267	-0.959	0.108
Threshold 3	-1.114	0.211	-1.116	0.218	1.076	0.158
Threshold 4	-0.958	0.175	-0.902	0.182	1.465	0.268
Threshold 5	-0.644	0.154	-0.609	0.161		
Threshold 6	-0.234	0.150	-0.262	0.152		
Threshold 7	0.209	0.161	0.111	0.155		
Threshold 8	0.623	0.190	0.484	0.171		
Threshold 9	0.946	0.239	0.829	0.200		
Threshold 10	1.115	0.301	1.120	0.245		
Threshold 11	1.068	0.345	1.330	0.304		
Threshold 12	0.742	0.332	1.432	0.368		

Note. The item threshold parameter indicates on the latent trait scale the point in which there is an equal probability of response to adjacent categories. Items disordered thresholds are highlighted in bold.

S5 Table. Factor loadings on the first Principal Component of the positively worded items.

Item	Item4	Item5	Item6	Item7	Item9	Item10	Item13
Factor Loading	0.725	0.537	-0.027	-0.515	-0.251	-0.634	-0.135

Note. The table displays the factor loadings of the items responses' residuals on the first principal component (i.e. the first residual component).

S6 Table. Residual correlations of the positively worded items.

		Item 4	Item 5	Item 6	Item 7	Item 9	Item 10	Item 13
Item 4		1						
Item 5	Obs	-0.057	1					
	Adj	0.103						
Item 6	Obs	-0.324	**0.058**	1				
	Adj	-0.164	**0.218**					
Item 7	Obs	-0.277	-0.194	-0.162	1			
	Adj	-0.117	-0.034	-0.002				
Item 9	Obs	-0.262	-0.199	-0.132	-0.188	1		
	Adj	-0.100	-0.039	0.028	-0.028			
Item 10	Obs	-0.308	-0.245	-0.093	**0.086**	-0.063	1	
	Adj	-0.148	-0.085	0.067	**0.246**	0.097		
Item 13	Obs	-0.190	-0.294	-0.100	-0.138	-0.116	-0.184	1
	Adj	-0.030	-0.134	0.060	0.022	0.044	-0.024	

Note. The residual correlations matrix displays the observed correlation between item responses after the influence of the latent trait ("Perceived Coping") was accounted by the model. It is also displayed the adjusted residual correlations, which are the differences between the observed residual correlations and the average residual correlation.

S7 Table. Analysis of DIF of the positively worded items.

	Perceived Stress (Class Interval)				Age[i]				Perceived Stress-by-Age			
	η2	$\eta_p 2$	*df*	Prob	η2	$\eta_p 2$	*df*	Prob	η2	$\eta_p 2$	*df*	Prob
Item 4	0.111	0.112	3	**0.000**	0.004	0.005	2	0.495	0.001	0.006	6	0.923
Item 5	0.051	0.053	3	0.008	0.008	0.009	2	0.253	0.006	0.039	6	0.060
Item 6	0.053	0.054	3	0.016	0.002	0.002	2	0.783	0.002	0.015	6	0.595
Item 7	0.034	0.034	3	0.013	0.000	0.000	2	0.937	0.002	0.011	6	0.774
Item 9	0.018	0.019	3	0.593	0.001	0.001	2	0.862	0.004	0.027	6	0.226
Item 10	0.054	0.054	3	0.004	0.003	0.004	2	0.577	0.002	0.012	6	0.705
Item 13	0.031	0.032	3	0.263	0.003	0.003	2	0.658	0.003	0.019	6	0.450
	Perceived Stress (Class Interval)				Socioeconomic position				Perceived Stress-by-Socioeconomic position			
	η2	$\eta_p 2$	*df*	Prob	η2	$\eta_p 2$	*df*	Prob	η2	$\eta_p 2$	*df*	Prob
Item 4	0.106	0.110	3	**0.000**	0.010	0.011	4	0.438	0.026	0.030	11	0.502
Item 5	0.045	0.047	3	0.010	0.002	0.003	4	0.926	0.031	0.032	11	0.423
Item 6	0.047	0.049	3	0.031	0.003	0.003	4	0.902	0.035	0.037	11	0.303
Item 7	0.031	0.032	3	0.018	0.016	0.017	4	0.212	0.014	0.015	11	0.930
Item 9	0.023	0.024	3	0.336	0.012	0.013	4	0.368	0.036	0.037	11	0.292
Item 10	0.041	0.043	3	0.023	0.005	0.005	4	0.790	0.028	0.029	11	0.512
Item 13	0.036	0.038	3	0.318	0.016	0.017	4	0.203	0.028	0.030	11	0.506
	Perceived Stress (Class Interval)				Education				Perceived Stress-by-Education			
	η2	$\eta_p 2$	*df*	Prob	η2	$\eta_p 2$	*df*	Prob	η2	$\eta_p 2$	*df*	Prob
Item 4	0.106	0.107	3	**0.000**	0.000	0.000	1	0.751	0.009	0.010	3	0.328

Item 5	0.045	0.046	3	0.009	0.001	0.001	1	0.522	0.012	0.013	3	0.213
Item 6	0.045	0.046	3	0.026	0.000	0.000	1	0.879	0.001	0.002	3	0.912
Item 7	0.031	0.031	3	0.017	0.009	0.010	1	0.063	0.008	0.008	3	0.396
Item 9	0.022	0.022	3	0.370	0.005	0.005	1	0.170	0.003	0.003	3	0.818
Item 10	0.041	0.041	3	0.024	0.000	0.000	1	0.837	0.014	0.014	3	0.166
Item 13	0.020	0.021	3	0.331	0.002	0.002	1	0.432	0.006	0.006	3	0.532
	Perceived Stress (Class Interval)				Smoking Status				Perceived Stress-by-Smoking Status			
	η2	η_p2	*df*	Prob	η2	η_p2	*df*	Prob	η2	η_p2	*df*	Prob
Item 4	0.099	0.100	3	**0.000**	0.002	0.002	2	0.672	0.005	0.005	6	0.933
Item 5	0.032	0.033	3	0.009	0.007	0.007	2	0.280	0.021	0.022	6	0.258
Item 6	0.026	0.027	3	0.025	0.003	0.003	2	0.554	0.027	0.027	6	0.142
Item 7	0.029	0.029	3	0.017	0.000	0.000	2	0.953	0.017	0.018	6	0.401
Item 9	0.009	0.009	3	0.389	0.000	0.000	2	0.986	0.022	0.022	6	0.253
Item 10	0.026	0.027	3	0.025	0.008	0.008	2	0.234	0.024	0.024	6	0.198
Item 13	0.009	0.010	3	0.343	0.005	0.005	2	0.386	0.011	0.011	6	0.679

i. For each of the 21 NSHT, a Bonferroni adjustment of 0.0024 was applied. Note. The table displays the results of the two-way ANOVA of the residuals according to class intervals (i.e. item-trait interaction) on the first four columns; according to subgroups defined by the exogenous variables (i.e. uniform DIF) on the next four columns; and according to the interaction between exogenous variables and class intervals (i.e. non-uniform DIF) on the last four columns. Statistically significant p-values are highlighted in bold.

ii.

S8 Table. The fit of the positively worded items ("Perceived coping") to the Rasch Model.

Item[i]	Location	SE	Fit Residual	*df*	$\chi 2$	*df*	Prob
4. dealt well with life hassles?	-0.192	0.059	**6.083**	302.81	52.751	7	**0.000**
5. coped well with important changes in your life?	0.236	0.065	-0.258	301.97	13.67	7	0.057
6. felt able to handle your personal problems?	0.47	0.066	-1.732	302.81	11.657	7	0.112
7. felt things were going your way?	-0.018	0.066	-0.363	301.97	9.438	7	0.223
9. felt able to control irritations in your life?	-0.346	0.063	0.669	302.81	7.169	7	0.412
10. felt you were on top of things?	-0.084	0.064	-1.318	302.81	11.197	7	0.130
13. felt able to control how you spend your time?	-0.065	0.060	-0.014	302.81	6.518	7	0.481

i. Every item started with the sentence "How often during the LAST YEAR have you….". Note. The second column displays the items' location on the latent trait scale (i.e. the item difficulty). Values of the Fit Residual statistic indicating item misfit (i.e. lower than -2.5 or higher than 2.5), as well as statistically significant $\chi 2$ indicating misfit due to item-trait interaction, were highlighted in bold.

S9 Table. The fit of the revised positively worded items ("Perceived coping") to the Rasch Model.

Item[i]	Location	SE	Fit Residual	*df*	$\chi 2$	*df*	Prob
Composite Item 3	0.427	0.044	-0.399	260.700	3.160	7.000	0.870
Composite Item 4	-0.044	0.044	-1.161	261.430	2.670	7.000	0.914
9. felt able to control irritations in your life?	-0.338	0.064	1.482	261.430	9.332	7.000	0.230
13. felt able to control how you spend your time?	-0.046	0.061	1.501	261.430	2.472	7.000	0.929

i. Every item started with the sentence "How often during the LAST YEAR have you….". Note. The second column displays the items' location on the latent trait scale (i.e. the item difficulty). Values of the Fit Residual statistic indicating item misfit (i.e. lower than -2.5 or higher than 2.5), as well as statistically significant $\chi 2$ indicating misfit due to item-trait interaction, were highlighted in bold.

S10 Table. Factor loadings on the first Principal Component of the two revised scales.

Item	Composite Item1	Composite Item2	Item8	Composite Item3	Composite Item4	Item9	Item13
Factor Loading	0.654	0.770	0.399	-0.689	-0.679	-0.665	-0.647

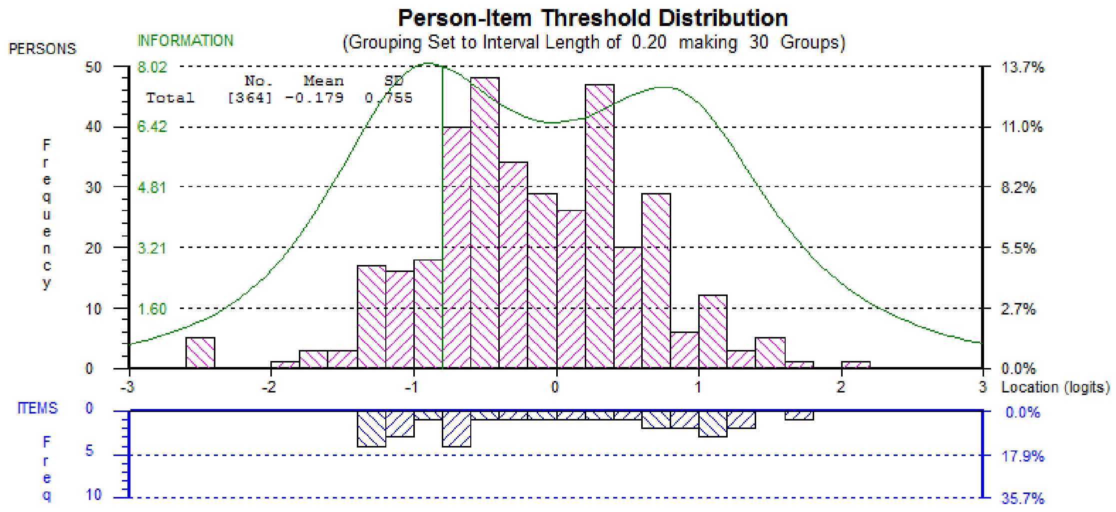
Note. The table displays the factor loadings of the items responses' residuals on the first principal component (i.e. the first residual component).

S11 Table. Score distribution of the revised PSS-14.

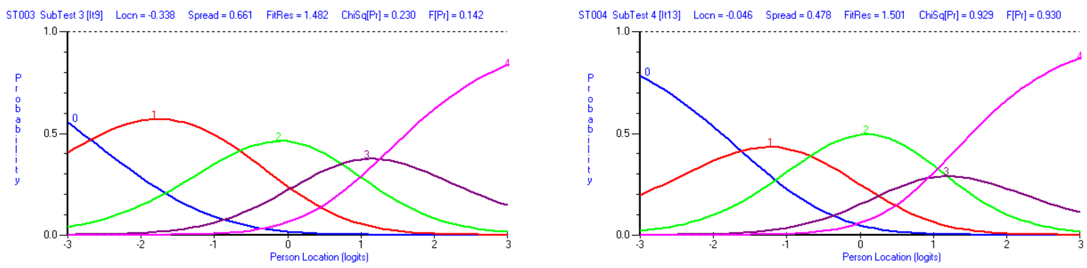
Perceived Distress subscale			Perceived Coping subscale		
Total Score	Frequency	Cumulative %	Total Score	Frequency	Cumulative %
0	5	1.4	0	7	1.9
2	3	2.2	1	6	3.5
3	3	3.0	2	3	4.4
4	11	6.0	3	11	7.4
5	6	7.6	4	10	10.1
6	10	10.4	5	16	14.4
7	6	12.0	6	28	22.1
8	18	16.9	7	28	29.7
9	18	21.8	8	35	39.2
10	21	27.5	9	30	47.4
11	20	33.0	10	27	54.8
12	26	40.1	11	34	64.0
13	33	49.0	12	33	73.0
14	27	56.4	13	26	80.1
15	26	63.5	14	18	85.0
16	25	70.3	15	14	88.8
17	22	76.3	16	8	91.0
18	19	81.5	17	6	92.6
19	18	86.4	18	10	95.4
20	10	89.1	19	3	96.2
21	6	90.7	20	3	97.0
22	6	92.4	22	3	97.8
23	6	94.0	24	1	98.1
24	3	94.8			
25	5	96.2			
26	1	96.5			
27	1	96.7			
Missing	12	100	Missing	7	100
Total	367	100	Total	367	100

Note. The table displays the score distribution of the revised Perceived Distress and Perceived Coping subscales. The items were responded on a five-point Likert scale (0=Not at all, 1=Rarely, 2=Sometimes, 3=Fairly often, 4=Very often).

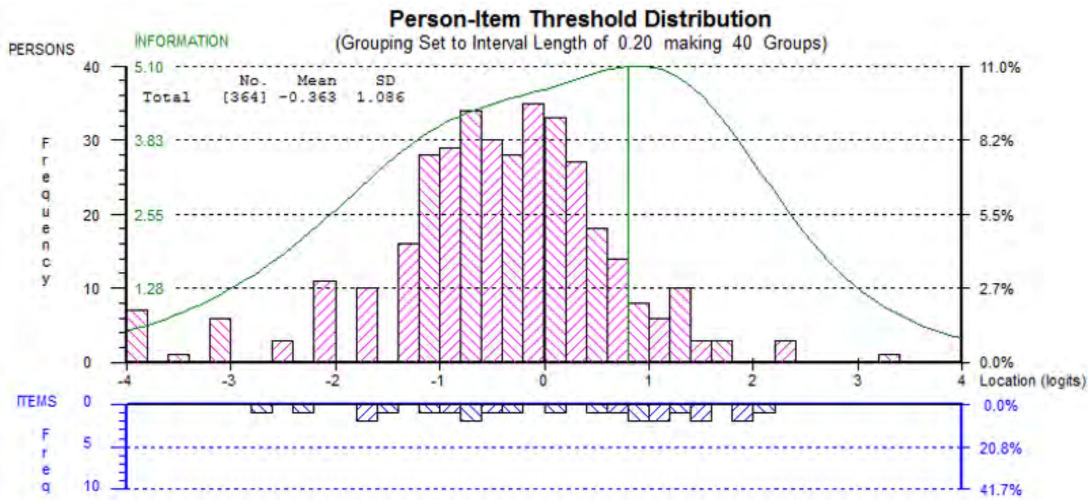
S1 Fig.



S2 Fig.



S3 Fig.



Chapter 4 – Measuring Stress in Australia: Validation of the Perceived Stress Scale (PSS-14) in a Nationally Representative Sample

Statement of Authorship

Title of Paper	Measuring Stress in Australia: Validation of the Perceived Stress Scale (PSS-14) in a Nationally Representative Sample
Publication Status	☐ Published ☐ Accepted for Publication ☒ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Submitted to Health and Quality of Life Outcomes.

Principal Author

Name of Principal Author (Candidate)	Pedro Henrique Ribeiro Santiago		
Contribution to the Paper	Conceptualized the idea, performed the analysis, wrote the primary draft and acted as corresponding author.		
Overall percentage (%)	80%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	21/11/2019

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);
ii. permission is granted for the candidate in include the publication in the thesis; and
iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Tine Nielsen		
Contribution to the Paper	Conceptualized the idea, supervised development of work, provided intellectual contribution, and critically reviewed the draft manuscript. Contributed to teaching the methods applied in this research (Rasch models and Graphical Log-linear Rasch models).		
Signature		Date	20/11/2019

Name of Co-Author	Lisa Smithers		
Contribution to the Paper	Conceptualized the idea, supervised development of work, provided intellectual contribution, and critically reviewed the draft manuscript.		
Signature		Date	21/11/2019

Name of Co-Author	Rachel Roberts		
Contribution to the Paper	Conceptualized the idea, supervised development of work, provided intellectual contribution, and critically reviewed the draft manuscript.		
Signature		Date	21/11/2019

Name of Co-Author	Lisa Jamieson		
Contribution to the Paper	Conceptualized the idea, supervised development of work, provided intellectual contribution, and critically reviewed the draft manuscript.		
Signature		Date	21/11/2019

Linkage to the body of work

After the development of an adapted version for an Aboriginal population (Study 1), one remaining question was the validity of the Perceived Stress Scale for non-Aboriginal Australians. Despite the Perceived Stress Scale being applied in national surveys such as the Stress & Well-Being Survey (SWBS) (Australian Psychological Society, 2015; Casey & Mathews, 2011), to the best of our knowledge validity for non-Aboriginal Australians has not been provided. Hence, in the second study, I examined the PSS-14 psychometric properties for a general non-Aboriginal Australian population. The findings indicated that a revised 8-item Perceived Stress Scale version should be preferred over the original PSS-14. The revised 8-item Perceived Stress Scale is readily available and can be applied to measure stress in Australia in non-Aboriginal populations.

The relevance of examining the Perceived Stress Scale psychometric properties for non-Aboriginal Australians surpassed establishing validity in this group; it also indicated that, in Australia, the psychometric properties of the Perceived Stress Scale were *different* between Aboriginal and non-Aboriginal people. Hence, the findings reinforce that validation of an instrument for Aboriginal Australians is necessary and should be conducted independently from non-Aboriginal Australians, since psychometric properties can differ across these groups.

Highlights

- The study provided initial evidence of validity of a revised 8-item version of the Perceived Stress Scale (PSS-14), composed of the Perceived Stress and Perceived Coping subscales, for the general non-Aboriginal Australian population.

- The four items excluded in the original validation by Cohen (1988), which proposed the PSS-10 instead of the PSS-14, displayed problems and were also excluded in Australia.
- We confirmed the differential item functioning (DIF) by gender of four items and hypothesize that this DIF *pattern* is a consequence of gender roles present in Western societies.

Research and Policy Implications

- A revised 8-item Perceived Stress Scale should be preferred to measure perceived stress in non-Aboriginal Australians instead of other versions (i.e. PSS-14, PSS-10). Total scores should be computed independently for the Perceived Stress and Perceived Coping subscales. The presence of DIF among four items indicates that adjustment of total scores through conversion tables should be conducted to avoid measurement bias.

Measuring Stress in Australia: Validation of the Perceived Stress Scale (PSS-14) in a Nationally Representative Sample

Pedro H Ribeiro Santiago

The University of Adelaide

Tine Nielsen

The University of Copenhagen

Lisa Smithers, Rachel Roberts, and Lisa Jamieson

The University of Adelaide

Author Note

Pedro H Ribeiro Santiago, Adelaide Dental School, The University of Adelaide; Tine Nielsen, Department of Psychology, The University of Copenhagen; Lisa Smithers, School of Public Health, The University of Adelaide; Rachel Roberts, School of Psychology, The University of Adelaide; Lisa Jamieson, Adelaide Dental School, The University of Adelaide.

This research was supported by a grant from Australia's National Health and Medical Research Council.

Correspondence concerning this article should be addressed to Pedro H Ribeiro Santiago, Adelaide Dental School, The University of Adelaide, AHMS Building, North Terrace, South Australia 5000. Email: pedro.ribeirosantiago@adelaide.edu.au

Abstract

In Australia, the stress levels have increased over the years, impacting on the physical and mental health of the general population. The aim of the present study was to evaluate the validity and reliability of the PSS-14 in an Australian population. The PSS-14 was applied to a representative sample comprising 3,857 Australians in the population-based cross-sectional study Australia's National Survey of Adult Oral Health 2004-2006. The psychometric properties analyzed with the Rasch model and Graphical Log-linear Rasch models were: model fit, item fit, local dependence, differential item functioning (DIF), unidimensionality, reliability, targeting and criterion validity. The Perceived Stress subscale displayed adequate psychometric properties after the deletion of two items; however, the majority of problems centered around the Perceived Control subscale. The presence of DIF among four items indicates that adjustment of total scores is required to avoid measurement bias. Recommendations for future applications in Australia are provided.

Keywords: Psychometrics; Perceived Stress Scale; Australia; Differential Item Functioning; Measurement Invariance; Psychological Stress; Rasch Analysis.

Introduction

In Australia, the Australian Psychological Society (APS) conducted a 'State-of-the-Nation' Stress & Well-Being Survey (SWBS) from 2011 to 2015 to investigate stress at a national level (Australian Psychological Society, 2015; Casey & Mathews, 2011). The results showed that almost two in three Australians (64%) reported that stress was impacting their mental health, while approximately one in five (17%) reported that stress was strongly impacting their physical health (Casey & Liang, 2014). The findings from the last survey, which had 1731 respondents, indicated that compared to 2011 the levels of stress increased, and the levels of well-being decreased in the Australian population. One of the concerning findings was that, among those with severe levels of distress, 61% drank alcohol, 41% gambled, 40% smoked and 31% used recreational drugs as a coping mechanism (Australian Psychological Society, 2015). The surveys also revealed gender differences. Women were consistently more affected by stress than men and reported financial and health issues as their main sources of concern (Casey & Liang, 2014).

One of the many psychological instruments used in the SWBS was the Perceived Stress Scale (PSS) (S. Cohen, Kamarck, & Mermelstein, 1983). The PSS is the world's most widely used instrument to measure perceived stress (Taylor, 2015) and since its development has been continuously applied in empirical research (Kechter et al., 2019; Manzar et al., 2019). The PSS was developed based on the theoretical perspective of Lazarus (1966), which rather than focusing on external environmental stressors, postulated that the stress response is determined by the *perception* of these environmental stressors. According to Lazarus (1966), life events, such as divorce or losing a job, only cause stress when they are *appraised as threatening (*e.g. "I don't have another job"*)* and there is a *perception of insufficient coping resources (*e.g. "I don't know anyone who could employ me"). The measurement of stress has then been operationalized in two ways: the *environmental perspective* (e.g. using life-event scales) and the *psychological perspective* (e.g. using perceived stress scales) (Sheldon Cohen, Janicki-Deverts, & Miller, 2007; Harkness & Monroe, 2016). The PSS was developed to measure stress from the *psychological perspective*, diverging from the life-event scales regularly used at that time (Holmes & Rahe, 1967). The initial validations conducted by Cohen (1983; 1988) led to the creation of two shortened scales derived from the original 14 item-version (PSS-14): the PSS-10 and the PSS-4.

The results of the SWBW surveys were nationally reported by the Australian media (see “Australian women feel more stressed than men, mental health survey finds” (Australia Associated Press, 2014)). However, the reports did not specify which PSS version was used and indicated only that the “level of stress was derived by summing the scores of the 11 scale items” (Australian Psychological Society, 2015, p. 43). Additionally, evidence of validity was not provided. Considering the high levels of stress reported in the Australian population, it is necessary to ensure that psychological measures applied to measure stress in Australians are valid and reliable, so it is possible to have confidence in the interpretation of test results. In the present study, we aim to investigate the psychometric properties of the PSS-14 in the general Australian population and examine whether this instrument can provide a valid measure of perceived stress for future research. To evaluate the PSS-14 validity and reliability we used data collected for Australia’s National Survey of Adult Oral Health (NSAOH) 2004-2006, a broad project originally aimed to determine the psychosocial determinants of oral health in Australia. Despite being conducted prior to the SWBW, the NSAOH 2004-2006 has a large nationally representative sample (n=3,857) that can provide evidence of the PSS-14 validity in the Australian general population.

The present research

The psychometric properties of the PSS have been evaluated in multiple countries (Lee, 2012). There are, however, two main limitations regarding the generalizability of its psychometric properties to an Australian population. Firstly, the majority of studies evaluated the PSS-14 in small and/or non-representative samples (Lee, 2012). For example, in China, the PSS-14 was evaluated in a sample of 1860 cardiac patients who smoked (Leung, Lam, & Chan, 2010), while the PSS-10 was evaluated in a sample of policewomen (Wang et al., 2011). Secondly, other studies were conducted in countries culturally and economically diverse from Australia, such as the application of the PSS-10 to 479 adults in Thailand (Wongpakaran & Wongpakaran, 2010), a country known for its "collectivist Eastern culture" (Patterson & Smith, 2001, p. 94); or the application of the PSS-14 to 941 adults in Greece (Andreou et al., 2011), which recently experienced financial crisis (Dudin, Gayduk, Sekerin, Bank, & Gorohova, 2016). Among all countries studied, Canada is the Western developed nation most similar to Australia due to its "large geography, low population density, and similar health care challenges" (Pong, DesMeules, & Lagacé, 2009, p. 58). However, the PSS-14 was initially applied in Canada to 96 psychiatric patients (Hewitt, Flett, & Mosher, 1992) and the PSS-4 was later evaluated in 217 pregnant women (Karam et al., 2012). The peculiarity of the samples from Canada (i.e. psychiatric patients) and most countries in general makes it difficult to generalize the results to typical members of the Australian general population. For the most part, the PSS has been validated in samples experiencing stressful environments (i.e. patients, students, policemen) rather than in general populations (Lee, 2012).

The most relevant study in a population similar to Australia continues to be the validation conducted by S. Cohen and Williamson (1988, p. 45) in a representative sample of 2387 Americans. Both countries, Australia and United States (US), are large high income countries (World Bank, 2017), with a history of English colonization (Archibald, 2006) and populations with similar demographic characteristics (Papanicolas, Woskie, & Jha, 2018) and morbidity patterns (WHO, 2015a, 2015b). Nevertheless, there are important dissimilarities in terms of social-political context between these countries. For example, in the US, the national health system is a private employer-based and individual insurance program that provides coverage to 90% of the population, while Australia has a universal public insurance program covering 100% of the individuals (Papanicolas et al., 2018). Although finances are the main source of stress both in Australia (Australian Psychological Society, 2015) and the US (American Psychological Association, 2015), these are structural differences regarding how these environmental stressors are experienced by each population (i.e. concerns with *health* costs are more prominent in the US).

One important characteristic of the Australian population is the cultural background of its Indigenous groups, namely Aboriginal Australians and Torres Strait Islanders (ABTSI). The Aboriginal Australians experiences of well-being are rather distinct from Western individuals (Garvey, 2008) and "Western psychological concepts are inappropriate and potentially damaging to Indigenous people" (Kowal, Gunthorpe, & Bailie, 2007, p. 2). For this reason, researchers such as Kowal et al. (2007) and Santiago, Roberts, Smithers, and Jamieson (2019) recommend that mainstream instruments should be validated *independently* for Aboriginal Australians due to cultural differences. One example is the PSS-14, which was recently validated for an Aboriginal population and the findings showed a weak latent correlation between the "Perceived Stress" and "Perceived Coping" subscales (r=0.14), a result distinct from the moderate (r=0.50) to strong (0.70) correlations found in Western societies (Santiago et al., 2019, p. 16). For these reasons, we followed the recent recommendations that ABTSI is a culturally distinct group in which psychological instruments should be evaluated separately from the general Australian population.

In summary, the validation conducted by S. Cohen and Williamson (1988) provided evidence of the psychometric properties of the PSS in a representative sample of a high-income country such as the US. Notwithstanding the similarities between the US and Australian population, the differences in terms of socio-political context indicate that validation for an Australian population is required.

The present study aims to (1) investigate the psychometric properties of the PSS-14 in the general Australian population. We hypothesize that the functioning of the PSS-14 in the Australian population is similar but not equal to its functioning in other high-income countries. In addition, we aim to (2) updated the evidence about the PSS-14 functioning in developed countries using a nationally representative sample. Since S. Cohen and Williamson (1988), only two studies had investigated the PSS properties with nationally representative samples, notably in German (Klein et al., 2016) and Denmark (M. G. Nielsen et al., 2016). However, similarly to the majority of PSS studies, these two studies focused on dimensionality and reliability (Lee, 2012) and “did not test for differential item functioning or local dependency” (M. G. Nielsen et al., 2016, p. 29). We intend to (3) further advance the knowledge regarding the PSS psychometric properties using item-response theory to investigate issues of differential item functioning and local dependence. The previous research about stress in Australia showed that “Australian women feel more stressed than men” (Australia Associated Press, 2014). Although this result is common in many Western countries, a long-established questioning is whether those differences are due to measurement bias (Lavoie & Douglas, 2012; Lee, 2012). Therefore, we aim to (4) investigate gender differences in PSS scores, and whether differences were due to measurement bias.

Finally, we aim to evaluate criterion validity by inspecting convergence and divergent validity with two psychological constructs (social support and stress at work) of the perceived stress’ nomological network (Cronbach & Meehl, 1955). Social support has been shown by a large body of research as a protective (or *buffering)* factor against stress (Lakey & Cohen, 2000). Social support refers to the functions performed by family, friends, and significant others when an individual encounters an external environmental stressor (Thoits, 1995). In this case, family, friends or significant others can help to change the situation (e.g. helping with a task at work) or change *the meaning* of the situation (e.g. help to interpret the event from a less distressing or extreme perspective) (Zimet, Dahlem, Zimet, & Farley, 1988). In both cases, the individual has additional resources to deal with the *environmental stressor* and this decreases his *perception* of how stressful the situation is (Cohen & Wills, 1985).

On the other hand, psychological stress can be experienced at work due to a demanding environment. One theoretical model that explains how the work environment generates stressful experiences is the *effort-reward imbalance* (Siegrist, 1996). The model indicates that when the rewards received at work did not correspond to the efforts employed ('high cost/low gain'), the imbalance can lead adverse stress responses (Siegrist, Wege, Pühlhofer, & Wahrendorf, 2009). Therefore, it is expected that participants with high perceived stress will have low social support from friends, family and significant others and experience more efforts with fewer rewards at work.

Methods

Participants and procedures

The sample comprised 3,857 non-Aboriginal Australians in the population-based cross-sectional study Australia's National Survey of Adult Oral Health 2004-2006. The NSAOH 2004-2006 was a broad project aimed to determine the psychosocial determinants of oral health in Australia. The survey had a three-stage (i.e. postcodes, households, people) stratified clustered sampling design to select a representative sample of Australian adults. Participants were contacted by study staff who conducted a computer-assisted telephone interview. Interviewees that agreed to undertake dental examinations were mailed the PSS-14 (Supplementary Table 1), along with the other complementary measures, as part of a larger questionnaire. The NSAOH 2004-2006 was approved by the University of Adelaide's Human Research Ethics Committee. All participants provided signed informed consent (Slade, Spencer, & Roberts-Thomson, 2004). The response rate of the PSS-14 was 77.4 % (Sanders & Slade, 2011). A sample of 42 Aboriginal Australians was removed from the analysis since the PSS-14 has been previously validated for this group (Santiago et al., 2019) and it is recommended that psychometric research with Indigenous groups should be conducted separately due to cultural differences (Kowal et al., 2007).

Psychometric properties of the Perceived Stress Scale

The psychometric properties of the PSS have been evaluated in multiple countries, including Spain, Canada, Brazil, Ethiopia and Japan, and its most studied property is dimensionality. There is a consensus, mostly from factor analytical studies, that the PSS has a two-dimensional structure, composed of negatively worded and positively worded items (Lee, 2012). These two dimensions are consistent with Lazarus's (1966) theory and were named the "Perceived Stress" and "Perceived Control" subscales, although other terminologies such as "Perceived Distress" and "Perceived Coping" have also been used (Hewitt, Flett, & Mosher, 1992).

Considering the robust evidence regarding dimensionality, a few psychometric studies have started to evaluate differential item functioning (DIF). One main hypothesis analysed is if the PSS items are biased according to gender (Cole, 1999; Dougherty, Cooley, & Davidorf, 2017; Gitchel, Roessler, & Turner, 2011; Lavoie & Douglas, 2012; Nielsen, Skogstad, & Dammeyer, 2017; Taylor, 2015). Since women have consistently scored higher than men in the Perceived Stress subscale (but not on the Perceived Control subscale (Gitchel et al., 2011; Hewitt et al., 1992; Lavoie & Douglas, 2012)), a long-lasting debate in the PSS literature is if score differences are "an artifact of measurement bias" or "true gender differences arising from social, biological, or psychological influences" (Lee, 2012, p. 126). The findings regarding DIF by gender are mixed (Cole, 1999; Dougherty et al., 2017; Gitchel et al., 2011; Lavoie & Douglas, 2012; Nielsen & Dammeyer, 2019; Sharp, Kimmel, Kee, Saltoun, & Chang, 2007; Taylor, 2015). Although some studies indicated no evidence of DIF (Dougherty et al., 2017; Lavoie & Douglas, 2012; Taylor, 2015), Cole (1999) reported that PSS-10 items 3, 6, 7, 8 and 10 had DIF with a small magnitude and suggested that the "combination of the potentially slightly biased items may explain the apparent test level bias towards women" (p. 320). Gitchel et al. (2011) found DIF by gender for PSS-10 items 1, 3, 4 and 6, a result partially confirmed by T. Nielsen and Dammeyer (2019) (i.e. which also reported DIF for Items 1 and 3). Other sources of DIF have also been investigated. Regarding education, DIF was found for the PSS-10 items 3, 4, 8 and 9 (Cole, 1999), while other studies analyzed age, ethnicity, and literacy (Cole, 1999; Sharp, Kimmel, Kee, Saltoun, & Chang, 2007).

The analysis of local dependence (LD) of PSS items has only recently started (Medvedev et al., 2017; Nielsen & Dammeyer, 2019). The investigation of LD is especially relevant for the PSS since, in many of the PSS-14 studies, the two-factor structure accounted for less than 50% of the total variance (Lee, 2012). These findings suggest that a high percentage of the variance of item responses is not explained by the latent trait, and the PSS literature is still not clear regarding what these other influences could be.

Finally, the PSS-14 has previously displayed adequate reliability in different samples. The internal consistency reliability, measured by the Cronbach's α (1951), was higher than .70 in 11 of 12 studies, while the test-retest reliability was higher than .70 in 2 of 3 studies (Lee, 2012). However, since Cronbach's α provides a lower-bound estimate of reliability when items are locally independent (Novick & Lewis, 1967), the analysis of LD of PSS items is required to ensure that reliability estimates are not inflated (Medvedev et al., 2017).

Complementary Measures

The Perceived Stress Scale (PSS). The PSS is a five-point scale (1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly Agree) with a two-factor structure of perceived Stress (PS) and perceived Coping (PC) which evaluates if a person's life is perceived as unpredictable, uncontrollable, or overloading (S. Cohen et al., 1983).

The two complementary measures used in this study in the analysis of criterion validity were:

The Multidimensional Scale of Perceived Social Support (MSPSS): The MSPSS is a 12 item five-point scale (1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly

Agree), with a three-factor structure of family (FA), friend (FR) and significant others (SO) (Zimet, Powell, Farley, Werkman, & Berkoff, 1990). The MSPSS containing all 12 items (α=0.93) and the FA (α=0.92), FR (α=0.92) and SO (α=0.95) subscales displayed excellent reliability.

The Efforts-Reward Imbalance Questionnaire (ERI): A shorter version of the five-point scale (1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly Agree) ERI questionnaire with 11 items was used. The ERI questionnaire has a three-factor structure composed of effort (EF), reward (RD) and over commitment (OC) (Siegrist, Wege, Pühlhofer, & Wahrendorf, 2009). The ERI containing all 11 items (α=0.75) and the ER (α=0.85) and RD (α=0.73) subscales displayed adequate reliability. The OC (α=0.52) subscale displayed poor reliability and for this reason, was not included in the analysis of criterion validity.

The Rasch measurement models

The Rasch model (RM) is part of the family of Item Response Theory (IRT) models and it has two distinctive features over other IRT models: (1) the sum score is a sufficient statistic for the person parameter, containing all the information that allows statistical inference about the latent trait (Christensen, Kreiner, & Mesbah, 2013); and (2) inference can be conducted on a conditional framework (Erling Bernhard Andersen, 1970), since person and item parameters can be eliminated by means of conditional probabilities (Fisher, 2010), a property that Rasch (1997) referred as *specific objectivity*.

A mathematical property of the RM is the conditional independence of item responses to exogenous variables (i.e. absence of DIF) and to other items (i.e. local independence). However, in most rating scales applied in health sciences, items often show evidence of LD and DIF. Therefore, items with LD or DIF do not fit the RM (T. Nielsen & Kreiner, 2013) and common practice has been the deletion of items solely to obtain statistical fit to the model

(Kreiner & Christensen, 2007, 2011). This practice is problematic; if the deleted items cover important aspects of the construct, there is a threat to content validity (Lynn, 1986) and can lead to "construct underrepresentation" (Messick, 1987). In addition, the revised scale might end up being composed of a small number of items, leading to reduced reliability (T. Nielsen & Kreiner, 2013).

For this reason, recent methodological advances consist of analysis by Graphical Loglinear Rasch Model (GLLRM), which extends the RM with additional parameters to incorporate uniform LD and uniform DIF (Kreiner & Christensen, 2011). The term *uniform* refers to when the magnitude of the conditional dependence between items (LD) or between an item and an exogenous variable (DIF) is constant across the trait level. GLLRM is a combination of two independently developed statistical methods. The first method is the log-liner IRT models developed by Kelderman (1984; 1994), which generalizes IRT models to relax the assumption of local independence. The assumption of local independence is restrictive and frequently not achieved by questionnaires in health sciences. Therefore, log-liner IRT models allow locally dependent items, while representing traditional IRT models with locally independent items (e.g. Partial Credit model) as a special case. The second method is the development of Graphical models (Whittaker, 2009), which graphically represent the structure of conditional dependence between variables. Since in the RM the total score is sufficient statistic for the person parameter, graphical models are suitable for the analysis of LD and DIF. For example, to evaluate DIF, items and exogenous variables should be conditionally independent given the total score. The structure of conditional dependence between items, latent trait, and exogenous variables can then be represented graphically.

The functional form of a general GLLRM (containing one LD and one DIF parameter) can be expressed as:

$$\ln(P(Y = (y_1, \dots, y_k)|\theta, C)) = \lambda_0(\theta, x) + \sum_i \left(\alpha^i_{y_i} + y_i\theta\right) + \sum_{i,j} \lambda^{i,j}_{y_i y_j} + \sum_{i,j} \delta^{i,j}_{y_i c_j}$$

which describes the conditional distribution of a vector of item responses ($y_1,\ldots,y_k$) given the latent trait θ and exogenous variables C. The terms $\lambda_0(\theta, x) + \sum_i \left(\alpha^i_{y_i} + y_i\theta\right)$ are equivalent to the RM for polytomous items (i.e. Partial Credit model), while $\lambda^{i,j}_{y_i y_j}$ represents the interaction parameter between *item i* and *item j* and $\delta^{i,j}_{y_i c_j}$ represents the interaction parameter between *item i* and *exogenous variable j*. For an in-depth technical discussion of GLLRMs, please see (Kreiner & Christensen, 2007).

The usefulness of GLLRM is that, when questionaries exhibit uniform LD and uniform DF, departures from the RM does not necessarily imply that items are flawed: locally dependent items convey less information than independent items and lead to reduced reliability; items with DIF require scores to be adjusted to allow comparison between subgroups. However, in both cases, the item serves its original purpose of measuring the latent trait, and retaining these items is important to preserve construct validity. Furthermore, in both cases, the distinctive feature of the RM is preserved: if the uniform LD parameter is included the sufficiency of the total score is retained; while, if the uniform DIF parameter is present, the sufficiency of the total score is retained within the DIF-defined subgroups (Kreiner & Christensen, 2007). Finally, the uniform LD and DIF parameters can inform how items deviated from ideal measurement requirements and become a starting point for modifications on an instrument level (T. Nielsen & Kreiner, 2013). This approach aims to investigate *why* items did not fit the RM; and when departures consist of uniform LD and uniform DIF, it is possible to retain the items and inform future modifications on the instrument (T. Nielsen & Kreiner, 2013).

Statistical analysis

Item analysis: Item analysis was conducted with the following steps: (1) initially testing if the items would fit the RM (Masters, 1982); (2) if fit to the RM was rejected, the departures were investigated and catalogued; and (3) in case of uniform LD and uniform DIF, the fit to a

GLLRM adjusting for these departures was tested. In the case of other types of departures, such as items displaying evidence of being a poor measure of the construct, the most problematic item was removed and the three previous steps repeated. The estimation method for the RM and GLLRM was conditional maximum likelihood (Erling Bernhard Andersen, 1970). Person parameters were estimated using weighted maximum likelihood (WML) (Warm, 1989). Since missing values for individual items ranged from 0.0% to 1.3%, multiple imputation was not required (Graham, 2009). All statistical analyses were conducted with the DIGRAM v4.05 (Kreiner, 2003; Kreiner & Nielsen, 2013). Descriptive statistics and graphs were created with R software (R Core Team, 2013). The item analysis included the evaluation of: a) model fit; b) global DIF; c) item fit; d) LD; e) DIF; and f) unidimensionality. After a measurement model was established,: g) reliability and h) targeting of the instrument in this sample was evaluated.

Model fit and global DIF: Overall fit of the model was evaluated through the Conditional Likelihood Ratio (CLR) test (Erling B Andersen, 1973). The CLR test evaluates if item parameters are *invariant* between subgroups. The subgroups were defined according to lower and higher scores (i.e. homogeneity) to evaluate overall model fit; and according to sex (Male; Female) and education (education level up to High School; TAFE[1] or University) to evaluate Global DIF (Christensen et al., 2013).

Item fit: Fit of individual items was evaluated by conditional infit and outfit statistics, which, differently from traditional infit and outfit statistics, have a known sampling distribution (Christensen & Kreiner, 2012).

LD and DIF: To investigate LD and/or DIF, Kelderman's (1984) likelihood ratio (LR) test was conducted to test if the additional uniform LD or uniform DIF parameter would better explain the item responses compared to the fitted model. In addition, the *magnitude* of the uniform LD or uniform[2] DIF was evaluated through the partial Goodman-Kruskal (1954) γ

[1] Technical and Further Education (or TAFE) is the biggest provider of post-secondary education in Australia. TAFE offers a broad range of courses, at the operative, trade and paraprofessional level, that can last from a few hours (refreshment courses) to three years (diploma courses). Unlike universities, which are composed mostly of full-time students, TAFE institutions allow students to combine study and work, and encourage programs of apprenticeships and traineeships (Goozee, 2001).

[2] For simplicity, the term uniform is omitted when referred to uniform LD or uniform DIF from now on.

rank correlation between items given the two restscores or between item and exogenous variable given the total score (Kreiner & Christensen, 2004). In case DIF was present, the scores were adjusted and conversion tables reported (Kreiner & Christensen, 2007). When multiple tests were performed, the Benjamini-Hochberg (1995) procedure was conducted to adjust for false discovery rate (FDR).

Dimensionality: Initially, the RM and subsequent GLLRMs were tested for the PSS-14 containing all items. In case no fit was found, we then proceeded to test the two subscales composed of negatively and positively worded items. Finally, if a RM or GLLRM was found for each subscale, a formal test of unidimensionality was conducted by comparing the observed γ correlation of the subscales with the expected γ correlation of the subscales under the unidimensional model. The rationale is that the correlation between two subscales measuring different traits is weaker than the expected correlation of subscales measuring the same trait (Horton, Marais, & Christensen, 2012). Markov graphs (Kreiner & Christensen, 2007) were reported to illustrate the final models.

Reliability: In case of fit to the RM, reliability was estimated using Cronbach's α (1951), since it provides a lower-bound estimate of reliability (Novick & Lewis, 1967) when items are locally independent. However, when LD was found, a Monte Carlo simulation method (Hamon & Mesbah, 2002) that adjusts for the LD between items was applied. Since DIF implies that the *item thresholds* (and, consequently, the *item difficulty*) change according to subgroup, the different item parameters influence the true score distribution so reliability was calculated for each subgroup independently (J. B. Nielsen, Kyvsgaard, Sildorf, Kreiner, & Svensson, 2017). In addition, the person separation probability was calculated, which is the probability that the total scores rank two random persons in the same way as the *true* value of their latent trait (i.e. rather than the *estimates*) (Kreiner & Christensen, 2012).

Targeting: Targeting was evaluated through the Test Target Information Index, which consists of the mean test information divided by the maximum obtained test information

(Kreiner & Christensen, 2012). In addition, targeting was evaluated graphically through the inspection of item maps.

Criterion validity: Since scores are ordinal, the convergent and divergent validity of the PSS with other psychological constructs pertaining to its nomological network (Cronbach & Meehl, 1955) was evaluated by calculating the non-parametric Kendall's τ (Kendall, 1948). For this analysis, the complementary measures were used A negative correlation of Perceived Stress with FA, FR, SG, and RW, and a positive correlation with EF and OC was anticipated. In addition, know-groups validity (Davidson, 2014) was assessed and it was expected that women would have higher scores on the Perceived Stress subscale (Lee, 2012) but no difference in scores on the Perceived Control subscale (Gitchel, Roessler, & Turner, 2011; Hewitt et al., 1992; Lavoie & Douglas, 2012). It was also expected that participants with less education would have higher scores on the Perceived Stress subscale (Lee, 2012).

Results

The demographic characteristics of the sample are found in Table 1. Participants age ranged from 18 to 82 years (M = 50.2, SD = 14.8). The majority of participants were women (61.9%), had tertiary education (67.5%) and were employed (59%).

(Table 1)

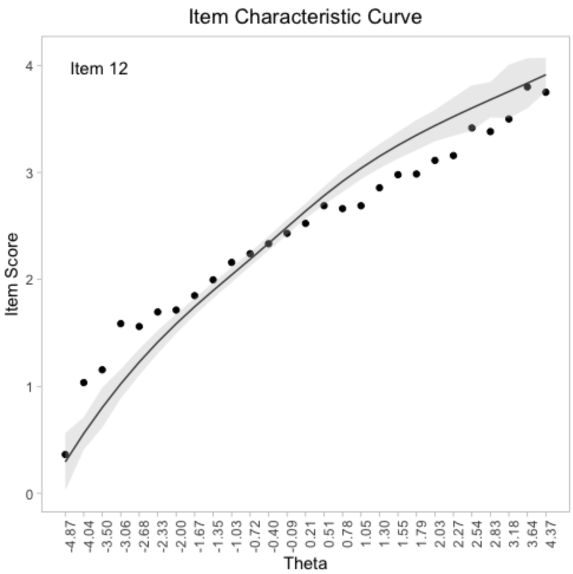
PSS-14: Fit of the PSS-14 to the RM was rejected (Table 2).

(Table 2)

The results indicated item misfit (Supplementary Table 2) among the majority of items. The analysis proceeded by sequentially excluding items, such as items 4, 5, 9, 12, 13, and 6 that displayed the highest misfit, while investigating departures in terms of LD and DIF with GLLRMs. However, it became clear that: a) LD and DIF could not explain the misfit to the RM and GLLRMs were not found; and b) the majority of excluded items were negatively worded, indicating that they would not form a unidimensional scale together with the positively worded items. At this point, we proceeded to the analysis of the subscales.

Perceived Stress subscale: Fit of the negatively worded items ("Perceived Stress") subscale to the RM was rejected (Table 2). The investigation of item fit statistics (Supplementary Table 3) indicated strong misfit of Item 12 ("…found yourself thinking about all the things you have to accomplish?") (Infit = 1.675, SE=0.023, p<0.001; Outfit=1.669, SE=0.023, p<0.001) (Figure 1).

Figure 1. Item characteristic curve for Item 12.



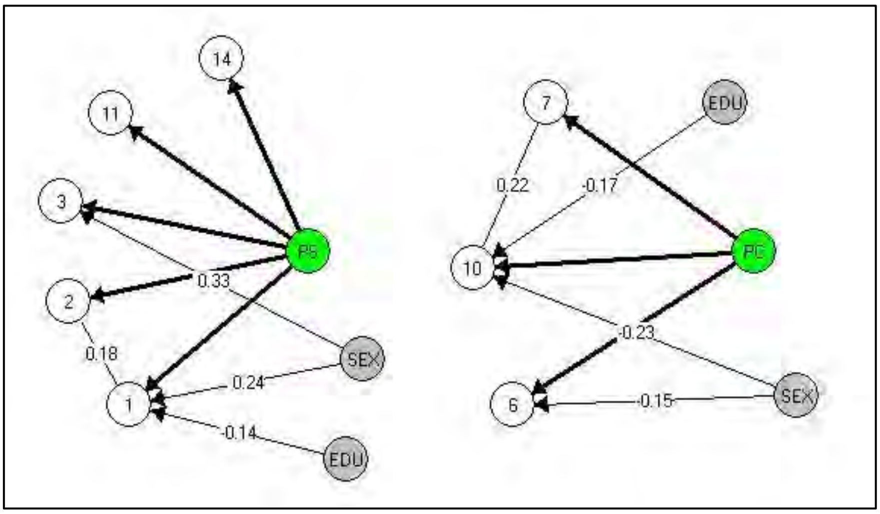
Note. The x-axis indicates the latent trait and the y-axis indicates the item score. The black points represent the observed item responses for each total score. The grey curve is the expected item responses and the grey shaded area is the 95% confidence region.

Figure 1 shows that the average observed scores exhibited a pattern of *under discrimination* since they formed a flat curve compared to the model expectations, indicating that item responses were less influenced by the latent trait ("perceived stress"). It was then evaluated whether Item 12 misfit could be a result of DIF or LD (i.e. although LD often results in *over discrimination*) but a GLLRM was not found. For these reasons, Item 12 was excluded.

After the deletion of Item 12, the CLR test rejected fit to the RM (χ^2 (23) = 312.9, p<0.001) and the next item that displayed misfit was Item 8 ("…felt unable to cope with all the things that you had to do?") (Infit = 1.145, SE=0.023, p<0.001; Outfit=1.155, SE=0.023, p<0.001). The analysis indicated that Item 8 misfit was also not a result of LD or DIF and Item 8 was also excluded.

GLLRM of the Perceived Stress subscale: After exclusion of the two items, the CLR test rejected fit to the RM but fit to a GLLRM was found (χ^2 (96) = 94.4, p=0.440) (Table 2) (Figure 2).

Figure 2. GLLRMs of the Perceived Stress subscale (left) and Perceived Control subscale (right).



Note. The Markov graph nodes represent the item numbers, the exogenous variables, and the latent trait. Disconnected nodes indicate that variables are conditionally independent and partial γ informs the magnitude of the LD and DIF.

LD was found between Item 1 ("…felt upset because of something that happened unexpectedly?") and Item 2 ("… felt unable to control the important things in your life?") (γ_{avg}=0.18). DIF was found between Item 1 and sex (γ=0.24); between Item 3 ("… felt either nervous or stressed?") and sex (γ=0.33); and between Item 1 and education (γ=-0.14). There was no item misfit (Table 3), and the Kelderman's LR test indicated no further evidence of DIF or LD (Supplementary Table 4).

(Table 3)

Considering that the GLLRM had overall model fit and there was no further evidence of global DIF, item misfit, DIF or LD, the measurement model for the "Perceived Distress" subscale was established.

Perceived Control subscale: Fit of the positively worded items ("Perceived Control") subscale to the RM was rejected (Table 2). Misfit was found among the majority of items (Supplementary Table 5). The item with the highest misfit was Item 9 ("…felt able to control irritations in your life?") (Infit = 1.367, SE=0.026, p<0.001; Outfit=1.237, SE=0.023, p<0.001) and it was excluded. On the subsequent analysis, substantial misfit was also found regarding Item 13 ("… felt able to control the way you spend your time?") (Infit = 1.363, SE=0.036, p<0.001; Outfit=1.180, SE=0.032, p<0.001), Item 4 ("… dealt successfully with irritating life hassles?") (Infit = 1.226, SE=0.024, p<0.001; Outfit=1.185, SE=0.024, p<0.001) and Item 5 ("...effectively coped with important changes in your life?") (Infit = 1.571, SE=0.024, p<0.001; Outfit=1.501, SE=0.024, p<0.001) and these items were removed.

GLLRM of the Perceived Control subscale: After the exclusion of the misfitting items, the CLR test indicates fit to a GLLRM (χ^2 (55) = 62.5, p=0.224) (Table 2) (Figure 2). LD was found between Item 7 ("… felt things were going your way?") and Item 10 ("… felt you were on top of things?") (γ_{avg}=0.22). DIF was found between Item 10 and sex (γ=-0.23); between Item 6 ("...felt confident about your ability to handle your personal problems?") and sex (γ=-0.15); and between Item 10 and education (γ=-0.17). There was no further evidence of item misfit (Supplementary Table 6) or LD/DIF (Supplementary Table 7). Considering that the GLLRM had overall model fit and there was no further evidence of global DIF, item misfit, LD or DIF, the measurement model for the "Perceived Control" subscale was established.

The table for adjusting scores after accounting for DIF is provided for both subscales (Supplementary Table 8).

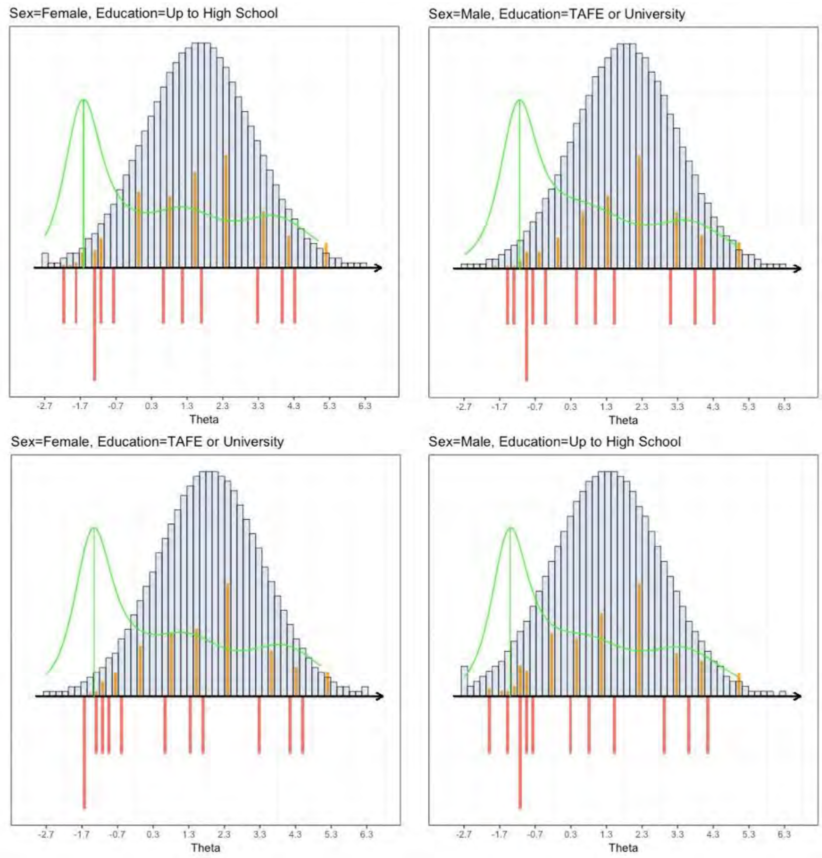
Dimensionality: Since the observed correlation between the Perceived Stress and Perceived Control subscales ($\gamma = 0.527$) was weaker than the expected correlation between the two subscales ($\gamma = 0.569$, SE = 0.009, $p < 0.001$) under a unidimensional model, the unidimensionality of the PSS-14 was rejected. Therefore, unidimensionality was confirmed *within* subscales but not *between* subscales, indicating that the Perceived Stress subscale and the Perceived Control subscale measure qualitatively distinct psychological traits.

Targeting and reliability: For the Perceived Stress subscale, the targeting was moderate. The overall Test Information Target Index indicates that for the Australian population the Perceived Stress subscale provided only 60% of the total information available if the instrument was perfectly targeted. Values ranged from 56% to 62% within subgroups (Table 4). For example, women who completed TAFE or University had an average total score of 8.48 (SD = 3.65), while the Perceived Stress subscale was perfectly targeted for a population with an average score of 14.79 (SE = 1.97). The overall reliability was 0.84. The overall person separation probability was 83%, indicating that if two respondents were randomly selected and then ranked on their total score, in 83% of cases they will be ranked correctly according to their true level of perceived stress.

(Table 4)

For the Perceived Control subscale, targeting was poor. The overall Test Information Target Index indicated that 34% of the total information was attained (Table 4) (Figure 3). The overall reliability was 0.74 and the overall person separation probability was 75%.

Figure 3. Item Map of the Perceived Control subscale according to subgroups



Note. The orange bars display the person parameters (WML estimates). The grey bars display the population distribution of Perceived Control under the assumption of normality. The red bars display the item thresholds and the green line is the information function.

Criterion validity: The Perceived Stress and Perceived Control subscales displayed the expected patterns of convergence and divergence regarding the complementary measures (Supplementary Table 9). The analysis of known-groups validity indicated that women had higher scores of perceived stress ($diff_{adj}$=0.67) but no substantial difference regarding perceived control ($diff_{adj}$=0.04). Participants with education up to high school had lower scores on perceived control ($diff_{adj}$=0.-50) but showed no substantial difference in perceived stress ($diff_{adj}$=0.05) (Table 5).

(Table 5)

Discussion

The aim of the present study was to evaluate if the PSS-14 constitutes a valid and reliable instrument to measure perceived stress in Australia. The results indicate that: 1) the revised version of the Perceived Stress subscale displayed adequate psychometric properties and provides a measure of perceived stress; however, 2) the majority of psychometric problems centered around the Perceived Control subscale. The implications for future use of the Perceived Stress Scale in Australia are discussed.

Dimensionality: The results from the present study indicated that the PSS-14 is not unidimensional but rather composed of two dimensions. The observed correlation between the Perceived Stress and Perceived Control subscales ($\gamma = 0.527$) was strong but weaker than expected under a unidimensional model. The conclusion towards two dimensions (rather than one) was based not only from the dimensionality analysis but also on the theoretical background of the PSS (Lee, 2012). The interpretation is that, although the two constructs of perceived stress and perceived control are *correlated* – as they are expected to be since according to Lazarus (1966) events are perceived as stressful when there is a perception of insufficient control over the situation – these constructs are nonetheless *qualitatively distinct*.

Perceived Stress subscale: The Perceived Stress subscale displayed adequate psychometric properties after the deletion of two items. The problems found with Item 12 ("…have you found yourself thinking about all the things you have to accomplish?"), which was excluded in the original validation conducted by Cohen (1988), have been extensively reported (de la Rubia & de León, 2014; Ezzati et al., 2014; Faro, 2015; Lavoie & Douglas, 2012; Pedrero Pérez & Olivar Arroyo, 2010; Ramírez & Hernández, 2007; Wu & Amtmann, 2013). It has been shown, for example, that Item 12 was endorsed by respondents with low *and* high levels of perceived stress, since "thinking about all the things you have to accomplish" does not necessarily means being overwhelmed by them but also constitutes a self-management behaviour (de la Rubia & de León, 2014). Studies that reported problems with Item 8 were less common (Taylor, 2015; Yokokura et al., 2017). Finally, the Perceived Stress subscale displayed the expected pattern of convergent/divergent validity and known-groups validity except for education, providing further support for construct validity in the Australian population.

DIF and Gender bias: The findings of the current study were also consistent with the recent PSS literature regarding DIF. When DIF was investigated in relation to sex, DIF was found for Item 1 (Gitchel et al., 2011; T. Nielsen, Skogstad, & Dammeyer, 2017), Item 3 (Cole, 1999; Gitchel et al., 2011; T. Nielsen et al., 2017), Item 6 (Gitchel et al., 2011) and Item 10 (Cole, 1999), similarly to previous studies. Rather than a characteristic specific to Australian respondents, the DIF of these items seems to be a consequence of *gender roles* present in Western societies, as documented by a robust body of psychological literature (Fleishman, Spector, & Altman, 2002; Lange, Thalbourne, Houran, & Lester, 2002; Shevlin, Bailey, & Adamson, 2002). The traditional female gender role prescribes emotional expressiveness and lack of assertiveness, while the traditional male role prescribes assertiveness and self-confidence (Matud, 2004). Matud (2004, p. 1403) explains that "The stress associated with gender role identification is different for each sex because women are more likely to identify with the feminine gender role, and men are more likely to identify with the masculine gender role". This is known as the *socialization hypothesis* (Rosario, Shinn, Mørch, & Huckabee, 1988) and the influence of gender roles on item response patterns has been previously reported in stress research. For example, Smith and Reise (1998) showed that, compared to men with the same level of stress, women more frequently endorse items regarding emotional vulnerability and sensitivity.

In the present study, this DIF pattern was found in Item 1 ("…felt upset because of something that happened unexpectedly?") (γ=0.24) and Item 3 ("…felt either nervous or stressed?") (γ=0.33), which were more frequently endorsed by women. An opposite pattern was found in Item 6 ("...felt confident about your ability to handle your personal problems?") (γ=-0.15) and Item 10 ("… felt you were on top of things?") (γ=-0.23), which were systematically endorsed by men. One possible explanation for these phenomena is that masculinity stereotypes in Western societies emphasize success, competition and *being in control*. Therefore, one possible explanation is that gender roles influenced response patterns and men were less likely to acknowledge negative emotions (Van Dam, Earleywine, & Forsyth, 2009) and more likely to acknowledge self-confidence (Matud, 2004). The pressure to hide vulnerabilities leads to underreporting of psychological symptoms among men and long-term consequences are under diagnosis and under treatment, creating a “silent epidemic” of mental illness (Real, 1998, p. 386; Strother, Lemberg, Stanford, & Turberville, 2012).

One main contribution of the present study is to provide evidence to the long-standing debate of “gender-related differences in PSS scores” (Lee, 2012, p. 126). The results demonstrated that women had higher levels of perceived stress even after scores were adjusted for measurement bias ($diff_{adj}$= 0.67; $diff_{obs}$ = 1.07), since bias was responsible for 37% of the difference. Therefore, the differences of perceived stress scores between men and women in Australia are not explained by measurement bias alone and can be interpreted as true differences arising from social, biological and psychological influences (Lavoie & Douglas, 2012). However, it is necessary for future studies to investigate the impact of these differences. For example, the impact generated by a 0.67 higher average score in terms of use of the health system, psychopathology, disability leave, among others.

When DIF was analysed with respected to education, DIF was found for Item 1 and Item 10 (“… felt you were on top of things?”). This result is congruent with Cole (1999), who also showed that, given the same level of perceived control, participants with higher education were more likely to believe they *were on top of things*. Recent findings have suggested that perceived control is affected by educational attainment and is a mediator of health behaviours. For example, individuals with more educational attainment had a stronger belief that their actions would produce desirable outcomes (e.g. exercise and dieting would prevent developing disease) and had less fatalism (Park, Cho, & Moore, 2018). Additionally, feeling *on top of things* might also be interpreted as the relationship between higher education and status in Western societies (Fantuzzo, 2018).

Since DIF was present among many of the PSS-14 items, a fundamental recommendation of the present study is that future applications of the Perceived Stress Scale in Australia need to use the conversion table (Supplementary Table 7) to adjust total scores and avoid measurement bias. The presence of DIF is a threat to construct validity since observed scores cannot be interpreted as reflecting true differences of perceived stress/perceived control. Therefore, if total scores are used without adjustment, the comparisons between subgroups are invalid.

Response dependence: The present study showed positive LD between Item 1 ("…felt upset because of something that happened unexpectedly?") and Item 2 ("… felt unable to control the important things in your life?") (γ_{avg} =0.18), and between Item 7 ("… felt things were going your way?") and Item 10 ("… felt you were on top of things?") (γ_{avg} =0.22). The dependence between Item 1 and 2 (Medvedev et al., 2017), and between Item 7 and 10 (Medvedev et al., 2017; T. Nielsen et al., 2017) have been previously reported; while the dependence between Item 7 and Item 10 found in Australia (γ_{avg}=0.22) was also found in Danish students with a similar magnitude (γ_{avg}=0.24) (T. Nielsen et al., 2017). In these two pairs of items, the dependence seems to be a case of *response dependence* (Andrich & Kreiner, 2010; Marais & Andrich, 2008). For example, given the same trait level, respondents who endorsed the Item 7 ("… felt things were going your way?") had a higher probability of endorsing the Item 10 ("… felt you were on top of things?") than those who did not endorse the former. This seems to happen because *feeling on top of the things* in most cases logically implies that *things were going your way*.

Problems with the Perceived Control subscale: The majority of psychometric problems were found on the Perceived Control subscale. Problems with the excluded Item 4 ("… dealt successfully with irritating life hassles?"), Item 5 ("...effectively coped with important changes in your life?") and Item 13 ("… felt able to control the way you spend your time?") have been reported by many (Benoy et al., 2018; Mondo, Sechi, & Cabras, 2019; Örücü & Demir, 2009; Perera et al., 2017). Therefore, in conjunction with Item 12 from the Perceived Stress subscale, the exclusion of these three items indicate that the four items that were removed in the original validation by Cohen (1988) that led to the creation of the PSS-10 once again performed poorly in Australia. For this reason, the application of the original PSS-14 in Australia is not warranted.

Furthermore, with the additional exclusion of Item 9 ("…felt able to control irritations in your life?"), there are two implications for future studies. Firstly, the Perceived Control subscale was initially developed to be a *seven-item measure* of perceived coping/control through the theoretical perspective of Lazarus (1966). However, with the majority of items performing poorly, it seems unclear whether the three remaining items are enough to cover this psychological construct and pose concerns regarding *construct underrepresentation* (Messick, 1987). Secondly, a subscale composed of three items might have reduced reliability, as happened in the current study, in which the overall reliability of the Perceived Control subscale was only moderate (R=0.74) (Nunnally & Bernstein, 1967). Therefore, the findings of this study suggest that: a) new items should be developed for the Perceived Control subscale to ensure construct validity for an Australian population; and b) if the 3-item Perceived Control subscale is applied, the results should be interpreted with caution.

Theoretical contributions and limitations. The current study provides theoretical contributions to the validity of the PSS and to stress measurement. This study confirms the well-established findings regarding the two-dimensional structure of the PSS ("Perceived Stress" and "Perceived Control" subscales) and the preference towards the PSS-10 over the PSS-14 version due to 4 misfitting items. We also confirmed recent findings of DIF by gender of items 1 and 3, more easily endorsed by women, and items 6 and 10, more easily endorsed by men. We hypothesize that this DIF *pattern* is a consequence of gender roles present in Western societies, a response pattern similar to what has been reported in other stress measures (Smith & Reise, 1998). We contribute to stress measurement by investigating whether score differences represent true gender differences or are solely a consequence of measurement bias. We showed that, although there is measurement bias due to DIF, this bias accounted for only 37% of score differences and the remaining difference in stress levels between men and women is real. This study also advances the literature of PSS validity by investigating local dependence and targeting. We revealed that items 1 and 2, and 7 and 10 showed patterns of positive local dependence and that, if not taken into account, the dependence can lead to inflated estimates of reliability. Furthermore, we showed that the PSS is poorly targeted for a general high-income country population and is possibly better targeted for groups at risk of stress, such as students (T. Nielsen & Dammeyer, 2019).

One limitation of the present study is that the data available was from a national study conducted from 2004 to 2006. Considering that stress levels have increased over the years (Australian Psychological Society, 2015), the difference in the population distribution limits the *norm-referenced* use of test scores (Bond, 1996). That is, the use of the current sample as a *normative sample* should be used with caution since the sample distribution does not correspond to the current population distribution in Australia. Furthermore, future studies should investigate *item parameter drift*, whether the functioning of items remains stable over time (Goldstein, 1983).

Conclusion

Research over half a decade has suggested high levels of stress in Australia, leading to critical consequences such as increased use of alcohol, cigarettes, and gambling as coping mechanisms. The present research showed that the Perceived Stress subscale is a valid and reliable measure of perceived stress after the deletion of two items. The majority of psychometric problems centered on the Perceived Control subscale. After the exclusion of four items, it is encouraged that new items should be developed to ensure construct representation or, if the short-form scale is applied, results should be interpreted with caution.

Finally, a fundamental recommendation is that future applications need to use the conversion table to adjust total scores for measurement bias. If total scores are used without adjustment, the comparisons between population groups in Australia are invalid.

References

American Psychological Association. (2015). Stress in America: Paying with our health.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. Journal of the Royal Statistical Society. Series B (Methodological), 283-301.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. Psychometrika, 38(1), 123-140.

Andreou, E., Alexopoulos, E. C., Lionis, C., Varvogli, L., Gnardellis, C., Chrousos, G. P., & Darviri, C. (2011). Perceived Stress Scale: reliability and validity study in Greece. International Journal of Environmental Research and Public Health, 8(8), 3287-3298. doi:10.3390/ijerph8083287

Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. Applied Psychological Measurement, 34(3), 181-192.

Archibald, L. (2006). Decolonization and healing: indigenous experiences in the United States, New Zealand, Australia and Greenland: Aboriginal Healing Foundation Ottawa.

Australia Associated Press. (2014). Australian women feel more stressed than men, mental health survey finds. The Guardian. Retrieved from https://www.theguardian.com/australia-news/2014/nov/09/australian-women-feel-more-stressed-than-men-mental-health-survey-finds

Australian Psychological Society. (2015). Stress & wellbeing: how Australians are coping with life: the findings of the Australian Psychological Society Stress and Wellbeing in Australia survey 2015. Melbourne, Victoria: Australian Psychological Society.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 289-300.

Benoy, C., Knitter, B., Knellwolf, L., Doering, S., Klotsche, J., & Gloster, A. T. (2018). Assessing psychological flexibility: Validation of the Open and Engaged State Questionnaire. Journal of Contextual Behavioral Science.

Bond, L. A. (1996). Norm-and criterion-referenced testing. Practical assessment, research & evaluation, 5(2), 120-125.

Casey, L., & Liang, R. P.-T. (2014). Stress and wellbeing in Australia survey 2014. Australian Psychological Society.

Casey, L., & Mathews, R. (2011). Stress and wellbeing in Australia in 2011: a state-of-the-nation survey. InPsych: The Bulletin of the Australian Psychological Society Ltd, 33(6), 32.

Christensen, K. B., & Kreiner, S. (2012). Item fit statistics. Rasch models in health, 83-104.

Christensen, K. B., Kreiner, S., & Mesbah, M. (2013). Rasch models in health. London: John Wiley & Sons.

Cohen, & Wills, T. A. (1985). Stress, social support, and the buffering hypothesis. Psychological Bulletin, 98(2), 310.

Cohen, S., Janicki-Deverts, D., & Miller, G. E. (2007). Psychological stress and disease. JAMA, 298(14), 1685-1687.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A Global Measure of Perceived Stress. Journal of Health and Social Behavior, 24(4), 385. doi:10.2307/2136404

Cohen, S., & Williamson, G. (1988). Psychological stress in a probability sample of the United States. In S. Spacapan & S. Oskamp (Eds.), The social psychology of health:

Claremont symposium on applied social psychology (pp. 31-37). Newbury Park, CA: Sage Publications.

Cole, S. R. (1999). Assessment of differential item functioning in the Perceived Stress Scale-10. Journal of Epidemiology and Community Health, 53(5), 319.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297-334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281.

Davidson, M. (2014). Known-groups validity. In Encyclopedia of quality of life and well-being research (pp. 3481-3482): Springer.

de la Rubia, J. M., & de León, F. C. (2014). VALIDACIÓN DE LA ESCALA DE ESTRÉS PERCIBIDO (PSS-14) EN LA POBLACIÓN DE DENTISTAS COLEGIADOS DE MONTERREY. Ansiedad Y Estrés, 20.

Dudin, M. N., Gayduk, V. I., Sekerin, V. D., Bank, S. V., & Gorohova, A. E. (2016). Financial crisis in Greece: challenges and threats for the global economy. International Journal of Economics and Financial Issues, 6(5S).

Ezzati, A., Jiang, J., Katz, M. J., Sliwinski, M. J., Zimmerman, M. E., & Lipton, R. B. (2014). Validation of the Perceived Stress Scale in a community sample of older adults. International Journal of Geriatric Psychiatry, 29(6), 645-652.

Fantuzzo, J. (2018). Admitting a Sense of Superiority: Aggrandized Higher Education Status as an Objection to Educational Inequality. Studies in Philosophy and Education, 1-15.

Faro, A. (2015). Confirmatory factor analysis of three versions of the Perceived Stress Scale (PSS): a population-based study. Psicologia: Reflexão e Crítica, 28(1), 21-30.

Fisher, W. P. (2010). Rasch, Frisch, and two Fishers: A social history of the econometric origins of some widely used psychometric models.

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 57(5), S275-S284.

Garvey, D. (2008). Review of the social and emotional wellbeing of Indigenous Australian peoples. Retrieved from

Gitchel, W. D., Roessler, R. T., & Turner, R. C. (2011). Gender effect according to item directionality on the perceived stress scale for adults with multiple sclerosis. Rehabilitation Counseling Bulletin, 55(1), 20-28.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. Journal of Educational Measurement, 369-377.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. Journal of the American statistical association, 49(268), 732-764.

Goozee, G. (2001). The development of TAFE in Australia. Adelaide: National Centre for Vocational Education Research.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. Annual Review of Psychology, 60, 549-576.

Hamon, A., & Mesbah, M. (2002). Questionnaire reliability under the Rasch model. In Statistical methods for quality of life studies (pp. 155-168): Springer.

Harkness, K. L., & Monroe, S. M. (2016). The assessment and measurement of adult life stress: Basic premises, operational principles, and design requirements. Journal of Abnormal Psychology, 125(5), 727.

Hewitt, P. L., Flett, G. L., & Mosher, S. W. (1992). The Perceived Stress Scale: Factor structure and relation to depression symptoms in a psychiatric sample. Journal of Psychopathology and Behavioral Assessment, 14(3), 247-257.

Horton, M., Marais, I., & Christensen, K. B. (2012). Dimensionality. Rasch models in health, 137-158.

Karam, F., Bérard, A., Sheehy, O., Huneau, M. C., Briggs, G., Chambers, C., . . . Koren, G. (2012). Reliability and validity of the 4-item Perceived Stress Scale among pregnant women: Results from the OTIS antidepressants study. Research in Nursing and Health, 35(4), 363-375.

Kechter, A., Black, D. S., Riggs, N. R., Warren, C. M., Ritt-Olson, A., Chou, C.-P., & Pentz, M. A. (2019). Factors in the Perceived Stress Scale Differentially Associate with Mindfulness Disposition and Executive Function among Early Adolescents. J Child Fam Stud, 1-8.

Kelderman, H. (1984). Loglinear Rasch model tests. Psychometrika, 49(2), 223-245.

Kendall, S. M. (1948). Rank correlation. Van Nostrand's Scientific Encyclopedia.

Klein, E. M., Brähler, E., Dreier, M., Reinecke, L., Müller, K. W., Schmutzer, G., . . . Beutel, M. E. (2016). The German version of the Perceived Stress Scale–psychometric characteristics in a representative German community sample. BMC Psychiatry, 16(1), 159.

Kowal, E., Gunthorpe, W., & Bailie, R. S. (2007). Measuring emotional and social wellbeing in Aboriginal and Torres Strait Islander populations: an analysis of a Negative Life Events Scale. International Journal for Equity in Health, 6(1), 18.

Kreiner, S. (2003). Introduction to DIGRAM: Department of Biostatistics, University of Copenhagen Copenhagen, Denmark.

Kreiner, S., & Christensen, K. B. (2004). Analysis of local dependence and multidimensionality in graphical loglinear Rasch models. Communications in Statistics-Theory and Methods, 33(6), 1239-1276.

Kreiner, S., & Christensen, K. B. (2007). Validity and objectivity in health-related scales: Analysis by graphical loglinear Rasch models. In Multivariate and mixture distribution Rasch models (pp. 329-346): Springer.

Kreiner, S., & Christensen, K. B. (2011). Item screening in graphical loglinear Rasch models. Psychometrika, 76(2), 228-256.

Kreiner, S., & Christensen, K. B. (2012). Person parameter estimation and measurement in Rasch models. Rasch models in health, 63-78.

Kreiner, S., & Nielsen, T. (2013). Item Analysis in Digram-Notes on the use of DIGRAM for item analysis by graphical loglinear Rasch models. Department of Biostatistics-University of Copenhagen.

Lakey, B., & Cohen, S. (2000). Social support and theory. Social support measurement and intervention: A guide for health and social scientists, 29.

Lange, R., Thalbourne, M. A., Houran, J., & Lester, D. (2002). Depressive response sets due to gender and culture-based differential item functioning. Personality and Individual Differences, 33(6), 937-954.

Lavoie, J. A., & Douglas, K. S. (2012). The Perceived Stress Scale: Evaluating configural, metric and scalar invariance across mental health status and gender. Journal of Psychopathology and Behavioral Assessment, 34(1), 48-57.

Lazarus, R. S. (1966). Psychological stress and the coping process. New York, NY: McGraw-Hill.

Lee, E.-H. (2012). Review of the psychometric evidence of the perceived stress scale. Asian Nursing Research, 6(4), 121-127.

Leung, D. Y., Lam, T.-h., & Chan, S. S. (2010). Three versions of Perceived Stress Scale: validation in a sample of Chinese cardiac patients who smoke. BMC Public Health, 10(1), 513.

Lynn, M. R. (1986). Determination and quantification of content validity. Nursing Research.

Manzar, M. D., Salahuddin, M., Peter, S., Alghadir, A., Anwer, S., Bahammam, A. S., & Pandi-Perumal, S. R. (2019). Psychometric properties of the perceived stress scale in Ethiopian university students. BMC Public Health, 19(1), 41.

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. J Appl Meas, 9(3), 200-215.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2), 149-174.

Matud, M. P. (2004). Gender differences in stress and coping styles. Personality and Individual Differences, 37(7), 1401-1415.

Medvedev, O. N., Krägeloh, C. U., Hill, E. M., Billington, R., Siegert, R. J., Webster, C. S., . . . Henning, M. A. (2017). Rasch analysis of the Perceived Stress Scale: Transformation from an ordinal to a linear measure. Journal of Health Psychology, 1359105316689603.

Messick, S. (1987). Validity. ETS Research Report Series, 1987(2), i-208.

Mondo, M., Sechi, C., & Cabras, C. (2019). Psychometric evaluation of three versions of the Italian Perceived Stress Scale. Current Psychology, 1-9.

Nielsen, J. B., Kyvsgaard, J. N., Sildorf, S. M., Kreiner, S., & Svensson, J. (2017). Item analysis using Rasch models confirms that the Danish versions of the DISABKIDS® chronic-generic and diabetes-specific modules are valid and reliable. Health and Quality of Life Outcomes, 15(1), 44.

Nielsen, M. G., Ørnbøl, E., Vestergaard, M., Bech, P., Larsen, F. B., Lasgaard, M., & Christensen, K. S. (2016). The construct validity of the Perceived Stress Scale. Journal of Psychosomatic Research, 84, 22-30.

Nielsen, T., & Dammeyer, J. (2019). Measuring higher education students' perceived stress: An IRT-based construct validity study of the PSS-10. Studies in Educational Evaluation, 63, 17-25.

Nielsen, T., & Kreiner, S. (2013). Improving items that do not fit the Rasch model. Rasch models in health, 317-334.

Nielsen, T., Skogstad, S., & Dammeyer, J. (2017). The Perceived Stress Scale ia higher education context: a Rasch-based construct validity study: Book of Abstracts page 95. Paper presented at the 14th European Conference on Psychological Assessment.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. Psychometrika, 32(1), 1-13.

Nunnally, J. C., & Bernstein, I. H. (1967). Psychometric theory (Vol. 226): McGraw-Hill New York.

Örücü, M. Ç., & Demir, A. (2009). Psychometric evaluation of perceived stress scale for Turkish university students. Stress and Health: Journal of the International Society for the Investigation of Stress, 25(1), 103-109.

Papanicolas, I., Woskie, L. R., & Jha, A. K. (2018). Health care spending in the United States and other high-income countries. JAMA, 319(10), 1024-1039.

Park, C. L., Cho, D., & Moore, P. J. (2018). How does education lead to healthier behaviours? Testing the mediational roles of perceived control, health literacy and social support. Psychology & Health, 1-14.

Patterson, P. G., & Smith, T. (2001). Modeling relationship strength across service types in an Eastern culture. International journal of service industry management, 12(2), 90-113.

Pedrero Pérez, E. J., & Olivar Arroyo, Á. (2010). Estrés percibido en adictos a sustancias en tratamiento mediante la escala de Cohen: propiedades psicométricas y resultados de su aplicación. Anales de psicología, 26(2).

Perera, M. J., Brintz, C. E., Birnbaum-Weitzman, O., Penedo, F. J., Gallo, L. C., Gonzalez, P., . . . Perreira, K. M. (2017). Factor structure of the Perceived Stress Scale-10 (PSS) across English and Spanish language responders in the HCHS/SOL Sociocultural Ancillary Study. Psychological Assessment, 29(3), 320.

Pong, R. W., DesMeules, M., & Lagacé, C. (2009). Rural–urban disparities in health: How does Canada fare and how does Canada compare with Australia? Australian Journal of Rural Health, 17(1), 58-64.

R Core Team. (2013). R: A language and environment for statistical computing.

Ramírez, M. T. G., & Hernández, R. L. (2007). Factor structure of the Perceived Stress Scale (PSS) in a sample from Mexico. The Spanish journal of psychology, 10(1), 199-206.

Rasch, G. (1997). On specific objectivity: An attempt of formalizing the generality and validity of scientific statements.

Real, T. (1998). I don't want to talk about it: Overcoming the secret legacy of male depression: Simon and Schuster.

Rosario, M., Shinn, M., Mørch, H., & Huckabee, C. B. (1988). Gender differences in coping and social supports: Testing socialization and role constraint theories. Journal of Community Psychology, 16(1), 55-69.

Sanders, A. E., & Slade, G. D. (2011). Gender modifies effect of perceived stress on orofacial pain symptoms: National Survey of Adult Oral Health. Journal of Orofacial Pain, 25(4).

Santiago, P. H. R., Roberts, R., Smithers, L. G., & Jamieson, L. (2019). Stress beyond coping? A Rasch analysis of the Perceived Stress Scale (PSS-14) in an Aboriginal population. PloS One, 14(5), e0216333.

Shevlin, M., Bailey, F., & Adamson, G. (2002). Examining the factor structure and sources of differential functioning of the Eysenck Personality Questionnaire Revised—Abbreviated. Personality and Individual Differences, 32(3), 479-487.

Siegrist, J. (1996). Adverse health effects of high-effort/low-reward conditions. Journal of Occupational Health Psychology, 1(1), 27.

Siegrist, J., Wege, N., Pühlhofer, F., & Wahrendorf, M. (2009). A short generic measure of work stress in the era of globalization: effort–reward imbalance. International Archives of Occupational and Environmental Health, 82(8), 1005.

Slade, G. D., Spencer, A. J., & Roberts-Thomson, K. F. (2004). Australia's dental generations. The national survey of adult oral health, 6(2007), 274.

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction scale. Journal of Personality and Social Psychology, 75(5), 1350.

Strother, E., Lemberg, R., Stanford, S. C., & Turberville, D. (2012). Eating disorders in men: underdiagnosed, undertreated, and misunderstood. Eating disorders, 20(5), 346-355.

Taylor, J. M. (2015). Psychometric analysis of the Ten-Item Perceived Stress Scale. Psychological Assessment, 27(1), 90.

Thoits, P. A. (1995). Stress, coping, and social support processes: Where are we? What next? Journal of Health and Social Behavior, 53-79.

Van Dam, N. T., Earleywine, M., & Forsyth, J. P. (2009). Gender bias in the sixteen-item Anxiety Sensitivity Index: An application of polytomous differential item functioning. Journal of Anxiety Disorders, 23(2), 256-259.

Wang, Z., Chen, J., Boyd, J. E., Zhang, H., Jia, X., Qiu, J., & Xiao, Z. (2011). Psychometric properties of the Chinese version of the Perceived Stress Scale in policewomen. PloS One, 6(12), e28610.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 54(3), 427-450.

Whittaker, J. (2009). Graphical models in applied multivariate statistics: Wiley Publishing.

WHO. (2015a). Australia: WHO statistical profile.

WHO. (2015b). United States of America: WHO statistical profile.

Wongpakaran, N., & Wongpakaran, T. (2010). The Thai version of the PSS-10: an investigation of its psychometric properties. Biopsychosocial Medicine, 4(1), 6.

World Bank. (2017). World Bank country and lending groups. In: World Bank Data Help Desk Washington (DC).

Wu, S. M., & Amtmann, D. (2013). Psychometric evaluation of the perceived stress scale in multiple sclerosis. ISRN Rehabilitation, 2013.

Yokokura, A. V. C. P., Silva, A. A. M. d., Fernandes, J. d. K. B., Del-Ben, C. M., Figueiredo, F. P. d., Barbieri, M. A., & Bettiol, H. (2017). Perceived Stress Scale: confirmatory factor analysis of the PSS14 and PSS10 versions in two samples of pregnant women from the BRISA cohort. Cadernos de Saúde Publica, 33, e00184615.

Zimet, G. D., Dahlem, N. W., Zimet, S. G., & Farley, G. K. (1988). The multidimensional scale of perceived social support. Journal of Personality Assessment, 52(1), 30-41.

Zimet, G. D., Powell, S. S., Farley, G. K., Werkman, S., & Berkoff, K. A. (1990). Psychometric characteristics of the multidimensional scale of perceived social support. Journal of Personality Assessment, 55(3-4), 610-617.

Table 1. Characteristic of the study participants.

	n	%
Age		
Mean	50.3	
SD	14.8	
Min/Max	18-82	
Missing	0	0%
Sex		
Female	2388	61.9%
Male	1469	38.1%
Missing	0	0%
Education		
High school or less	1252	32.5%
TAFE or university	2605	67.5%
Missing	0	0%
Employed		
Yes	2274	59%
No	1583	41%
Missing	0	0

Note. Mean values, minimum, maximum and standard deviations; numbers and percentages.

TAFE, Technical and Further Education (trade school/college).

Table 2. Conditional likelihood ratio test of overall model fit and Global DIF

	Model	Homogeneity	DIF by sex	DIF by education
PSS-14	RM	$\chi^2(55)=3828.3$, $p<0.001$	$\chi^2(55)=575.1$, $p<0.001$	$\chi^2(55)=320.9$, $p<0.001$
Perceived Stress	RM	$\chi^2(27)=1409.7$, $p<0.001$	$\chi^2(27)=177.8$, $p<0.001$	$\chi^2(27)=82.2$, $p<0.001$
	GLLRM	$\chi^2(96)=94.4$, $p=0.440$	$\chi^2(80)=111.8$, $p=0.012$	$\chi^2(88)=104.1$, $p=0.080$
Perceived Control	RM	$\chi^2(27)=713.4$, $p<0.001$	$\chi^2(27)=197.2$, $p<0.001$	$\chi^2(27)=104.1$, $p<0.001$
	GLLRM	$\chi^2(55)=62.5$, $p=0.224$	$\chi^2(39)=39.0$, $p=0.469$	$\chi^2(47)=70.9$, $p=0.014$

Note. The subgroups were defined according to lower and higher scores (i.e. homogeneity) to evaluate overall model fit; and according to sex (men; women) and education (Up to high school; TAFE or University) to evaluate Global DIF.

Table 3. Item fit statistics for the GLLRM of the negatively worded items (“Perceived Stress”)

	Conditional Outfit			Conditional Infit		
Item	Observed	SE	*p*-value	Observed	SE	*p*-value
Item 1	1.021	0.029	0.482	1.024	0.028	0.386
Item 2	0.950	0.031	0.108	0.948	0.026	0.049
Item 3	0.993	0.027	0.783	0.991	0.025	0.726
Item 11	1.015	0.026	0.550	1.024	0.025	0.355
Item 14	0.991	0.024	0.702	0.994	0.024	0.806

Note. The Conditional Outfit and Conditional Infit statistics have expected values equal to one under the Rasch model.

Table 4. Targeting and reliability information of the Perceived Stress and Perceived Control subscales.

Subgroup			Score			Target Index	Reliability	Probability of Person Separation
Education	Sex	n	Mean	SD	Target			
			Perceived Stress subscale					
Up to High School	Male	392	7.51	3.99	14.83	0.56	0.85	0.83
TAFE or Uni	Male	1075	7.41	3.70	14.85	0.58	0.83	0.82
Up to High School	Female	858	8.53	4.02	14.79	0.60	0.86	0.84
TAFE or Uni	Female	1525	8.48	3.65	14.79	0.62	0.82	0.82
			Perceived Control Subscale					
Up to High School	Male	392	4.29	2.45	9.18	0.36	0.77	0.75
TAFE or Uni	Male	1070	3.72	2.18	9.07	0.34	0.73	0.74
Up to High School	Female	857	4.14	2.20	9.28	0.34	0.75	0.75
TAFE or Uni	Female	1526	3.91	2.12	9.20	0.34	0.71	0.73

Note. The mean score is the average score for each subgroup. The target is the score which maximizes the information function. Reliability is the proportion of true score variance in relation to the total score variance. The probability of person separation is the probability that the scores of two random persons have the same rank order as their true person parameters.

Table 5. Observed and adjusted scores accounting for DIF

	Observed		Adjusted		Bias
	Mean	SE	Mean	SE	
Perceived Stress					
Education					
Up to High School	8.21	0.11	7.94	0.12	0.26
TAFE or University	8.04	0.07	7.89	0.07	0.15
Sex					
Female	8.50	0.08	8.16	0.08	0.34
Male	7.43	0.10	7.49	0.10	-0.06
Perceived Control					
Education					
Up to High School	7.81	0.06	7.92	0.07	-0.11
TAFE or University	8.17	0.04	8.42	0.04	-0.25
Sex					
Female	8.01	0.04	8.27	0.06	-0.26
Male	8.13	0.06	8.23	0.06	-0.11

Note. It is displayed the average score for each subgroup before and after adjustment for DIF. The bias indicates the differences between observed and adjusted scores.

Supplementary Materials

Table S1

The PSS-14 items divided into Perceived Stress and Perceived Control subscales

Item number	Item content
	Perceived Stress (PS) subscale
1	… felt upset because of something that happened unexpectedly?
2	… felt unable to control the important things in your life?
3	… felt either nervous or stressed?
8	… felt unable to cope with all the things that you had to do?
11	… felt angered because of things that happened outside of your control?
12	… found yourself thinking about all the things that you have to accomplish?
14	… felt difficulties were piling up so high that you could not overcome them?
	Perceived Control (PC) subscale
4	… dealt successfully with irritating life hassles?
5	… effectively coped with important changes in your life?
6	… felt confident about your ability to handle your personal problems?
7	… felt things were going your way?
9	… felt able to control irritations in your life?
10	… felt you were on top of things?
13	… felt able to control the way you spend your time?

Note. Every item started with the sentence "How often during the last year have you…"

Table S2

Item fit statistics for the PSS-14

	Conditional Outfit			Conditional Infit			Item-Restscore correlation		
Item	Observed	SE	*p*-value	Observed	SE	*p*-value	Observed	Expected	*p*-value
Item 4	1.556	0.024	<0.001	1.449	0.024	<0.001	0.266	0.471	<0.001
Item 5	1.550	0.024	<0.001	1.404	0.024	<0.001	0.319	0.477	<0.001
Item 6	0.808	0.024	<0.001	0.810	0.025	<0.001	0.657	0.468	<0.001
Item 7	0.818	0.023	<0.001	0.805	0.023	<0.001	0.624	0.466	<0.001
Item 9	1.493	0.023	<0.001	1.329	0.023	<0.001	0.399	0.483	<0.001
Item 10	0.661	0.024	<0.001	0.660	0.024	<0.001	0.747	0.471	<0.001
Item 13	1.181	0.023	<0.001	1.134	0.023	<0.001	0.444	0.481	<0.001
Item 1	0.959	0.024	0.090	0.958	0.024	0.080	0.525	0.478	<0.001
Item 2	0.731	0.023	<0.001	0.727	0.023	<0.001	0.657	0.489	<0.001
Item 3	0.820	0.023	<0.001	0.822	0.023	<0.001	0.616	0.488	<0.001
Item 8	0.887	0.023	<0.001	0.876	0.023	<0.001	0.592	0.477	<0.001
Item 11	0.951	0.023	0.036	0.949	0.024	0.029	0.521	0.475	<0.001
Item 12	1.452	0.022	<0.001	1.437	0.022	<0.001	0.257	0.481	<0.001
Item 14	0.726	0.023	<0.001	0.728	0.024	<0.001	0.657	0.478	<0.001

Note. The Conditional Outfit and Conditional Infit statistics have expected values equal to one under the Rasch model. The Item-Restscore correlation compares the observed item-restscore correlation with the expected item-restscore correlation under the model.

Table S3

Item fit statistics for the Perceived Stress subscale

	Conditional Outfit			Conditional Infit			Item-Restscore correlation		
Item	Observed	SE	*p*-value	Observed	SE	*p*-value	Observed	Expected	*p*-value
Item 1	0.942	0.024	0.016	0.946	0.023	0.021	0.653	0.618	0.001
Item 2	0.750	0.023	<0.001	0.751	0.023	<0.001	0.739	0.622	<0.001
Item 3	0.801	0.024	<0.001	0.805	0.023	<0.001	0.717	0.624	<0.001
Item 8	1.007	0.023	0.748	1.003	0.023	0.904	0.654	0.615	<0.001
Item 11	1.010	0.023	0.659	1.020	0.023	0.398	0.617	0.615	0.821
Item 12	1.675	0.023	<0.001	1.669	0.023	<0.001	0.367	0.617	<0.001
Item 14	0.804	0.023	<0.001	0.813	0.023	<0.001	0.714	0.615	<0.001

Note. The Conditional Outfit and Conditional Infit statistics have expected values equal to one under the Rasch model. The Item-Restscore correlation compares the observed item-restscore correlation with the expected item-restscore correlation under the model.

Table S4

Local dependence of the revised PSS-14 items

Item pair	Partial γ
Item 1– Item 2	0.18
Item 1 – Item 11	-0.23
Item 2 – Item 3	0.05
Item 3 – Item 14	0.03
Item 6 – Item 10	0.05
Item 7 – Item 10	0.22

Note. Average partial gamma given the items' restscores.

Table S5

Kelderman's likelihood ratio tests for the GLLRM of Perceived Stress subscale

Items	Conditional Likelihood Ratio test			Obs γ
	Differential Item Functioning			
Item 2 & Education:	lr = 12.23	df = 4	p = 0.016	0.11
Item 3 & Education:	lr = 3.35	df = 4	p = 0.501	
Item 11 & Education:	lr = 3.98	df = 4	p = 0.408	
Item 14 & Education:	lr = 3.97	df = 4	p = 0.410	
Item 2 & Sex:	lr = 3.93	df = 4	p = 0.416	
Item 11 & Sex:	lr = 6.66	df = 4	p = 0.155	
Item 14 & Sex:	lr = 6.80	df = 4	p = 0.147	
	Local Dependence			
Item 1 & Item 3:	lr = 81.60	df = 16	p <0.001	0.11 0.08
Item 1 & Item 14:	lr = 52.69	df = 16	p <0.001	-0.31 -0.26
Item 2 & Item 11:	lr = 96.56	df = 16	p <0.001	-0.23 -0.07
Item 11 & Item 14:	lr = 91.18	df = 16	p <0.001	0.03 0.18
Item 3 & Item 11:	lr = 92.19	df = 16	p <0.001	-0.14 -0.03
Item 11 & Item 14:	lr = 56.11	df = 16	p <0.001	0.05 0.02

Note. After the Benjamini-Hochberg procedure, statistical significance was adjusted as $p < 0.02692$ for a 5% FDR.

Table S6

Item fit statistics for the Perceived Control subscale

	Conditional Outfit			Conditional Infit			Item-Restscore correlation		
Item	Observed	SE	*p*-value	Observed	SE	*p*-value	Observed	Expected	*p*-value
Item 4	1.201	0.024	<0.001	1.134	0.024	<0.001	0.482	0.505	0.086
Item 5	1.152	0.024	<0.001	1.083	0.024	<0.001	0.521	0.509	0.372
Item 6	0.680	0.025	<0.001	0.704	0.025	<0.001	0.704	0.500	<0.001
Item 7	0.930	0.023	0.003	0.907	0.024	<0.001	0.585	0.501	<0.001
Item 9	1.367	0.023	<0.001	1.237	0.023	<0.001	0.457	0.513	<0.001
Item 10	0.763	0.024	<0.001	0.779	0.024	<0.001	0.661	0.505	<0.001
Item 13	1.222	0.023	<0.001	1.187	0.023	<0.001	0.442	0.511	<0.001

Note. The Conditional Outfit and Conditional Infit statistics have expected values equal to one under the Rasch model. The Item-Restscore correlation compares the observed item-restscore correlation with the expected item-restscore correlation under the model.

Table S7

Item fit statistics for the GLLRM of the Perceived Control subscale

	Conditional Outfit			Conditional Infit			Item-Restscore correlation		
Item	Observed	SE	*p*-value	Observed	SE	*p*-value	Observed	Expected	*p*-value
Item 6	0.979	0.033	0.532	0.994	0.033	0.867	0.719	0.711	0.470
Item 7	0.981	0.033	0.572	0.998	0.030	0.946	0.744	0.739	0.633
Item 10	1.022	0.042	0.594	1.023	0.038	0.538	0.781	0.783	0.877

Note. The Conditional Outfit and Conditional Infit statistics have expected values equal to one under the Rasch model. The Item-Restscore correlation compares the observed item-restscore correlation with the expected item-restscore correlation under the model.

Table S8

Kelderman's likelihood ratio tests for the GLLRM of the Perceived Control subscale

Items	Conditional Likelihood Ratio test			Obs γ
	Differential Item Functioning			
Item 6 & Education:	lr = 12.59	df = 4	p = 0.014	
Item 7 & Education:	lr = 3.36	df = 4	p = 0.449	
Item 7 & Sex:	lr = 0.64	df = 4	p = 0.982	
	Local Dependence			
Item 6 & Item 7:	lr = 128.04	df = 16	p<0.001	-0.16 -0.28

Note. After the Benjamini-Hochberg procedure, statistical significance was adjusted as $p < 0.005$ for a 5% FDR.

Table S9

Conversion table for score adjustment

Score	Adjusted Scores			
Perceived Stress Subscale	Sex: Male Education: Up to high school	Sex: Male Education: TAFE or University	Sex: Female Education: Up to high school	Sex: Female Education: TAFE or University
1	1.00	0.99	0.72	0.71
2	2.00	2.00	1.5	1.50
3	3.00	3.03	2.38	2.41
4	4.00	4.07	3.36	3.45
5	5.00	5.11	4.43	4.56
6	6.00	6.14	5.52	5.67
7	7.00	7.14	6.6	6.74
8	8.00	8.11	7.68	7.78
9	9.00	9.07	8.73	8.79
10	10.00	10.04	9.72	9.74
11	11.00	11.02	10.66	10.67
12	12.00	12.01	11.61	11.62
13	13.00	13.03	12.59	12.6
14	14.00	14.05	13.59	13.64
15	15.00	15.08	14.63	14.71
16	16.00	16.11	15.70	15.8
17	17.00	17.11	16.79	16.9
18	18.00	18.08	17.89	17.97
19	19.00	19.04	18.96	19
Perceived Control subscale	Sex: Male Education: Up to high school	Sex: Male Education: TAFE or University	Sex: Female Education: Up to high school	Sex: Female Education: TAFE or University
1	1.00	1.49	0.67	0.99
2	2.00	2.81	1.39	2.09
3	3.00	3.81	2.30	3.16
4	4.00	4.62	3.45	4.21
5	5.00	5.39	4.80	5.24
6	6.00	6.22	6.11	6.33
7	7.00	7.13	7.31	7.41
8	8.00	8.08	8.30	8.36
9	9.00	9.07	9.26	9.32
10	10.00	10.12	10.30	10.42

11	11.00	11.10	11.20	11.28

Note. The table indicates conversion values for comparison among subgroups. For example, if a woman with education up to high school had on the Perceived Stress subscale an observed total scores of 10, the score should be adjusted to the (true) value of 9.72.

Table S10

Convergent and divergent validity of the PSS-14

	FR	95% CI	FA	95% CI	SO	95% CI	EF	95% CI	RW	95% CI
Perceived Stress	-0.15	[-0.20, -0.10]	-0.22	[-0.27, -0.17]	-0.16	[-0.21, -0.11]	0.19	[0.12, 0.25]	-0.26	[-0.29, -0.16]
Perceived Control	0.23	[0.18, 0.28]	0.25	[0.30, 0.21]	0.28	[0.20, 0.29]	-0.09	[-0.16, -0.03]	0.28	[0.22, 0.34]

Note. MS = Mastery; PC = Perceived Constraints; SSS = Social Support Scale; FR = Friends; FA = Family; SO = Significant others; EF = Effort; RW = Rewards.

Chapter 5 – Networks of support: psychometric properties of the Social Support Scale (SSS) in two Aboriginal samples

Statement of Authorship

Title of Paper	Networks of support: psychometric properties of the Social Support Scale (SSS) in two Aboriginal samples
Publication Status	☐ Published ☐ Accepted for Publication ☒ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Submitted to Social Psychiatry and Psychaitric Epidemiology.

Principal Author

Name of Principal Author (Candidate)	Pedro Henrique Ribeiro Santiago		
Contribution to the Paper	Conceptualized the idea, performed the analysis, wrote the primary draft and acted as corresponding author.		
Overall percentage (%)	80%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	21/11/2019

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);
ii. permission is granted for the candidate in include the publication in the thesis; and
iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Lisa Smithers		
Contribution to the Paper	Conceptualized the idea, supervised development of work, provided intellectual contribution, and critically reviewed the draft manuscript.		
Signature		Date	21/11/2019

Name of Co-Author	Rachel Roberts
Contribution to the Paper	Conceptualized the idea, supervised development of work, provided intellectual contribution, and critically reviewed the draft manuscript.
Signature	Date 21/11/2019

Name of Co-Author	Lisa Jamieson
Contribution to the Paper	Conceptualized the idea, supervised development of work, provided intellectual contribution, and critically reviewed the draft manuscript.
Signature	Date 21/11/2019

Linkage to the body of work

Social support has a prominent role in the health of Aboriginal Australians since Aboriginal cultures foster a strong sense of communal responsibility. Historically, Aboriginal society was a complex kinship system, in which extended family members supported each other when confronted with hardship and illness. However, the process of colonization and the subsequent decades of assimilation policies has somewhat disassembled Aboriginal societies and mitigated the social support derived from their communities (Waterworth et al., 2014). One Western-developed psychological instrument that potentially could be used in Aboriginal research is the Social Support Scale (SSS). For purposes of this research, the SSS was examined by a 15-member Aboriginal Reference Group, comprising Aboriginal community members and Aboriginal Infant Care workers, who indicated the instrument had content and face validity to measure social support in Aboriginal people. In this study, we evaluated the psychometric properties of the Social Support Scale (SSS) in two Aboriginal samples.

Highlights

- The results indicate that the 4-item SSS is a measure that is valid and reliable to measure social support in Aboriginal Australians, and the replication of the findings across samples provides confidence in the robustness of the results.
- The SSS is unidimensional and a total score summing all four items can be used as a measure of social support.
- No DIF was found regarding education and sex, consequently, scores of men/women and respondents with high school/university education can be compared without bias due to gender or education.

Research and Policy Implications

- The present study showed that the original 4-item SSS is a construct valid and reliable instrument to measure social support in Aboriginal Australians.

Networks of support: psychometric properties of the Social Support Scale (SSS) in two Aboriginal samples

Pedro Henrique Ribeiro Santiago[a]*, Lisa Gaye Smithers[b], Rachel Roberts[c] and Lisa Jamieson[a]

[a]Adelaide Dental School, The University of Adelaide, Adelaide, Australia; [b]School of Public Health, The University of Adelaide, Adelaide, Australia; [c]School of Psychology, The University of Adelaide, Adelaide, Australia;

*Pedro Henrique Ribeiro Santiago is the corresponding author. Present address: AHMS Building, The University of Adelaide, North Terrace, Adelaide SA 5000. Telephone: +61 0478199813. ORCID: 0000-0002-1267-8340. E-mail address: pedro.ribeirosantiago@adelaide.edu.au.

Authorship: All persons who meet authorship criteria are listed as authors. Pedro Henrique Ribeiro Santiago conducted the formal analysis. Lisa Jamieson conceptualized the study. Pedro Henrique Ribeiro Santiago, Lisa Gaye Smithers, Rachel Roberts, and Lisa Jamieson designed the methodology. Lisa Gaye Smithers, Rachel Roberts, and Lisa Jamieson supervised the work. Pedro Henrique Ribeiro Santiago wrote the original draft and Lisa Gaye Smithers, Rachel Roberts and Lisa Jamieson reviewed and edited the manuscript.

Abstract

Objective: In Australia, despite social support increasingly being reported as playing an important role in influencing health outcomes of Aboriginal and Torres Strait Islanders, measures of social support have not yet been validated for Aboriginal people. The aim of the current study was to evaluate the validity and reliability of the Social Support Scale in an Aboriginal population. The Social Support Scale (SSS) is a 4-item measure developed to evaluate the emotional, appraisal, instrumental and informational domains of social support. Methods: Data was collected from two different samples: participants of the (1) Teeth Talk Study (n=317), an oral-health randomized controlled trial (RCT) conducted among Aboriginal adults; and (2) the Baby Teeth Talk Study (n=367), an RCT conducted among pregnant Aboriginal women. The psychometric properties of the SSS were evaluated with Graphical Loglinear Rasch Models (GLLRM). Results: Overall fit to a GLLRM was found ($\chi^2(96)_{sample1}$= 52.7, p=0.06; $\chi^2(25)_{sample2}$= 22.2, p=0.62) after the inclusion of local dependence between items 3 and 4. Item 2 displayed differential item functioning by employment in Sample 1. Unidimensionality was confirmed in both samples (γ_{obs1}=0.80; γ_{exp1}=0.78, p=0.65; γ_{obs2}=0.75, γ_{exp2}=0.77, p=0.16) and the SSS displayed good reliability ($R_{sample1}$=0.82, $R_{sample2}$=0.84). The same unidimensional structure was found in both samples, providing confidence in the robustness of results. Conclusions: The current study consisted of the first validation of a Western-developed psychological instrument to measure social support in Aboriginal Australians. The results show that the SSS is a culturally-valid and reliable instrument that can be applied in future health research among Aboriginal and Torres Strait Islanders.

Keywords: Social support; Aboriginal Australians; Rasch analysis; Differential Item Functioning; Item-response theory.

Introduction

In contemporary Australia, despite improvements due to governmental, non-governmental and private organisations efforts over the past decades, the health inequalities between Aboriginal and Torres Strait Islanders and non-Aboriginal Australians persist. For example, Aboriginal people have approximately a 10 years shorter life expectancy, are 3.3 times more likely to have diabetes, 1.2 times more likely to have respiratory and cardiovascular disease and almost 2 times more likely to commit suicide than non-Aboriginal Australians. A large proportion of this gap can be explained by social determinants of health (e.g. education, employment, discrimination) (31%), while behavioural and biomedical factors (e.g. substance use, physical activity, blood pressure) account for a smaller fraction (11%) (AIHW, 2015). Among the social determinants of health, social support has been recently receiving attention in Indigenous research (King et al., 2009).

Social support is defined as four classes of supportive behaviour: *emotional support*, the provision of trust, empathy, and love by others (e.g. listening to difficulties experienced by a colleague in their job); *appraisal support*, the extent which individuals receive adequate evaluation (e.g. receiving feedback from a work supervisor if performance was good/poor); *instrumental support*, instrumental behaviour directed to address specific needs (e.g. borrowing money from a friend to pay a bill); and *informational support*, receiving information that can help with solving problems (e.g. informing a person of a job opportunity) (House, 1981). This definition maintained prominence within the literature since other theoretical conceptualizations can be mapped under these four broad categories of supportive behaviour (Malecki and Demaray, 2003).

The effects of social support on health have been extensively reported in non-Indigenous populations. High social support is associated with better outcomes in coronary heart disease (Barth et al., 2010), breast cancer (Nausheen et al., 2009), type 2 diabetes (Van

Dam et al., 2005), improved sleep outcomes (Kent de Grey et al., 2018), pain management (Che et al., 2018), among many others. Over the past 30 years, two theories have explained how social support influences health: *the main effect model*, in which social support promotes positive psychological states (e.g. security, stability, belonging) and greater motivation for self-care; and the *stress-buffering model*, which suggests that the provision of resources by others diminish the perceived harm of a situation and reduces the negative biological reactions of stress (Rueger et al., 2016). In recent years, studies have also investigated the effects of social support on Indigenous health. For example, in Canada social support was consistently associated with better self-reported general health in First Nations, Métis and Inuit (Richmond et al., 2007), while *Yuuyaraq,* social support from extended family and peers, reduced drinking alcohol as a coping response to trauma in Alaska Natives (Mohatt et al., 2004).

Social support has a prominent role in the health of Aboriginal Australians since Aboriginal cultures foster a strong sense of communal responsibility. Historically, Aboriginal society was a complex kinship system, in which extended family members supported each other when confronted with hardship and illness. These connections provided a defined social structure and reinforced their cultural identity. However, the process of colonization and the subsequent decades of assimilation governmental policies, which culminated in the removal of Aboriginal children from their parents (i.e. the "Stolen Generations") from the 1900s to the 1970s, disassembled the Aboriginal societies and mitigated the social support derived from their communities (Waterworth et al., 2014).

The research of social support in Aboriginal Australians, however, has been faced with conceptual challenges: the Aboriginal experiences of social and emotional well-being (SEWB) are connected to the well-being of the land and the community; and, for this reason, are not directly comparable to more individualistic Western conceptualizations of "*mental* health" (Garvey, 2008). For this reason, to avoid cultural bias, the recommendation is that

Western-developed psychological instruments should always be validated in an Aboriginal population prior to usage in health research (Santiago et al., 2019).

One Western-developed psychological instrument that potentially could be used in Aboriginal research is the Social Support Scale (SSS) (Supplementary Table 1). Following the recommendations for cultural adaptation of psychological instruments, the SSS was examined by a 15-member Aboriginal Reference Group, comprising Aboriginal community members and Aboriginal Infant Care workers, which indicated the instrument had content and face validity for Aboriginal Australians. For this reason, the SSS was chosen to be applied and its psychometric properties evaluated in an Aboriginal population.

Psychometric properties of the Social Support Scale

The SSS was originally developed by Peeters et al. (1995) to evaluate *emotional* (e.g. "There are people in my life who pay attention to my feelings and problems"), *appraisal* (e.g. "There are people in my life who appreciate what I do"), *instrumental* (e.g. "There are people in my life who I can get help from if I need it ") and *informational support* (e.g. "There are people in my life who I can talk to about how to handle things") according to House's (1981) framework. Additionally, Peeters et al. (1995) included two items (e.g. "we had a casual chat" and "we made jokes and had fun") to evaluate *rewarding companionship*. The two items included in the SSS to evaluate reward companionship were not incorporated in the current study since several authors have argued about the importance of "making a clear distinction between supportive interactions and rewarding companionship" (Buunk and Verhoeven, 1991).

In the original validation of the SSS, the questionnaire was applied to a sample of 41 female secretaries in the Netherlands. A Principal Component Analysis (PCA) was conducted and three components emerged. The first contained the *instrumental* and *informational support* items and was interpreted as "Instrumental Support", while the second contained the *emotional* and *appraisal support* items and was interpreted as "Intimate Support. The two overarching dimensions of Instrumental and Emotional support have been consistently identified in empirical research (Morelli et al., 2015; Semmer et al., 2008). The reason is that informational support (i.e. providing information) is usually given as a way to help someone solve a problem or complete a task and, therefore, is closely related to instrumental support (i.e. directly helping). In parallel, providing appraisal support (i.e. constructive feedback or praise) can be perceived as a form of emotional support (Shakespeare-Finch and Obst, 2011). The third component found by Peeters et al. (1995) contained the two *reward companionship* items. The study also evaluated the reliability ($\alpha_{Instrumental}$=0.80; $\alpha_{Intimate}$=0.77) and the correlation between the Instrumental and the Intimate Support subscales (r=0.56; $p < 0.001$) (Peeters et al., 1995). Finally, to investigate criterion validity, the *stress-buffering model* was chosen and the results showed that Instrumental Support buffered the effects of daily stressful events on negative affect ($\beta_{Inst \times Stress}$ = -0.16; $p < 0.01$).

Despite reporting adequate psychometric properties for a limited Western population of female secretaries, the restricted sample of the original study, combined with the cultural differences of Aboriginal Australians, limits the generalizability of the results (number of dimensions, magnitude of the loadings) to an Aboriginal population. After the initial validation conducted by Peeters et al. (1995), the SSS was applied in epidemiological studies, including studies with Indigenous populations (Lawrence et al., 2016), but no further validation has been conducted.

The present research

Research on social support can provide insight into how social determinants contribute to the existing health gap between Aboriginal and non-Aboriginal Australians. To the best of our knowledge, there are no available validated instruments to measure social support in any Indigenous population. Therefore, the aim of the current study was to evaluate whether the SSS is a valid and reliable measure of social support in Aboriginal Australians. To answer this question, we used Rasch models and extended Rasch models (Kreiner and Christensen, 2007) to provide a robust analysis of the SSS psychometric properties. The analysis included: a) model fi; b) item fit; c) local dependence; d) differential item functioning; and e) unidimensionality. After a measurement model was established, f) reliability and g) targeting of the SSS was assessed. Finally, the h) criterion validity was evaluated by inspecting the convergent/divergent validity of the SSS with stress (i.e. Perceived Stress Scale (SSS)) and sense of personal control (i.e. Sense of Personal Control Scale (SPCS)).

Material and methods

Participants and procedures

Sample 1: The sample was composed of 317 Aboriginal Australians in the Teeth Talk Study (TT), a randomized controlled trial (RCT) aimed at improving oral health literacy among Aboriginal adults in Port Augusta. The study was promoted via posters in community centres and advertisements on a local Aboriginal radio station. Participants were recruited through a variety of methods, including self-nomination, word of mouth, home visits and referrals. The TT study received ethical approval from the Aboriginal Health Council of South Australia, the Human Research Ethics Committee of the University of Adelaide, the Board of Management of the Pika Wiya Health Service (PWHS) and the local community-controlled Indigenous health service (Parker et al., 2012).

Sample 2: The sample was composed of 367 pregnant Aboriginal women in the Baby Teeth Talk Study (BTT). The Baby Teeth Talk Study is a randomized controlled trial aimed at reducing early childhood caries of Aboriginal children in South Australia. The BTT study received approval from the University of Adelaide Human Research Ethics Committee, the Aboriginal Health Council of South Australia, the Government of South Australia and the

Human Research Ethics Committees of three participating South Australian hospitals. The psychological instruments were administered as part of a broader questionnaire by four research staff (three Indigenous and one non-Indigenous) at baseline (Jamieson et al., 2014).

Written informed consent was obtained from all individual participants included in both studies. Additionally, all procedures performed in the BTT and TT studies were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Complementary Measures

The Perceived Stress Scale (PSS): The PSS is the most widely used psychological instrument to measure perceived stress, which evaluates if a person's life is perceived as unpredictable, uncontrollable, or overloading. The PSS is composed of 14 items in its original form (PSS-14) and the subscales of Perceived Stress (PS) and Perceived Coping (PC). A revised version with 13 items was validated in an Aboriginal population (Santiago et al., 2019).

The Sense of Personal Control Scale (SPCS): The SPCS is a 12 items five-point scale (1=Not at all, 2=Rarely, 3=Sometimes, 4=Fairly often, 5=Very often), composed of the two factors of Mastery (MS) and Perceived Constraints (PCN) (Lachman and Weaver, 1998). A revised version with 7 items was validated for Aboriginal Australians (Santiago et al., 2017).

The Rasch measurement models

The Rasch model (RM) is part of the family of Item Response Theory (IRT) models and has the two distinctive features that: (1) the sum score is a sufficient statistic for the person parameter; and (2) measurement is objective, since the comparison of two individuals is independent of the particular items or psychological instrument being used. See Tennant and Conaghan (2007) for a non-technical introduction, and Kreiner (2007) for a technical discussion.

The RM implies that items are conditionally independent of other items (i.e. local independence) and to exogenous variables (i.e. absence of differential item functioning (DIF)) given the latent trait. However, these mathematical requirements of the model are strict and it is common for items of rating scales applied in health sciences to exhibit local dependence (LD) and/or DIF. For this reason, these items do not fit the RM and might be deleted during the validation process to obtain statistical fit to the model. The deletion of items can be problematic since many of the items were originally developed to evaluate certain domains of the construct and the exclusion can lead to "construct underrepresentation". Furthermore, an instrument with a small number of items might exhibit decreased reliability (Nielsen and Kreiner, 2013). For this reason, models called Graphical Loglinear Rasch Model (GLLRM) which extend the RM with log-linear parameters were developed to incorporate items with uniform LD and uniform DIF (i.e. for simplicity, the term uniform is omitted when referred to uniform LD or uniform DIF from now on). Items with LD convey less information than independent items and items with DIF require adjustment of scores between subgroups. However, in both cases, the items serve the purpose of measuring the latent trait and retaining them contributes to construct validity.

Statistical analysis

Item analysis: Initially, it was tested the fit of the items to the RM. In case there was no fit, it was investigated whether the departures in terms of positive LD and/or DIF could be accounted by a GLLRM. The estimation method was conditional maximum likelihood (Andersen, 1970) and person parameters were estimated with weighted maximum likelihood (WML) (Warm, 1989). All statistical analyses were conducted with the DIGRAM v4.05 (Kreiner and Nielsen, 2013). Descriptive statistics and graphs were created with R software (R

Core Team, 2013). Since missing values for individual items ranged from 0.0% to 0.005% considering both studies, multiple imputation was not required (Graham, 2009).

Model fit and item fit: Overall fit of the model was evaluated through the Conditional Likelihood Ratio (CLR) test (Andersen, 1973). The CLR test evaluates *measurement invariance* within subgroups, providing a test of overall model fit (i.e. subgroups defined by higher and lower scores) and overall DIF (e.g. subgroups defined by exogenous variables, such as males and females). In the current study, the characteristics analysed for DIF were sex (Male; Female), education (education level up to High School; TAFE or University) and employment status (Unemployed; Employed). Technical and Further Education (or TAFE) is the biggest provider of post-secondary education in Australia. TAFE offers a broad range of courses, at the operative, trade and paraprofessional level, that can last from a few hours (refreshment courses) to three years (diploma courses). Unlike universities, which are composed mostly of full-time students, TAFE institutions allow students to combine study and work, and encourage programs of apprenticeships and traineeships. Item fit was evaluated by conditional infit and outfit statistics (Christensen and Kreiner, 2012).

LD and DIF: LD was evaluated through the matrix of residual correlations. Since residuals correlations of the Rasch model are known to be *negatively biased*, the mean-adjusted residual correlation matrix was used (Christensen et al., 2017). In addition, Kelderman's (1984) likelihood ratio (LR) test was conducted to evaluate LD and/or DIF, and the *magnitude* of the LD/DIF was informed through partial Goodman-Kruskal γ rank correlations. Due to the multiple testing, the Benjamini-Hochberg (1995) procedure was performed to adjust for false discovery rate (FDR).

Dimensionality: The SSS was divided into two subscales composed of Items 1 and 2 ("Emotional Support"), and Items 3 and 4 ("Instrumental Support"), and a formal test of dimensionality was conducted by comparing the observed γ correlation of the subscales with the expected γ correlation of the subscales under a unidimensional model. The rationale of this analysis was that the correlation between two subscales measuring distinct traits (i.e.

"Emotional Support" and "Instrumental Support") should be weaker than the expected correlation of subscales measuring the same trait (i.e. "Social Support") (Horton et al., 2012).

Reliability and targeting: Since Cronbach's α (1951) provides a lower-bound estimate of reliability *when items are locally independent*, a Monte Carlo simulation method (Hamon and Mesbah, 2002) that accounts for LD between items was applied. The targeting was evaluated through the Test Target Information Index, which is the mean test information divided by the maximum obtained test information (Kreiner and Christensen, 2012).

Criterion validity: Considering that scores are ordinal, the non-parametric Kendall's τ correlation of the SSS with the PSS and SPCS was evaluated. It was expected a positive correlation of the SSS with PC and MA, and negative correlation with the PS and PCN subscales.

Results

The sample characteristics of both studies are found in Table 1.

(Table 1)

Sample 1: The CLR indicated that item parameters were not invariant between participants with low scores and high scores (χ^2=69.6; df=15; p<0.001) and, therefore, there was no overall fit to the RM. When item parameters were compared across subgroups, invariance was not achieved between employed and unemployed (χ^2=31.2; df=15; p=0.01) indicating DIF by employment status (Table 2).

(Table 2)

The conditional outfits and infits disclosed misfit of Item 1 (Outfit = 1.350, SE=0.103, $p<0.001$; Infit=1.32, SE=0.10, p=0.001) (Supplementary Table 2).

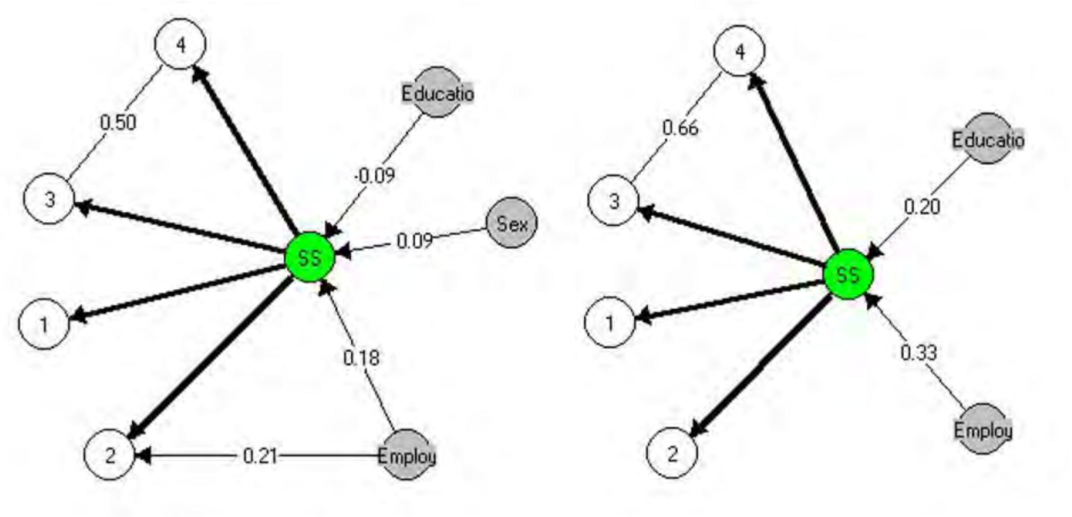
Before considering removing Item 1, it was investigated whether the departures from the RM would constitute LD and/or DIF. The analysis of the residual correlations (Table 3) suggested that, after the influence of the trait ("social support") was accounted, there was LD between Item 3 ("There are people in my life who I can get help from if I need it") and Item 4 ("There are people in my life who I can talk to about how to handle things") (Adj=0.22; $p<0.001$), and between Item 1 ("There are people in my life who pay attention to my feelings and problems") and Item 2 ("There are people in my life who appreciate what I do") (Adj=0.07; p=0.17).

(Table 3)

The Kelderman's LR test supported the evidence of LD between items 3 and 4 (LR = 57.26, df=16, $p<0.001$) (γ_{avg}=0.50) and items 1 and 2 (LR = 68.46, df=16, $p<0.001$; γ_{avg}=0.02). However, in the last case, the magnitude of the dependence (γ_{avg}=0.02) was unsubstantial. Furthermore, the results from Kelderman's LR tests disclosed that Item 2 had DIF by employment status (LR = 12.60, df=4, p=0.01; γ=0.21).

Fit to a GLLRM adjusting for these departures was found (χ^2 = 52.7, df = 96, p=0.06) (Table 2) (Figure 1).

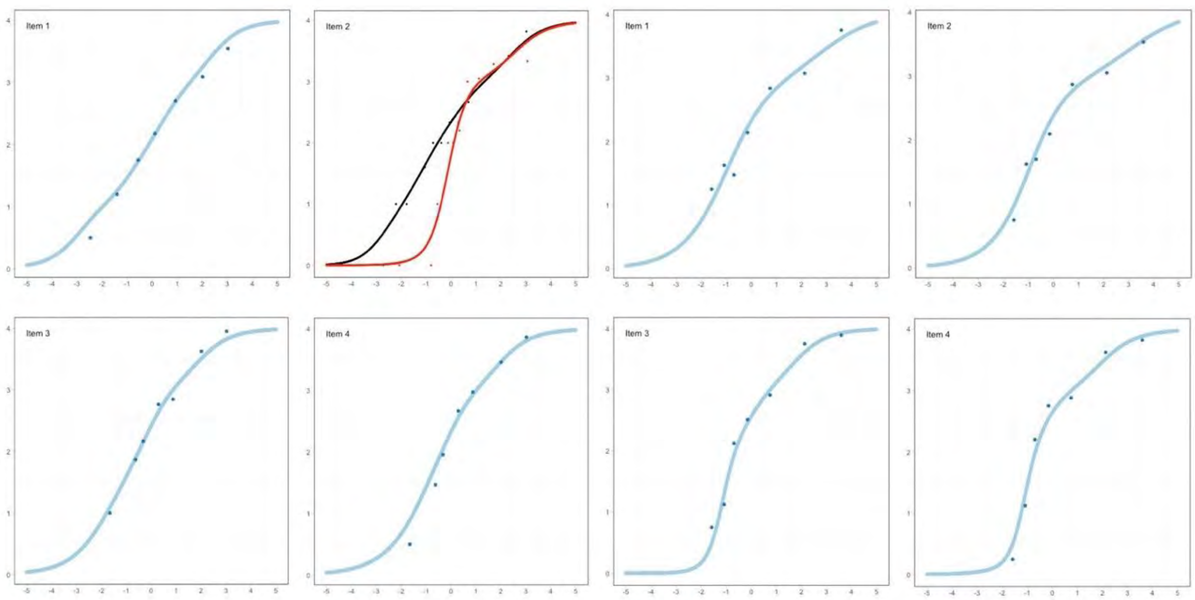
Figure 1. GLLRMs of the SSS for Sample 1 (left) and Sample 2 (right).



Note. The Markov graph nodes represent the item numbers, the exogenous variables, and the latent trait. Disconnected nodes indicate that variables are conditionally independent and partial γ informs the magnitude of the LD and DIF. When LD/DIF was unsubstantial (γ<0.1), the edges were omitted (see Supplementary Figure 1).

Considering that there was no further evidence of item misfit (Supplementary Table 2) (Figure 2), and Kelderman's LR indicated no substantial evidence of DIF or LD (Supplementary Table 3), the measurement model for the SSS in Sample 1 was established.

Figure 2. Item characteristic curves of the GLLRM for Sample 1 (left) and Sample 2 (right).



Note. The x-axis indicates the latent trait ("Social support") and the y-axis indicates the item score. The points represent the *average* observed item responses in each of the seven class intervals. Since Item 2 had DIF by employment, the Item Characteristic Curves (ICCs) for participants employed (black) and unemployed (red) are displayed.

Sample 2: The CLR indicated weak evidence against the RM (χ^2=25.1; df=15; p=0.05). Invariance of item parameters was found within subgroups, and there was no evidence of DIF by education (χ^2=19.3; df=15; p=0.20) or employment (χ^2=12.8; df=15; p=0.62) (Table 2). In addition, there was no item misfit (Supplementary Table 2). However, the examination of the residual correlations showed a similar pattern of Sample 1, suggesting LD between Item 3 and Item 4 (Adj=0.32; p<0.001) and between Item 1 and Item 2 (Adj=0.14; p<0.001). Kelderman's LR test confirmed these results by disclosing LD between items 3 and 4 (LR = 68.15, df=16, p<0.001; γ_{avg}=0.66) and items 1 and 2 (LR = 44.33, df=16, p<0.001; γ_{avg} =0.15). Finally, Kelderman's LR test showed no evidence of DIF.

The strong LD between items 3 and 4 was the reason why the CLR initially showed evidence against the RM since after this LD parameter was incorporated into a GLLRM, the model fitted (χ^2 = 22.2, df = 25, p=0.62) (Table 2) (Figure 2). Since there was no further evidence of item misfit (Supplementary Table 2) and LD or DIF (Supplementary Table 3), the measurement model for the SSS in Sample 2 was established.

Dimensionality: The correlation between subset 1 (Item 1 and Item 2) and subset 2 (Item 3 and Item 4) was in accordance with the expected correlation under an unidimensional model in Sample 1 (γ_{obs}=0.80, γ_{exp}=0.78, p=0.65) and Sample 2 (γ_{obs}=0.75, γ_{exp}=0.77, p=0.16). Therefore, the results indicated that the SSS is a unidimensional scale.

Reliability and targeting: The SSS displayed good overall reliability ($R_{sample1}$=0.82, $R_{sample2}$=0.84) and adequate overall probability of person separation ($P_{sample1}$=0.77, $P_{sample2}$=0.78) (Supplementary Table 4). Therefore, if we order two randomly chosen participants of Sample 2 based on their total score in 78% of the cases they will be correctly ordered with respect to their *true* level of social support.

Targeting was poor in both samples (Supplementary Table 4). For example, the Test Information Target Index showed that in Sample 1 the SSS provided only 28% of the total information available regarding social support in comparison to a perfectly targeted instrument. The reason is that the mean score in Sample 2 was 12.49 (SD=3.15) while the SSS was perfectly target for a group with a mean score of 7.47 (Supplementary Figure 2).

Criterion validity: The SSS score displayed the expected pattern of convergent and divergent validity regarding stress ($r_{sample2}$=-0.21, 95% C.I. [-0.35, -0.06]), perceived coping ($r_{sample2}$=0.28, 95% C.I. [0.13, 0.41]) and mastery ($r_{sample1}$=0.35, 95% C.I. [0.22, 0.46], $r_{sample2}$=0.34, 95% C.I. [0.20, 0.46]). The only exception was perceived constraints, which displayed the expected negative correlation in Sample 2 ($r_{sample2}$=-0.27, 95% C.I. [-0.40, -0.12]) but in Sample 1 no correlation was found ($r_{sample1}$=0.03, 95% C.I. [-0.12, 0.15]) (Supplementary Table 5).

Discussion

The aim of the present study was to evaluate the construct validity and reliability of the Social Support Scale in an Aboriginal population. The results indicate that the SSS is a measure that is valid and reliable to measure social support in Aboriginal Australians, and the replication of the findings across samples provides confidence in the robustness of the results. The implications for future research are: (1) the SSS is unidimensional, therefore the total score (i.e. summated score) can be used as a measure of social support; (2) no DIF was found regarding education and sex, consequently scores of men/women and respondents with high school/university education can be compared without been confounded by gender or education.

Dimensionality and local dependence: The results showed that the SSS is a unidimensional scale with strong local dependence between Item 3 (“There are people in my life who I can get help from if I need it”) which measures *instrumental support* and Item 4

("There are people in my life who I can talk to about how to handle things") which measures *informational support*. The dependence between the instrumental and informational items is in accordance with psychological theory since providing useful information is closely related to directly helping the completion of a task (Shakespeare-Finch and Obst, 2011).

The result of unidimensionality seems to contrast the original SSS validation in which a PCA indicated the two dimensions of Instrumental and Emotional support (Peeters et al., 1995). However, Peeters et al. (1995) original validation had a small (n=43) and restricted sample and PCA without a sufficient size can lead to the extraction of random factors (Osborne & Costello, 2004) (Osborne and Costello, 2004). Furthermore, PCA (and Exploratory Factor Analysis do not allow for locally dependent items (i.e. correlated uniqueness) and can lead to the extraction of factors that do not necessarily correspond to substantive psychological traits (i.e. artifactors).

In this study, the investigation of dimensionality *and* local dependence showed that, despite the dependence between the instrumental/informational items, a single dimension of social support was measured by the four items. The unidimensionality of the SSS was indicated by the correlation between subscales: not only the observed correlations were not distinct from the expected correlations under a unidimensional model, but the strong *ordinal* γ correlations (γ_{obs1}=0.796, γ_{obs2}=0.747) implied poor discriminant validity across subscales. That is, the correlation between instrumental/informational and appraisal/emotional subscales was too strong for them to be considered qualitatively different and constitute different dimensions.

The unidimensionality of the SSS is consistent with recent research on social support. Semmer et al. (2008) argued that the distinction between the two dimensions of Emotional and Instrumental support, previously reported in empirical research (Kirrane and Buckley, 2004), is tenuous since many supportive behaviours described as instrumental (e.g. helping to clean the house) have also an emotional meaning by communicating empathy, caring and respect. Furthermore, Semmer et al. (2008: 247) explained that: "Skillfully provided social support will, therefore, often be both instrumental in behavior and emotional in symbolic meaning. At least for support that is interpreted as helpful, a high correlation between instrumental and emotional support from the same source is, therefore, very likely. This might explain the observation that instrumental and emotional support are often strongly correlated in previous research".

DIF: Item 2 ("There are people in my life who appreciate what I do") displayed DIF by employment. This indicates that employed Aboriginal respondents systematically endorsed that there are people in their life who appreciate what they do in comparison to non-employed respondents given the same level of social support. High unemployment rates are one of the main social inequalities experienced by Aboriginal Australians; for example, 75% of participants in Sample 1 and 87% in Sample 2 were unemployed. Considering that the causes of unemployment among Aboriginal Australians are deeply rooted in historical, structural and cultural factors (Altman, 2018), it seems plausible that individuals who were employed systematically felt more appreciated by those in their life. It should be noted that DIF happened only for *appraisal support*, but the same pattern was not found for instrumental, informational and emotional support.

Moreover, there is one possible explanation why Item 2 DIF by employment was not replicated in Sample 2, which comprised pregnant Aboriginal women. In Australia, approximately 23% of the women leave their job during pregnancy and 92% take paid or unpaid leave (Australian Bureau of Statistics, 2017). Since it is common to be unemployed or absent from work during pregnancy, Aboriginal mothers who were employed during their gestation might not have felt systematically more appreciated by others in comparison to the non-employed mothers. Nonetheless, we recommend that future studies should try to replicate the results found in Sample 1 and investigate again if Item 2 displays DIF by employment before practical recommendations for the SSS application are made.

Reliability and targeting: Reliability and probability of person separation were good in both samples ($R_{sample1}$=0.82, $R_{sample2}$=0.84, $P_{sample1}$=0.77, $P_{sample2}$=0.78). Reliability values between .70 and .80 are usually deemed adequate for research purposes, while values between .80 and .85 indicate that the instrument can be used for individual testing in low-stakes scenarios (Wells and Wollack, 2003). Therefore, the SSS can be used to measure social support of Aboriginal respondents not only as an average group level but also on an individual level. For example, the SSS can inform us of the average social support of an Aboriginal population in health research, but it can also be used to distinguish social support between two individual Aboriginal respondents in a low-stake context (i.e. no critical decision will be made based on their scores).

Targeting was poor in both samples since participants had more social support than the levels the instrument was developed to measure. In Indigenous groups, high levels of social support do not implicate *only* health protection. The high-density social networks (e.g. extended families) provide support and reinforce belonging but also exert conformity pressure and over-obligations. Furthermore, the low income and poor material circumstances confine individuals within their immediate social context, making it harder to avoid harmful relationships (e.g. domestic violence) (Richmond and Ross, 2008). For example, in Western Australia, Aboriginals have reported the difficulties created by the lifestyle of their extended family members ("We've got family members who drink; I've got three brothers and a sister who are alcoholics"), while discrimination at work impeded new relationships ("Meeting people through work, I thought a lot of people would reject me, being black. […] Like what happened one time, they thought I was stealing petrol out of my own truck") (Waterworth et al., 2014). To improve targeting, future studies can further develop the SSS by creating *culturally-sensitive difficult* items to measure higher levels of social support. For example, an item could measure whether Aboriginal Australians experience social support on the workplace or from the wider Australian community (i.e. non-Indigenous Australians), and this would be a difficult item that could improve the targeting of the SSS.

Criterion validity: The analysis indicated a negative association with perceived stress (since Aboriginal Australians with higher social support perceived themselves to be less stressed) and a positive association of social support with perceived control. A positive association was also found between social support and *personal*-mastery (e.g. “I can do just anything I really set my mind to”). In Indigenous cultures, previous research has shown that *communal*-mastery (e.g. “I am successful by virtue of my social attachments”) has a higher impact than *personal*-mastery on health outcomes (Hobfoll et al., 2002). This reflects the collectivistic aspect of Indigenous societies when compared to more individualistic Western cultures. For this reason, future validation studies should include *communal*-mastery as an additional criterion for validity of the SSS in Aboriginal Australians (and other Indigenous populations). In general, the patterns of convergent and divergent validity were congruent with the theoretical expectations, providing further evidence of the construct validity of the SSS.

Strengths and limitations: The strengths of the present study comprised of the modern psychometric methodology applied and the use of two Aboriginal samples to ensure the replicability of the results. One fundamental limitation is that the main sample of this study which contained Aboriginal men and women was a convenience sample in a rural setting; while the second sample was recruited in metropolitan areas but interviewed only pregnant Aboriginal women. Future studies need to further investigate the construct validity of the SSS in a sample more representative of the Aboriginal population.

Conclusion

The relationship between social support and health of Aboriginal Australians is nuanced. The social support derived from extended families provides health benefits, but the high-density networks and lack of financial resources can create over-obligations and force harmful relationships. One main challenge to epidemiological research was the lack of psychological instruments validated to measure social support in this group. To the best of our knowledge, this is the first study to validate a measure of social support for an Indigenous population. The present study showed that SSS is a construct valid and reliable instrument to measure social support in Aboriginal Australians.

Word count: 4901.

Declaration of interest

The authors declare that they have no conflict of interest.

Funding

The Teeth Talk study was funded by the National Health and Medical Research Council of Australia (NHMRC, Project Grant 627101). The Baby Teeth Talk study was funded by an International Collaborative Indigenous Health Research Partnership grant from the National Health and Medical Research Council of Australia (NHMRC, Project Grant 627350).

References

AIHW. (2015) The Health and Welfare of Australia's Aboriginal and Torres Strait Islander peoples 2015. AIHW Canberra.

Altman JC. (2018) *Aboriginal employment equity by the year 2000*: Canberra, ACT: Centre for Aboriginal Economic Policy Research, Research ….

Andersen EB. (1970) Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*: 283-301.

Andersen EB. (1973) A goodness of fit test for the Rasch model. *Psychometrika* 38: 123-140.

Australian Bureau of Statistics. (2017) Pregnancy and Employment Transition.

Barth J, Schneider S and Von Känel R. (2010) Lack of social support in the etiology and the prognosis of coronary heart disease: a systematic review and meta-analysis. *Psychosomatic Medicine* 72: 229-238.

Benjamini Y and Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*: 289-300.

Buunk BP and Verhoeven K. (1991) Companionship and support at work: A microanalysis of the stress-reducing features of social interaction. *Basic and Applied Social Psychology* 12: 243-258.

Che X, Robin C, Sungwook C, et al. (2018) Investigating the influence of social support on experimental pain and related physiological arousal: A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews*.

Christensen KB and Kreiner S. (2012) Item fit statistics. *Rasch models in health*: 83-104.

Christensen KB, Makransky G and Horton M. (2017) Critical values for Yen's Q 3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement* 41: 178-194.

Cronbach LJ. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297-334.

Garvey D. (2008) Review of the social and emotional wellbeing of Indigenous Australian peoples. *Australian Indigenous Health InfoNet.*

Graham JW. (2009) Missing data analysis: Making it work in the real world. *Annual Review of Psychology* 60: 549-576.

Hamon A and Mesbah M. (2002) Questionnaire reliability under the Rasch model. *Statistical methods for quality of life studies.* Springer, 155-168.

Hobfoll SE, Jackson A, Hobfoll I, et al. (2002) The impact of communal-mastery versus self-mastery on emotional outcomes during stressful conditions: A prospective study of Native American women. *American Journal of Community Psychology* 30: 853-871.

Horton M, Marais I and Christensen KB. (2012) Dimensionality. *Rasch models in health*: 137-158.

House JS. (1981) *Work stress and social support,* Reading, PA: Addison Wesley.

Jamieson LM, Parker EJ, Roberts-Thomson KF, et al. (2014) Self-efficacy and self-rated oral health among pregnant aboriginal Australian women. *BMC Oral Health* 14: 29.

Kelderman H. (1984) Loglinear Rasch model tests. *Psychometrika* 49: 223-245.

Kent de Grey RG, Uchino BN, Trettevik R, et al. (2018) Social support and sleep: A meta-analysis. *Health Psychology* 37: 787.

King M, Smith A and Gracey M. (2009) Indigenous health part 2: the underlying causes of the health gap. *The Lancet* 374: 76-85.

Kirrane M and Buckley F. (2004) The influence of support relationships on work-family conflict: Differentiating emotional from instrumental support. *Equal Opportunities International* 23: 78-96.

Kreiner S. (2007) Validity and objectivity: Reflections on the role and nature of Rasch models. *Nordic Psychology* 59: 268-298.

Kreiner S and Christensen KB. (2007) Validity and objectivity in health-related scales: Analysis by graphical loglinear Rasch models. *Multivariate and mixture distribution Rasch models.* Springer, 329-346.

Kreiner S and Christensen KB. (2012) Person parameter estimation and measurement in Rasch models. *Rasch models in health*: 63-78.

Kreiner S and Nielsen T. (2013) Item Analysis in Digram-Notes on the use of DIGRAM for item analysis by graphical loglinear Rasch models. *Department of Biostatistics-University of Copenhagen*.

Lachman ME and Weaver SL. (1998) The sense of control as a moderator of social class differences in health and well-being. *Journal of Personality and Social Psychology* 74: 763.

Lawrence HP, Cidro J, Isaac-Mann S, et al. (2016) Racism and oral health outcomes among pregnant Canadian Aboriginal women. *Journal of Health Care for the Poor and Underserved* 27: 178-206.

Malecki CK and Demaray MK. (2003) What type of support do they need? Investigating student adjustment as related to emotional, informational, appraisal, and instrumental support. *School Psychology Quarterly* 18: 231.

Mohatt GV, Rasmus SM, Thomas L, et al. (2004) " Tied together like a woven hat:" Protective pathways to Alaska native sobriety. *Harm Reduct J* 1: 10.

Morelli SA, Lee IA, Arnn ME, et al. (2015) Emotional and instrumental support provision interact to predict well-being. *Emotion* 15: 484.

Nausheen B, Gidron Y, Peveler R, et al. (2009) Social support and cancer progression: a systematic review. *Journal of Psychosomatic Research* 67: 403-415.

Nielsen T and Kreiner S. (2013) Improving items that do not fit the Rasch model. *Rasch models in health*: 317-334.

Osborne JW and Costello AB. (2004) Sample size and subject to item ratio in principal components analysis. *Practical assessment, research & evaluation* 9: 8.

Parker EJ, Misan G, Chong A, et al. (2012) An oral health literacy intervention for Indigenous adults in a rural setting in Australia. *BMC Public Health* 12: 461.

Peeters MC, Buunk BP and Schaufeli WB. (1995) Social interactions, stressful events and negative affect at work: A micro-analytic approach. *European Journal of Social Psychology* 25: 391-401.

R Core Team. (2013) R: A language and environment for statistical computing.

Richmond CA and Ross NA. (2008) Social support, material circumstance and health behaviour: Influences on health in First Nation and Inuit communities of Canada. *Social Science and Medicine* 67: 1423-1433.

Richmond CA, Ross NA and Egeland GM. (2007) Social support and thriving health: a new approach to understanding the health of indigenous Canadians. *American Journal of Public Health* 97: 1827-1833.

Rueger SY, Malecki CK, Pyun Y, et al. (2016) A meta-analytic review of the association between perceived social support and depression in childhood and adolescence. *Psychological Bulletin* 142: 1017.

Santiago PHR, Roberts R, Smithers LG, et al. (2017) Psychometric properties of the Sense of Personal Control Scale (SPCS), Perceived Stress Scale (PSS) and Social Support Scale (SSS) in an Aboriginal Population. *2017 PHAA SA State Population Health Conference.* Adelaide, South Australia.

Santiago PHR, Roberts R, Smithers LG, et al. (2019) Stress beyond coping? A Rasch analysis of the Perceived Stress Scale (PSS-14) in an Aboriginal population. *PloS One* 14.

Semmer NK, Elfering A, Jacobshagen N, et al. (2008) The emotional meaning of instrumental social support. *International Journal of Stress Management* 15: 235.

Shakespeare-Finch J and Obst PL. (2011) The development of the 2-way social support scale: A measure of giving and receiving emotional and instrumental support. *Journal of Personality Assessment* 93: 483-490.

Tennant A and Conaghan PG. (2007) The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research* 57: 1358-1362.

Van Dam HA, van der Horst FG, Knoops L, et al. (2005) Social support in diabetes: a systematic review of controlled intervention studies. *Patient Education and Counseling* 59: 1-12.

Warm TA. (1989) Weighted likelihood estimation of ability in item response theory. *Psychometrika* 54: 427-450.

Waterworth P, Rosenberg M, Braham R, et al. (2014) The effect of social support on the health of Indigenous Australians in a metropolitan community. *Social Science and Medicine* 119: 139-146.

Wells CS and Wollack JA. (2003) An instructor's guide to understanding test reliability. *Testing & Evaluation Services. University of Wisconsin*.

Table 1. Characteristics of the study participants.

	Sample 1		Sample 2	
	n	%	n	%
Age				
Mean	36.4		24.9	
SD	14.0		5.9	
Min/Max	18/82		14/43	
Missing	0	0%	0	0%
Sex				
Female	214	76%	367	100%
Male	103	24%	0	0%
Missing	0	0%	0	0%
Education				
High school or less	236	74%	266	73%
TAFE or university	81	26%	98	27%
Missing	0	0%	0	0%
Employment				
Yes	80	25%	45	12.4%
No	237	75%	316	87.3%
Missing	0	0%	1	0.3%

Note. Mean values, range, and standard deviations; numbers and percentages.
TAFE, Technical and Further Education (trade school/college).

Table 2. Conditional likelihood ratio test of overall model fit and global DIF

Model	Sample 1				Sample 2		
	Homogeneity	DIF by sex	DIF by education	DIF by employment status	Homogeneity	DIF by education	DIF by employment status
RM	$\chi^2(15)=69.6$, $p<0.001$	$\chi^2(15)=15.8$, p=0.39	$\chi^2(15)=13.6$, p=0.55	$\chi^2(15)=31.2$, p=0.01	$\chi^2(15)=25.1$, p=0.05	$\chi^2(15)=19.3$, p=0.20	$\chi^2(15)=12.8$, p=0.62
GLLRM	$\chi^2(38)=52.7$, p=0.06	$\chi^2(38)=50.4$, p=0.09	$\chi^2(38)=39.5$, p=0.40	$\chi^2(30)=43.9$, p=0.05	$\chi^2(25)=22.2$, p=0.62	$\chi^2(25)=10.3$, p=0.99	$\chi^2(25)=16.4$, p=0.90

Note. The subgroups were defined according to lower and higher scores (i.e. homogeneity) to evaluate overall model fit; and according to sex (Men; Women), education (Up to high school; TAFE or University) and employment (Unemployed; Employed) to evaluate overall DIF.

Table 3. Matrix of residual correlations of the Social Support Scale.

		Sample 1			
		Item 1	Item 2	Item 3	Item 4
Item 1		1			
Item 2	Obs	**-0.254**	1		
	Adj	**0.073**			
Item 3	Obs	-0.482	-0.309	1	
	Adj	-0.155	0.017		
Item 4	Obs	-0.523	-0.286	**-0.105**	1
	Adj	-0.196	0.041	**0.221**	
		Sample 2			
		Item 1	Item 2	Item 3	Item 4
Item 1		1			
Item 2	Obs	**-0.185**	1		
	Adj	**0.140**			
Item 3	Obs	-0.413	-0.484	1	
	Adj	-0.088	-0.159		
Item 4	Obs	-0.453	-0.407	**-0.009**	1
	Adj	-0.127	-0.082	**0.316**	

Note. The matrix includes: a) the observed correlation between standardized residuals after the influence of the latent trait ("Social Support") was accounted by the Rasch model; and b) the mean-adjusted residual correlations, which are the difference between the observed residual correlations and the average residual correlation.

Supplementary Materials

Supplementary Table 1. The SSS items.

Item number	Item content
1	There are people in my life who pay attention to my feelings and problems
2	There are people in my life who appreciate what I do
3	There are people in my life who I can get help from if I need it
4	There are people in my life who I can talk to about how to handle things

Supplementary Table 2. Item fit statistics for the RM and GLLRM of the Social Support Scale (SSS).

	Conditional Outfit			Conditional Infit		
	Observed	SE	p-value	Observed	SE	p-value
	Sample 1					
RM						
Item 1	1.350	0.103	<0.001	1.319	0.100	0.001
Item 2	0.794	0.110	0.060	0.815	0.106	0.080
Item 3	0.912	0.104	0.400	0.920	0.103	0.438
Item 4	0.954	0.107	0.665	0.978	0.104	0.823
GLLRM						
Item 1	1.254	0.133	0.056	1.234	0.124	0.058
Item 2	0.730	0.124	0.029	0.773	0.130	0.079
Item 3	0.965	0.120	0.714	0.952	0.121	0.691
Item 4	0.988	0.121	0.921	1.046	0.113	0.585
	Sample 2					
RM						
Item 1	1.234	0.095	0.014	1.150	0.090	0.096
Item 2	1.213	0.108	0.049	1.118	0.096	0.217
Item 3	0.804	0.102	0.054	0.832	0.100	0.092
Item 4	0.785	0.103	0.037	0.808	0.105	0.068
GLLRM						
Item 1	0.826	0.112	0.121	0.879	0.097	0.212
Item 2	0.890	0.122	0.367	0.991	0.100	0.931
Item 3	1.239	0.149	0.110	1.170	0.120	0.156
Item 4	1.080	0.140	0.566	1.011	0.119	0.976

Note. The Conditional Outfit and Conditional Infit statistics have expected values equal to one under the Rasch model.

Supplementary Table 3. Kelderman's likelihood ratio tests for the GLLRM of the Social Support Scale.

	Conditional Likelihood Ratio test			Obs γ
Sample 1[a]	Differential Item Functioning			
Item 1 & Sex:	lr = 4.65	df = 4	p = 0.325	
Item 2 & Sex:	lr = 5.67	df = 4	p = 0.225	
Item 3 & Sex:	lr = 2.59	df = 4	p = 0.629	
Item 4 & Sex:	lr = 4.01	df = 4	p = 0.404	
Item 1 & Education:	lr = 6.86	df = 4	p = 0.143	
Item 2 & Education:	lr = 3.00	df = 4	p = 0.557	
Item 3 & Education:	lr = 2.29	df = 4	p = 0.582	
Item 4 & Education:	lr = 3.20	df = 4	p = 0.524	
Item 1 & Income:	lr = 7.81	df = 4	p = 0.989	
Item 3 & Income:	lr = 11.61	df = 4	p = 0.020	
Item 4 & Income:	lr = 4.93	df = 4	p = 0.294	
	Local Dependence			
Item 1 & Item 3:	lr = 48.88	df = 16	p < 0.001	0.26 -0.45
Item 1 & Item 4:	lr = 57.75	df = 16	p < 0.001	-0.36 -0.54
Item 2 & Item 3:	lr = 9.50	df = 16	p = 0.892	
Item 2 & Item 4:	lr = 39.97	df = 16	p = 0.001	-0.03 0.15
Sample 2	Differential Item Functioning			
Item 1 & Education:	lr = 2.44	df = 4	p = 0.655	
Item 2 & Education:	lr = 12.71	df = 4	p = 0.013	
Item 3 & Education:	lr = 6.94	df = 4	p = 0.139	
Item 4 & Education:	lr = 2.39	df = 4	p = 0.665	
Item 1 & Income:	lr = 1.62	df = 4	p = 0.806	
Item 2 & Income:	lr = 6.11	df = 4	p = 0.191	
Item 3 & Income:	lr = 2.04	df = 4	p = 0.728	
Item 4 & Income:	lr = 3.29	df = 4	p = 0.511	
	Local Dependence			
Item 1 & Item 2:	lr = 45.34	df = 16	p < 0.001	0.04 0.27
Item 1 & Item 3:	lr = 52.69	df = 16	p = 0.104	
Item 1 & Item 4:	lr = 96.56	df = 16	p = 0.027	
Item 2 & Item 3:	lr = 91.18	df = 16	p = 0.029	
Item 2 & Item 4:	lr = 92.19	df = 16	p = 0.562	

Note. After the Benjamini-Hochberg procedure, statistical significance was adjusted as $p < 0.010$ for Sample 1 and $p < 0.004$ for Sample 2 for a 5% FDR.

Supplementary Table 4. Targeting and reliability information of the SSS.

Sample	Score				Test Target Index	Cronbach's α	Reliability	Probability of Person Separation
	n	Mean	SD	Target				
Unemployed (Sample 1)	237	12.38	3.18	7.47	0.57	0.89	0.83	0.79
Employed (Sample 1)	80	13.16	2.89	7.64	0.34	0.86	0.80	0.75
Sample 2	365	12.49	3.15	5.43	0.28	0.88	0.84	0.78

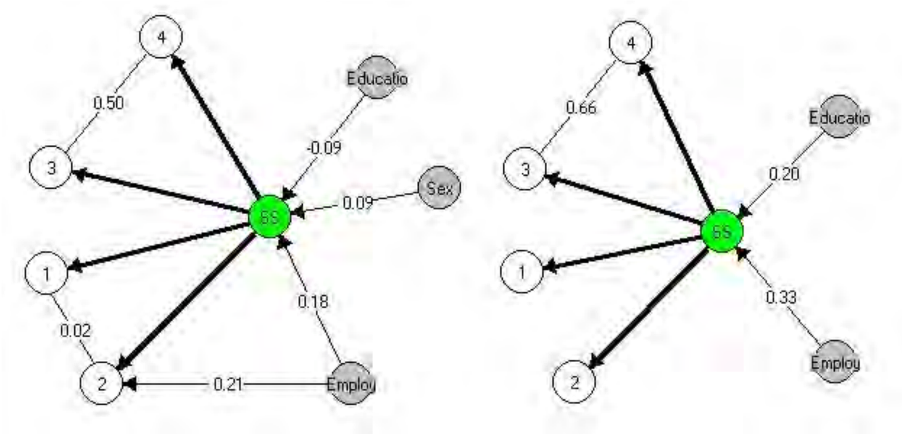
Note. The mean score is the average score for each subgroup. The target is the score which maximizes the information function. Reliability is the proportion of true score variance in relation to the total score variance. The probability of person separation is the probability that the scores of two random persons have the same rank order as their true person parameters.

Supplementary Table 5. Convergent and divergent validity of the SSS.

	MA	95% C.I	PCN	95% C.I	PS	95% C.I	PC	95% C.I
Sample 1	0.35	[0.22, 0.46]	0.03	[-0.12, 0.15]	-	-	-	-
Sample 2	0.34	[0.20, 0.46]	-0.27	[-0.40, -0.12]	-0.21	[-0.35, -0.06]	0.28	[0.13, 0.41]

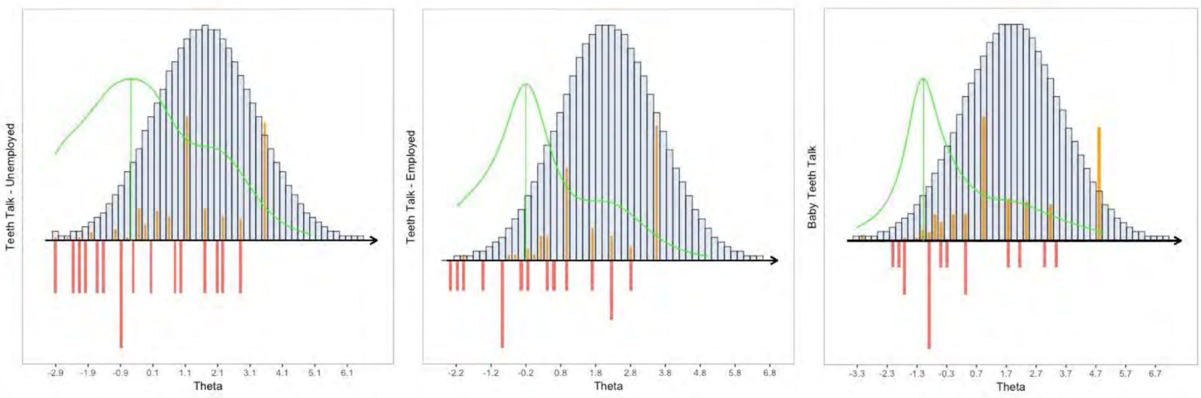
Note. The table displays the score correlations between SSS and complementary measures. MS = Mastery; PCN = Perceived Constraints; PS = Perceived Stress; PLC = Perceived Lack of Control.

Supplementary Figure 1. GLLRMs of the SSS for Sample 1 (left) and Sample 2 (right).



Note. The Markov graph nodes represent the item numbers, the exogenous variables, and the latent trait. Disconnected nodes indicate that variables are conditionally independent and partial γ informs the magnitude of the LD and DIF.

Supplementary Figure 2. Item Map of the SSS in Sample 1 (left and center) and Sample 2 (right).



Note. The orange bars display the person parameters (WML estimates). The grey bars display the theoretical population distribution of Perceived Control under the assumption of normality. The red bars display the item thresholds and the green line is the information function.

Chapter 6 – Sense of Personal Control: can it be assessed fairly across Aboriginal and non-Aboriginal Australians?

Statement of Authorship

Title of Paper	Sense of Personal Control: can it be assessed fairly across Aboriginal and non-Aboriginal Australians?
Publication Status	☐ Published ☐ Accepted for Publication ☒ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Submitted to PlOS One.

Principal Author

Name of Principal Author (Candidate)	Pedro Henrique Ribeiro Santiago		
Contribution to the Paper	Conceptualized the idea, performed the analysis, wrote the primary draft and acted as corresponding author.		
Overall percentage (%)	70%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	21/11/2019

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);
ii. permission is granted for the candidate in include the publication in the thesis; and
iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Tine Nielsen		
Contribution to the Paper	Conceptualized the idea, supervised development of work, provided intellectual contribution, and critically reviewed the draft manuscript.		
Signature		Date	21/11/2019

Name of Co-Author	Rachel Roberts		
Contribution to the Paper	Conceptualized the idea, supervised development of work, provided intellectual contribution, and critically reviewed the draft manuscript.		
Signature		Date	21/11/2019

Name of Co-Author	Lisa Smithers
Contribution to the Paper	Conceptualized the idea, supervised development of work, provided intellectual contribution, and critically reviewed the draft manuscript.
Signature	Date 21/11/2019

Name of Co-Author	Lisa Jamieson
Contribution to the Paper	Conceptualized the idea, supervised development of work, provided intellectual contribution, and critically reviewed the draft manuscript.
Signature	Date 21/11/2019

Linkage to the body of work

In Australia, the legacy of colonization and subsequent decades of assimilation policies had a direct impact on the sense of personal control of Aboriginal people. Aboriginal Australians were marginalized from participation in major social and political decisions and their society was disassembled during the 20th century. The undermining of self-determination in social matters, both in the country and in their communities, led individuals to lose the sense of control over their lives (Daniel, Brown, Dhurrkay, Cargo, & O'Dea, 2006). The recommendations regarding validating psychological instruments specifically for Aboriginal Australians seem particularly important for measures of personal control since personal control is influenced by culture (Cheng, Cheung, Chio, & Chan, 2013). For example, the association of personal control with anxiety symptoms is weaker in collectivist societies compared to individualistic (Western) societies (Cheng et al., 2013); moreover, individuals from collectivist cultures (e.g. China) are more likely to exert control through cultivating relationships, while individuals from individualistic cultures tend to exert control through personal effort (e.g. United States) (Spector, Sanchez, Siu, Salgado, & Ma, 2004). Previous studies applied sense of personal control measures without validation for Aboriginal people and one stated that “construct validity has been confirmed” (but in the original country) (Daniel et al., 2006). In this study, we evaluated the psychometric properties of the Sense of Personal Control Scale (SPCS) for Aboriginal Australians and whether this instrument could be used to provide cross-cultural measurement with non-Aboriginal Australians. Similarly to what happened with the previous instruments, the SPCS was initially examined by a 15-member Aboriginal Reference Group, who indicated that the instrument had content and face validity for an Aboriginal population.

Highlights

- The study provided initial evidence of validity of a revised 5-item Perceived Constraints subscale. However, the overlap in content among the items led to poor reliability. Future studies should consider including new culturally-specific items to improve reliability.
- The revised Mastery subscale had only 2 items and new culturally-specific items need to be developed for this subscale.
- The SPCS did not provide culturally unbiased measurement across Aboriginal and non-Aboriginal samples.

Research and Policy Implications

- The revised 7-item Sense of Personal Control Scale is an instrument that requires adaptations, such as the inclusion of new culturally-specific items, to improve its psychometric properties before it can be applied to measure personal control in Aboriginal Australians.

Sense of Personal Control: can it be assessed fairly across Aboriginal and non-Aboriginal Australians?

Pedro Henrique Ribeiro Santiago(1)*‡, Tine Nielsen(2)‡, Rachel Roberts(3), Lisa Smithers(4), Lisa Jamieson(5)

(1)Australian Research Centre for Population Oral Health (ARCPOH), Adelaide Dental School, The University of Adelaide, Australia
(2) Department of Psychology, The University of Copenhagen, Denmark
(3)School of Psychology, The University of Adelaide, Australia
(4)School of Public Health, The University of Adelaide, Australia

* Corresponding author
‡ These authors are joint first authors on this work and have contributed equally to the work.

Email: pedro.ribeirosantiago@adelaide.edu.au

Author contributions
Conceptualization: TN, PHRS, LJ
Data Curation: LJ
Formal analysis: TN, PHRS
Funding Acquisition: TN, LJ
Investigation: TN, PHRS, RR, LS, LJ
Methodology: TN, PHRS, RR, LS, LJ
Supervision: TN, RR, LS, LJ
Validation: TN, RR, LS, LJ
Writing – original draft: TN, PHRS
Writing – review and editing: TN, PHRS, RR, LS, LJ

Abstract

In recent decades, several studies have emphasized sense of personal control as a prominent aspect of Aboriginal health. However, one limitation is that instruments available to measure personal control were originally developed in Western countries and validation for Aboriginal Australians has not been conducted. The aims of the current study were to evaluate whether the Sense of Personal Control Scale (SPCS) can be used to obtain fair and culturally unbiased measurement of personal control across Aboriginal and non-Aboriginal Australians and to assess the psychometric properties of the SPCS in samples of Aboriginal and non-Aboriginal Australian . Methods: The current study utilized two Australian subsamples retrieved from the Teeth Talk Study (n=317) and the National Survey of Adult Oral Health 2004-2006 (n=3,857) in which the SPCS was included. The psychometric properties were evaluated with Graphical Loglinear Rasch Models (GLLRM). Results: The Perceived Constraints subscale achieved fit to a GLLRM for Aboriginal Australians after the exclusion of three items, while fit to any RM/GLLRM model could not found in the non-Aboriginal sample. The Mastery subscale achieved fit to a GLLRM in the non-Aboriginal sample after the exclusion of one item. In the Aboriginal sample, two items of the Mastery subscale achieved fit to a RM, however, two items cannot be considered as a scale. Conclusion: In the present study, we showed that the development of new culturally-specific items is crucial before the revised SPCS might constitute a valid and reliable measure of sense of personal control in both Aboriginal and non-Aboriginal Australian populations, and it is possible to assess whether the SPCS can measure fairly across these two populations.

Introduction

A topic of on-going research in Australia is how social determinants of health contribute to the large health inequalities between Aboriginal and non-Aboriginal Australians (Carson, Dunbar, Chenhall, & Bailie, 2007). Among different social determinants, recent studies have emphasized sense of personal control as a fundamental aspect of the contemporary Aboriginal experience (Reilly et al., 2008). Sense of personal control is the generalized expectation that outcomes are contingent on individual behaviors (Rotter, 1966). Therefore, it has operationalized into two dimensions: *mastery*, beliefs that individual behaviors will produce the desired outcomes (Pearlin, Menaghan, Lieberman, & Mullan, 1981); and *perceived constraints*, beliefs that outcomes are beyond individual influence, being determined by external factors (Seligman & Maier, 1967).

The effects of personal control on health have been extensively studied in non-Indigenous populations. Meta-analysis associated sense of personal control with general well-being (mental well-being, life satisfaction, and physical health), higher job satisfaction (Wang, Bowling, & Eschleman, 2010), lower burnout (Lee, Lim, Yang, & Lee, 2011) and lower depression (Presson & Benassi, 1996). Individual studies have also associated personal control with anxiety (Clark & Watson, 1991) and longevity (Krause & Shaw, 2000).

Sense of personal control of Aboriginal Australians

In Australia, the legacy of colonization and subsequent decades of assimilation policies had a direct impact on the sense of personal control of Aboriginal people. Aboriginal Australians were marginalized from participation in major social and political decisions and their society was disassembled during the 20th century. The undermining of self-determination in social matters, both in the country and in their communities, led individuals to lose the sense of control over their lives (Daniel, Brown, Dhurrkay, Cargo, & O'Dea, 2006).

A few recent studies have investigated the effects of personal control on Aboriginal health. Daniel et al. (2006) examined personal control in a remote Aboriginal community with poor living conditions (e.g. limited access to transportation, communication, food storage) and found it to be negatively associated with stress. Furthermore, considering that racism against Aboriginals creates unfair and unpredictable demands that can undermine personal control (Ziersch, Gallaher, Baum, & Bentley, 2011), a recent study by Paradies and Cunningham (2012) showed that personal control mediated the effects of racism on depression. Finally, Reilly et al. (2008, p. 351) suggested personal control as a potential protective factor of cardiovascular disease and recommended "further empirical investigation".

Unbiased measurement of personal control of Aboriginal and non-Aboriginal Australians

While research on the effects of personal control on Aboriginal health is on-going, there are two main gaps that this study aims to address. Firstly, recent recommendations by Santiago, Roberts, Smithers, and Jamieson (2019) emphasized the importance of validating psychological instruments specifically for Aboriginal Australians. These recommendations seem particularly important for measures of personal control since personal control is influenced by culture (Cheng, Cheung, Chio, & Chan, 2013). For example, the association of personal control with anxiety symptoms is weaker in collectivist societies compared to individualistic (Western) societies (Cheng et al., 2013); moreover, individuals from collectivist cultures (e.g. China) are more likely to exert control through cultivating relationships, while individuals from individualistic cultures tend to exert control through personal effort (e.g. US) (Spector, Sanchez, Siu, Salgado, & Ma, 2004). Since Aboriginal Australians comprise several collectivist cultures and the general (non-Aboriginal) Australian population form a Western individualistic society (in most part due to its European descendants), these cultural differences raise questions whether Western-developed measures of personal control are appropriate for Aboriginal Australians. Previous studies applied sense of personal control measures without validation for Aboriginal people and one stated that "construct validity has been confirmed" (but in the original country) (Daniel et al., 2006). It is necessary to ensure that psychological instruments have also construct validity *specifically* for Aboriginal and Torres Strait Islanders[3] in Australia, otherwise the item responses can be subject to construct bias.

Secondly, it is important to investigate whether these instruments provide *unbiased* measurement of personal control compared with non-Aboriginal Australians. The validation of an instrument for Aboriginal Australians can only inform the level of the construct measured (e.g. level of personal control) *within* the Aboriginal community. However, it is the development of unbiased instruments, considering that some forms of bias can be adjusted statistically and others need to be addressed at the instrument level, which can inform the real impact of social inequalities on Aboriginal Australians' emotional well-being by *comparing it* to a non-Aboriginal group. That is, an unbiased instrument can inform how much personal control Aboriginal Australians experience by contrasting their personal control with non-Aboriginal Australian individuals.

The present research

The aims of the current study were to (1) investigate whether the SPCS can be used to obtain fair and culturally unbiased measurement of personal control across Aboriginal and non-Aboriginal Australians and (2) evaluate the psychometric properties of the Sense of Personal Control Scale (SPCS) for Aboriginal Australians and non-Aboriginal Australian sample. To achieve these aims, we employed state-of-the-art item response theory methods in

[3] The term "Torres and Strait Islanders" refer to the Indigenous people of Torres Strait Islands in Queenland, Australia. Since they are ethnically distinct from the continental Indigenous people of Australia, referred as Aboriginals, the term "Torres Strait Islanders" is usually combined with the term "Aboriginal Australians" to emphasize the two distinct groups (Cox & Taua, 2012).

the form of Rasch models and graphical log-linear Rasch models to evaluate the construct validity, the issue of bias and fairness, and the reliability of the SPCS across samples of Aboriginal and non-aboriginal Australians.

Methods

Measures

Sense of Personal Control Scale (SPCS): The SPCS is a 12 item scale intended to measure sense of personal control (Lachman & Weaver, 1998). It consists of two subscales; Mastery (MS) and Perceived Constraints (PC), and a five-point response scale is used (1=Not at all, 2=Rarely, 3=Sometimes, 4=Fairly often, 5=Very often) to rate items.

The SPCS was developed based on a widely used instrument to measure personal control; the seven-item Pearlin (1981) Mastery Scale, which was expanded with five items to make up the SPCS (Supplementary Table 1). In the development study of the SPCS, Exploratory Factor Analysis (EFA) found a two-dimensional structure, interpreted as Mastery (MS) and Perceived Constraints (PC). Internal consistency of the MS subscale was α=.70 and for the PC subscale α=.86 (Lachman & Weaver, 1998).

The validity of Pearlin's (1981) original 7-item Mastery Scale has been investigated within several cultures, for example in countries such as Sweden, Iran, China, Japan, among others (Chen, Hsiung, Chung, Chen, & Pan, 2013; Eklund, Erlandsson, & Hagell, 2012; Shateri et al., 2018; Togari & Yonekura, 2015). Despite the investigation of the Mastery Scale psychometric properties in several countries, no previous study conducted cross-cultural validation. Pearlin's scale has also been examined with modern psychometric methods, including Rasch models (Eklund et al., 2012), and shown to be associated with biological markers such as interleukin-6[4] (Lundgren et al., 2018). Furthermore, Pearlin's scale has been translated into Indigenous languages such as the Yolngu Matha, an Aboriginal language spoken in northeast Australia (Daniel et al., 2006). In contrast, the psychometric properties of the extended 12-item SPCS have not been evaluated since the original study (Lachman & Weaver, 1998).

Finally, prior to application in Aboriginal participants, we followed recommendations for the cultural adaptation of psychological instruments and consulted an Aboriginal Reference Group composed of 15 members, comprising Aboriginal community members and Aboriginal Infant Care workers. The group examined the SPCS and indicated that the instrument had content and face validity for the Aboriginal culture. For this reason, the SPCS was chosen as a potentially valid instrument for Aboriginal Australians.

Exogenous variables

Education was measured through the categories "Up to High school" and "Tertiary education". Considering that the items measuring education had different number of categories in each sample (i.e. Aboriginal and non-Aboriginal Australians), we dichotomized the variable into two categories to enable the comparison across samples. I ncome was measured through the categories "Employed" and "Unemployed". Once again, we dichotomized this variable to enable comparison across samples.

Samples

The current study utilized two Australian subsamples retrieved from other studies, where the SPCS had been included in the collected data. The first sample was composed of

[4] The inflammatory cytokine interleukin (IL)-6 is an established risk marker of several diseases, including coronary heart disease (CHD) (Lundgren, Garvin, Andersson, Jonasson, & Kristenson, 2018).

317 Aboriginal Australians that participated in the Teeth Talk study (Parker et al., 2012). . The Teeth Talk study was a randomized controlled trial (RCT) aimed at improving oral health literacy among Aboriginal adults in South Australia. The study was promoted via posters in community centers and advertisements on a local radio station. The participants were recruited through various methods, including home visits, referrals, self-nomination and word of mouth.

The second sample was composed of 3,857 non-Aboriginal Australians in the population-based cross-sectional study Australia's National Survey of Adult Oral Health (NSAOH) 2004-2006 (Sanders & Slade, 2011). The study used a questionnaire that was mailed to participants that undertook dental examination.

Both studies received ethical approval and all participants provided signed informed consent. The demographic characteristics of each sample are included in Supplementary Table 2.

Rasch measurement models

The class of Rasch models belongs to the larger family of item response theory (IRT) models. The simplest is the original Rasch Model (RM) for dichotomous items (Rasch, 1960). In the current study, we used the Partial Credit Model (PCM) (Masters, 1982), which generalize the RM to ordinal items, and Graphical Log-Linear Rasch models (GLLRM) (Kreiner, 2007; Kreiner & Christensen, 2002; Kreiner & Christensen, 2004). As both the dichotomous RM and the ordinal PCM adhere to the same requirements for measurement ((Kreiner, 2012; Mesbah & Kreiner, 2012), we use the term "RM" for Rasch model in the remainder of the paper. The five basic requirements for measurement are: 1) *unidimensionality*, the items of a scale measure a single underlying latent construct; 2) *monotonicity*, the expected item scores increase with increasing values on the latent variable; 3) *local independence of items* (or no LD), the item responses are conditionally independent given the latent variable; 4) *absence of differential item functioning* (no DIF), item responses and exogenous variables (i.e. relevant background variables) are conditionally independent given the latent variable, and 5) *homogeneity*; the rank order of item parameters (item "difficulties") is the same for all persons no matter their level on the latent variable. The first four requirements provide criterion-related construct validity according to Rosenbaum (1989) and are common for all IRT models. The fifth requirement of homogeneity pertains only to the RM. Fulfillment of all five requirements provides ideal measurement, as the raw summed score is then a sufficient statistic for the person parameter. The sufficiency of the raw sum score distinguishes scales fitting Rasch models from scales fitting other IRT models (Kreiner, 2013). Sufficiency is desirable when summed raw scores are used, such as is the case with sense of control in population surveys (Slade, Spencer, & Roberts-Thomson, 2007). However, it is also possible to convert the sum scores to Rasch scores (i.e. the estimated person parameters), which are on a logit scale.

If fit to the RM is rejected, it is still possible to achieve close to optimal measurement, if the only departures from the model are in the form of uniform DIF and/or uniform LD (Kreiner & Christensen, 2002, 2004, 2007). Uniform implies that the LD or DIF is the same across all levels of the latent construct. Uniform LD and DIF can be adjusted in GLLRM, which are extensions of the RM that allow these two specific departures. When a GLLRM adjusts only for uniform LD, the sufficiency of the sum score is not affected, but most probably the reliability of the instrument will appear lower than when LD is not taken into account. If a GLLRM includes uniform DIF, the sum score is no longer a sufficient statistic for the person parameter; however, adjusting the sum scores for DIF enables subsequent comparisons of subgroup scores without measurement bias (Kreiner & Christensen, 2007), and similarly person parameters can be estimated for each subgroup to overcome the bias.

Item analysis

Considering the cultural differences between the two subsamples, our approach was to first analyse the MS and PC subscales independently in the samples, and second to analyse the samples jointly. With this approach, it would be easier to determine any DIF related to cultural differences in the joint sample, as any other measurement issue with items had already been discovered in each of the samples. The item analysis of each subscale in each sample followed the same overall strategy. Initially, fit to the RM was assessed, and if a scale did not fit the RM, we proceeded to catalog the departures, and if possible we proceeded to test the fit of the item responses to a GLLRM. When we were not able to successfully define a GLLRM, we eliminated the most (statistically and content-wise) problematic item and proceeded again to test fit to the RM for the reduced scale and so on, in an iterative process.

Overall tests of fit (i.e. tests of global homogeneity by comparison of item parameters in low and high scoring groups, and tests no DIF) were conducted using Andersen (1973) conditional likelihood ratio test (CLR). The fit of individual items was tested by comparing the observed item-rest-score correlations with the expected item-restscore correlations under the model (Kreiner & Christensen, 2004) and with conditional infit and outfit statistics (Christensen & Kreiner, 2012). The lack of local independence and DIF was tested in two ways: (a) conditional tests of independence using partial Goodman-Kruskal gamma coefficients for the conditional association between item pairs (indicating presence of LD) or between items and exogenous variables (indicating presence of DIF) given the restscores (Kreiner & Christensen, 2004); and (b) Kelderman’s (1984) conditional likelihood ratio test of no DIF/no LD. In addition, Kelderman’s (1984) test was used for confirmatory tests that the LD and DIF included in GLLRMs was warranted (Kreiner & Nielsen, 2013). Evidence of overall homogeneity and no global DIF found in the global tests was rejected if this was not supported by individual item fit and absence of LD and/or DIF at an item level. The Benjamini-Hochberg procedure was used to adjust for false discovery rate (FDR) due to multiple testing (Benjamini & Hochberg, 1995). In line with recommendations by Cox et al. (1977), we did not apply a critical limit of 5% for p-values as a deterministic decision criterion, but we used p-values as a continuous measure of evidence against the null, distinguishing between weak ($p < 0.10$), moderate ($p < 0.05$), and strong ($p < 0.01$) evidence.

Reliability was estimated using Hamon and Mesbah (2002) Monte Carlo method, which takes into account the conditional dependence between items in a GLLRM and adjusts the reliability accordingly (in contrast to Cronbach's α, which require local independence of items, as it will otherwise set the lower bound of reliability too high). Targeting was assessed numerically by two indices (Kreiner & Christensen, 2012): the Test Information (TI) target index, which is the mean test information divided by the maximum test information, and the Root Mean Squared Error (RMSE) target index, which is the minimum standard error of measurement divided by the mean standard error of measurement. Both indices should have a value close to one. We also estimated the target of the observed score and the standard error of measurement (SEM) of the observed score (i.e. the raw sum score). For a graphical illustration of targeting and test information, we plotted item maps showing the distribution of the item thresholds against weighted maximum likelihood estimates of the person parameters and the person parameters assuming a normal distribution (i.e. the theoretical distribution) and included the information function. Item analysis was conducted with Digram software (Kreiner, 2003; Kreiner & Nielsen, 2013) and item maps were created with R software (R Core Team, 2013).

Results

Item analysis with Aboriginal and non-Aboriginal samples

Aboriginal sample

Sense of Personal Control scale: The 12-item SPCS scale did not fit the RM. The analysis proceeded by investigating whether departures in terms of LD/DIF could be adjusted by GLLRM and removing misfitting items. The items 11, 4, 1 and 3 were sequentially excluded and all these items belonged to the MA subscale. Therefore, it became clear that the items from the MA subscale did not measure the same construct as the items from the PC subscale and both subscales could not form a unidimensional model. As a result of these findings, we analyzed each subscale separately.

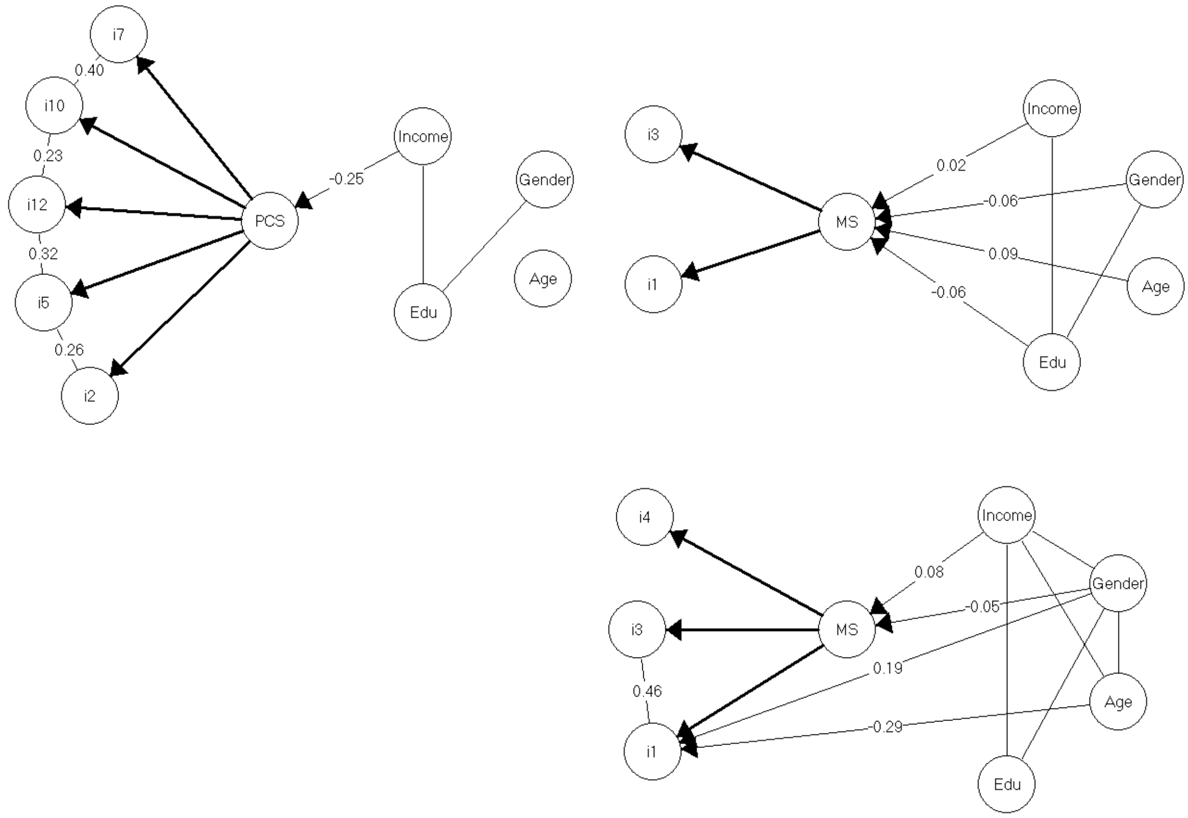
Perceived Constraints subscale: The 8-item PC subscale did not achieve overall fit the RM (Supplementary Table 3). We proceeded to investigate whether the departures consisted of LD and DIF and could be adjusted with GLLRM. However, we were unable to fit a GLLRM for the complete PC subscale with all 8 items. After several iterations investigating model departures, the misfitting items 6 ("I often feel helpless in dealing with life's problems"), 8 ("I have little control over the things that happen to me") and 9 ("There is really no way I can solve all the problems I have") were excluded and overall fit to a GLLRM with LD among all items pairs, but no DIF according to sex, age, education and income was found (Table 1, Figure 1, Supplementary Table 4). There were no issues with item fit for the five retained items (Table 2).

Table 1. Overall fit statistics for the GLLRMs of the PC and MA subscales§

Overall Tests	PC (Aboriginal Australians)			MA (Aboriginal Australians)			MA (non-Aboriginal Australians)		
	CLR	*Df*	*p*	CLR	*df*	*p*	CLR	*df*	*p*
Homogeneity	89.6	80	.22	12.5	6	.05	52.6	34	.02
Global DIF relative to:									
Sex	92.1	80	.17	5.4	6	.50	29.3	26	.30
Age	96.8	80	.10	6.9	6	.33	38.2	30	.14
Education	115.9	80	.02	8.2	6	.22	42.4	34	.15
Income	107.9	80	.005	12.0	6	.06	56.6	34	0.009

Notes. PC: Personal Constraints Scale. MA: Mastery Scale. CLR: Conditional likelihood ratio. df: degrees of freedom. p: p-value. DIF: differential item function. Overall homogeneity test compares item parameters in approximately equal-sized groups of high and low scoring persons, while the global DIF test for DIF across the entire set of items. The critical limits for the p-values after adjusting for false discovery rate in the GLLRM were: 5% limit p = .01 and 1% limit p = .002 for the Aboriginal sample; 5% = 0.05 and 1% limit p = .002 for the non-Aboriginal sample. §The results displayed in this table refer to the reduced subscales after the exclusion of misfitting items.

Figure 1. GLLRM of the Perceived Constraints and Mastery subscales.



Note. GLLRM of the Perceived Constraints subscale for Aboriginal Australians (top left), Mastery subscale for Aboriginal Australians (top right) and Mastery subscale for non-Aboriginal Australians (bottom right). Disconnected nodes indicate that variables are conditionally independent and partial gamma coefficients (Goodman & Kruskal's γ)informs the magnitude of the LD and DIF.

Table 2. Item fit statistics for the GLLRM of the PC and MA subscales

	Conditional Infit			Conditional Outfit			Item-restscore association		
Items[§]	Observed	SE	*p*	Observed	SE	*p*	Observed γ	Expected γ	*p*
PC GLLRM									
Aboriginal Australians									
2	0.958	0.080	0.600	0.936	0.080	0.427	0.384	0.351	0.517
5	1.033	0.088	0.705	1.039	0.091	0.671	0.448	0.456	0.874
7	0.998	0.081	0.981	0.989	0.079	0.892	0.409	0.384	0.622
10	1.031	0.084	0.709	1.085	0.090	0.349	0.463	0.451	0.799
12	1.004	0.088	0.965	0.991	0.098	0.924	0.469	0.470	0.983
MA RM									
non-Aboriginal Australians									
1	1.091	0.112	0.418	0.997	0.145	0.991	0.809	0.799	0.784
3	1.091	0.112	0.418	0.997	0.145	0.991	0.809	0.799	0.784
MA GLLRM									
non-Aboriginal Australians									
1	1.073	0.032	0.02	1.004	0.048	0.94	0.701	0.711	0.490
3	0.936	0.038	0.08	0.859	0.057	0.01	0.786	0.754	0.021
4	1.011	0.025	0.66	0.962	0.043	0.38	0.596	0.579	0.334

Notes. γ = Goodman & Kruskal's gamma coefficients. PC: Perceived Constraints Scale. MA: Mastery Scale. §The results displayed in this table refer to the reduced subscales after the exclusion of misfitting items.

Mastery subscale: The 4-item MA subscale did not fit the Rasch model either (Supplementary Table 3). Since the model departures did not consist uniquely of LD and DIF, these departures could not be adjusted in a GLLRM. After the exclusion of two misfitting items, item 4 (“Whether or not I am able to get what I want was in my own hands”) and item 11 (“What happens to me in the future mostly depends on me”), the two remaining items, item 1 (“I can do just about anything I really set my mind to”) and item 3 (“When I really want to do something I usually find a way to”), fit the RM (Table 1, Figure 1). However, only two items cannot be considered as a scale, and so be interpreted with caution.

Non-Aboriginal sample

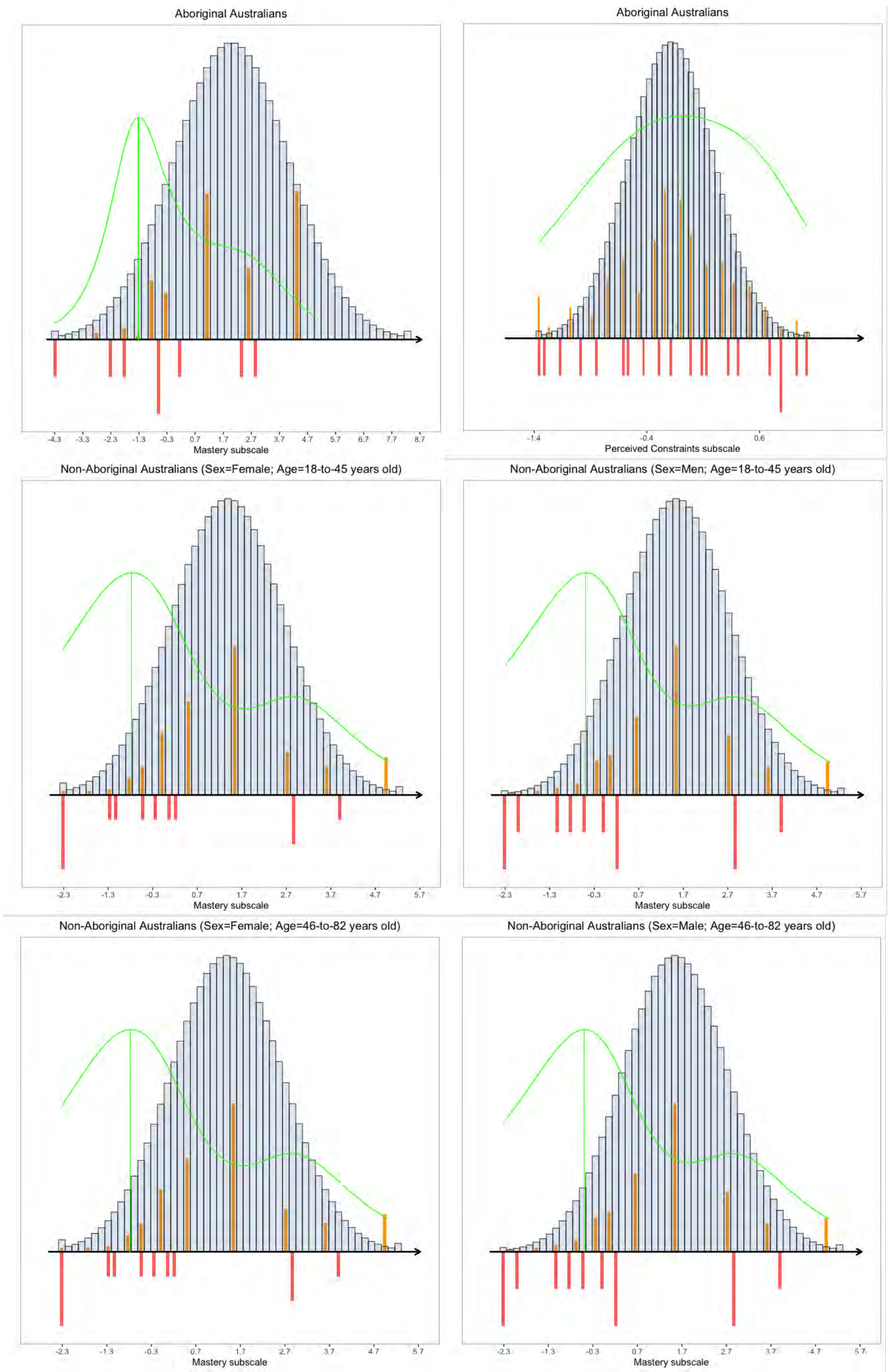
Sense of Personal Control scale and Perceived Constraints subscale: Similar to the results for the Aboriginal sample, the 12-item SPCS did not fit the RM. The item analysis again indicated that the items from the MA subscale did not measure the same construct as the items from the PC subscale, so we proceeded to evaluate the subscales separately. The 8-item PC subscale did not fit the RM (Supplementary Table 3). Despite several iterations investigating model departures, it was not possible to fit any model (RM or GLLRM) for the PC subscale in the non-Aboriginal sample.

Mastery subscale: The 4-item MA subscale did not fit the RM (Supplementary Table 3) and a GLLRM adjusting for LD and/or DIF could not be established either. After the exclusion of item 4 (“Whether or not I am able to get what I want is in my own hands”), which displayed misfit, the remaining 3 items fitted a GLLRM with age and gender DIF for item 1, as well as LD between items 1 and 3 (Table 1, Figure 1, Table 2).

Joint item analysis across Aboriginal and non-Aboriginal samples

Considering that only two items, Item 1 and 3 of the MA subscale, functioned for both Aboriginal and non-Aboriginal Australians, we were not able to proceed with the cross-cultural analyses of any of the two SPCS subscales, including testing for DIF across both cultures. Therefore, the SPCS did not provide unbiased measurement across Aboriginal and non-Aboriginal Australians and we were not able to achieve the main aim of this study. Nonetheless, it was possible to observe differences in targeting across the two samples (Figure 2).

Figure 2. Item Maps of the GLLRM.



Note. The orange bars display the person parameters (WML estimates). The grey bars display the population distribution under the assumption of normality. The red bars display the item thresholds and the green line is the Fisher's information function.

The targeting of the PC subscale for Aboriginal Australians was excellent (TI_{TI}=.94; $RMSE_{TI}$=0.96). For example, the TI target index indicated that for the Aboriginal sample the PC subscale provided 94% of the total information available if the instrument was perfectly targeted. The targeting of the 2-item MA subscale for Aboriginal Australians was poor, ranging from .42 to .56 across the DIF-defined subgroups. Furthermore, since this scale was composed of only 2 items, there was little information regarding where the majority of individuals were located. On the other hand, the targeting of the MA subscale for non-Aboriginal Australians was less than ideal and the TI target index ranged from .72 to .75 across the DIF-defined subgroups.

Discussion

The main aim of the present study was to evaluate whether the SPCS could be used to obtain fair and culturally unbiased measurement of personal control across Aboriginal and non-Aboriginal Australians, and thus also to investigate the psychometric properties of the SPCS in an Aboriginal and a non-aboriginal population. The findings indicated that: (a) a revised 5-item PC subscale was a measure of perceived constraints, however, the overlap in content among items led to poor reliability; (b) the revised MA subscale had only 2 items and new culturally-specific items should be developed before its application in Aboriginal Australians; (c) the SPCS did not provide culturally unbiased measurement across Aboriginal and non-Aboriginal samples. Culturally unbiased measurement was not possible since the SPCS psychometric properties in non-Aboriginal Australians were poor and only two items functioned properly in both cultures. In general, new culturally-specific items need to be developed and included to create a valid and reliable psychological instrument to measure sense of personal control in Aboriginal Australians.

Psychometric properties of the SPCS for Aboriginal Australians.

The findings indicated that the SPCS is composed of two subscales, PC and MA. In future use of the SPCS with Aboriginal Australians, total scores need to be computed for the PC and MA subscales separately. However, there was no conclusive evidence of the dimensionality across the two cultural samples, this thus remains a topic for future research.

Perceived Constraints subscale: One main finding of the current study was that *all five retained items* of the Perceived Constraints subscale were locally dependent. The content evaluation confirmed a large conceptual overlap between the PC items. For example, item 10 states that “I sometimes feel as I am being pushed around in my life” while item 7 states that “There are many things that interfere with that I want to do”. It seems reasonable that if a person has been “pushed around in life”, there will be “many things” interfering with what that person wants to do. Thus, we consider this a case of positive *response dependence*, in which endorsement of one item logically implies the endorsement of another (Marais, 2013) and the same is the case with the remaining local dependence. These results highlight the importance of investigating psychometric properties of instruments with models (such as GLLRM) that relax assumptions of local independence, which is, in fact, a requirement of all standard IRT and factor analytical models, since these assumptions are rarely met in practice.

One consequence of local dependence is inflated estimates of reliability. The reason being is that, since items are *too* conceptually similar, they are not different enough to provide one item worth of information (Linacre, 2000). For example, in our study, if a reliability index that assumes local independence, such as Cronbach’s (1951) alpha, was applied to the 5 items of the revised PC subscale, it would have indicated adequate reliability by standard conventions (α=0.71 – 95% CI [0.66, 0.76]). However, after adjusting for LD among all items pairs, Hamon and Mesbah’s (2002) Monte Carlo method indicated that the true reliability was poor (R=0.54). This means that, although the revised 5-item PC subscale provides a potentially valid measure of perceived constraints among Aboriginal Australians, measurement was not reliable in this sample, not even for research purposes and even less for individual assessment (Wells & Wollack, 2003). This result is worrisome since, although Aboriginal Australians comprise several culturally distinct groups, this population is notably homogeneous due to their experience of social inequalities as a whole (Carson et al., 2007). Hence, when psychological assessment is performed with Aboriginal Australians, the low trait variance (respondents are similar) needs to be compensated with higher measurement precision (Santiago, Roberts, Smithers, & Jamieson, 2019). The implication of these findings is that future studies need to develop culturally sensitive items to improve the PC subscale.

The 5 PC items displayed no DIF by sex, age, income or education. Therefore, scores (and person parameters) can be compared across these groups and will reflect true differences in perceived constraints rather than measurement bias. Nonetheless, the Aboriginal sample had a moderate size, so future studies should investigate DIF issues in a larger sample (i.e. with more power) to ensure that there was no DIF by these items. Finally, the targeting of the PC subscale was excellent, meaning that the items' "difficulty" was close to perfect for this population. That is, the PC subscale ranged from "easy" items (Item 10 – "I *sometimes* feel I am being pushed around in my life"), which were endorsed by participants with low perceived constraints, to "difficult" items (Item 7 – "There are *many* things that interfere with what I want to do"), which were endorsed by those with high perceived constraints. Although the local dependence between all item pairs complicates the disclosure of items' difficulty, the item maps (Figure 2) showed that the 5 items covered the whole range of perceived constraints in the Aboriginal population.

Mastery subscale: The analysis indicated that 2 MA items fit the RM. However, although in this case, the items were locally independent, two items cannot be considered as a scale and the results need to be interpreted with caution. The development of the SPCS by Lachman and Weaver (1998) included item 3 ("When I really want to do something I usually find a way to do it") and item 4 ("Whether or not I am able to get what I want is in my own hands"), which were added to the original items 1 ("I can do just about anything I really set my mind to") and 11 ("What happens to me in the future mostly depends on me") present in the Pearlin (1981) Mastery Scale. Despite the inclusions, the items 4 and 11 display misfit and were excluded. Problems with item 11 have been previously reported. For example, when evaluating the Pearlin (1981) Mastery Scale with Rasch analysis, Eklund et al. (2012, p. 387) showed that item 11 had the most pronounced misfit among the items and that this item may represent a different construct than the one measured by the scale as a whole". Therefore, in agreement with Eklund et al. (2012), we recommend item 11 to be excluded.

Considering the exclusion of two misfitting items, the two remaining items do not constitute a scale and it is implausible that these two items would cover enough content of a multifaceted construct such as mastery (Fok, Allen, Henry, & Mohatt, 2012), posing immediate concerns of construct underrepresentation (Messick, 1987). For this reason, it is required that future studies include culturally-specific items to evaluate mastery in Aboriginal Australians. Among these new items, one recommendation is the inclusion of items to measure *communal mastery* rather than *personal mastery*. While personal mastery promotes coping through individualized strategies, communal mastery improves coping through the use of the social network (Hobfoll, Schröder, Wells, & Malek, 2002). In a study with Indigenous American women, Hobfoll, Jackson, Hobfoll, Pierce, and Young (2002) showed that, while personal mastery was a strong predictor for coping with stress in individualistic cultures, communal mastery is more effective in enhancing coping in collectivistic cultures such as Indigenous populations. Due to these considerations, psychological instruments that measure both personal and communal mastery have been developed and one was recently validated in a *Yup'ik* population, an Indigenous group of Alaska natives (Fok et al., 2012). One example of an item measuring communal mastery is "What happens to me in the future depends on my ability to work well with others" (Jackson, MacKenzie, & Hobfoll, 2000), contrasting directly with item 11 ("What happens to me in the future mostly depends on me") which was eliminated due to misfit in the current study. In addition to items measuring communal mastery, future studies should also include "harder" items measuring personal mastery in an

Aboriginal population (i.e. items that require a higher degree of mastery to be endorsed). For example, items can evaluate whether participants feel they have control in face of overwhelming demands for Aboriginal Australians, such as racism and removal from their land (Reilly et al., 2008; Ziersch et al., 2011).

Psychometric properties of the SPCS for non-Aboriginal Australians

The psychometric properties of the SPCS were poor for non-Aboriginal Australians. It was not possible to obtain fit to a model with the Perceived Constraints subscale and the 3-item Mastery subscale had several problems in terms of DIF and LD.

Unbiased measurement across Aboriginal and non-Aboriginal Australians

One challenge for the research on social determinants of Aboriginal health is that their culture is different from European-descendent cultures and psychological instruments developed in Western countries shouldn't be assumed as valid and reliable. In the current study, the fact that we were not able to conduct the cross-cultural comparisons across Aboriginal and non-Aboriginal Australians reinforce the need for validation of psychological instruments specifically for Aboriginal Australians, and that comparability of scores between Aboriginal and non-Aboriginal groups should not be assumed and needs to be investigated. We recommended that future studies are undertaken aimed at first modifying and extended the SPCS scales based on the current results (e.g. using Nielsen & Kreiner's (2013) strategy for item improvement), and second aimed at investigating the validity and reliability of these new scales, before attempting to assess whether they are suitable for unbiased measurement across Aboriginal and non-Aboriginal populations.

Strengths and limitations. The strengths of the present study include the use of item response theory methods to evaluate issues of DIF and LD. Another strength is the size of the sample used for validation of psychological instruments for Aboriginal Australians, due to notably difficulty in recruiting participants from Indigenous populations, this data is one of the best available data for investigating the psychometric properties of a sense of personal control measure in an Aboriginal population. Moreover, we also employed a large non-Aboriginal sample for the analysis of cross-cultural validity. Limitations include the fact that the Aboriginal sample was a convenience sample in a rural setting and was composed mostly of women. Therefore, it is unclear whether the analysis had enough power to detect DIF by sex, and the absence of DIF by sex needs to be replicated in independent Aboriginal samples. Furthermore, many exogenous variables present in the original studies were not comparable across Aboriginal and non-Aboriginal Australians, which limited our possibilities of analysis of cross-cultural validity, and thus the issue of fairness in measurement across the two cultures.

Conclusions

In the present study, we showed that the development of new culturally-specific items is needed before the revised SPCS might constitute a valid and reliable measure of sense of personal control in both Aboriginal and non-aboriginal Australian populations, thus making it possible to assess whether the SPCS can measure fairly across these two populations.

Acknowledgments

The authors gratefully acknowledge the support of participants in both the pilot and parent Teeth Talk study. We also acknowledge the members of the advisory group, Teeth Talk study staff, and Pika Wiya Health Service.

Funding

The study was partially funded by The Carlsberg Foundation (project grant CF18-0384).

The Teeth Talk study was funded by the National Health and Medical Research Council of Australia (NHMRC Project Grant #627101).

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*(1), 123-140.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological), 57*(1), 289-300.

Carson, B., Dunbar, T., Chenhall, R. D., & Bailie, R. (2007). *Social determinants of Indigenous health*. Crows Nest, NSW: Allen & Unwin.

Chen, Y.-L., Hsiung, P.-C., Chung, L., Chen, S.-C., & Pan, A.-W. (2013). Psychometric properties of the Mastery Scale-Chinese version: Applying classical test theory and Rasch analysis. *Scandinavian Journal of Occupational Therapy, 20*(6), 404-411.

Cheng, C., Cheung, S. F., Chio, J. H.-m., & Chan, M.-p. S. (2013). Cultural meaning of perceived control: A meta-analysis of locus of control and psychological symptoms across 18 cultural regions. *Psychological Bulletin, 139*(1), 152.

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. *Journal of Abnormal Psychology, 100*(3), 316.

Cox, D. R., Spjøtvoll, E., Johansen, S., van Zwet, W. R., Bithell, J., Barndorff-Nielsen, O., & Keuls, M. (1977). The role of significance tests [with discussion and reply]. *Scandinavian Journal of Statistics*, 49-70.

Cox, L., & Taua, C. (2012). Cultural safety: Cultural considerations. In *Jarvis's physical examination and health assessment* (pp. 40-60): Elsevier.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Daniel, M., Brown, A., Dhurrkay, J., Cargo, M. D., & O'Dea, K. (2006). Mastery, perceived stress and health-related behaviour in northeast Arnhem Land: a cross-sectional study. *International Journal for Equity in Health, 5*(1), 10.

Eklund, M., Erlandsson, L.-K., & Hagell, P. (2012). Psychometric properties of a Swedish version of the Pearlin Mastery Scale in people with mental illness and healthy people. *Nordic Journal of Psychiatry, 66*(6), 380-388.

Fok, C. C. T., Allen, J., Henry, D., & Mohatt, G. V. (2012). Multicultural Mastery Scale for Youth: Multidimensional assessment of culturally mediated coping strategies. *Psychological Assessment, 24*(2), 313.

Hamon, A., & Mesbah, M. (2002). Questionnaire reliability under the Rasch model. In *Statistical methods for quality of life studies* (pp. 155-168): Springer.

Hobfoll, S. E., Jackson, A., Hobfoll, I., Pierce, C. A., & Young, S. (2002). The impact of communal-mastery versus self-mastery on emotional outcomes during stressful conditions: A prospective study of Native American women. *American Journal of Community Psychology, 30*(6), 853-871.

Hobfoll, S. E., Schröder, K. E., Wells, M., & Malek, M. (2002). Communal versus individualistic construction of sense of mastery in facing life challenges. *Journal of Social and Clinical Psychology, 21*(4), 362-399.

Jackson, T., MacKenzie, J., & Hobfoll, S. E. (2000). Communal aspects of self-regulation. In *Handbook of self-regulation* (pp. 275-300): Elsevier.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49*(2), 223-245.

Krause, N., & Shaw, B. A. (2000). Role-specific feelings of control and mortality. *Psychology and Aging, 15*(4), 617.

Kreiner, S. (2003). *Introduction to DIGRAM*: Department of Biostatistics, University of Copenhagen Copenhagen, Denmark.

Kreiner, S. (2007). Validity and objectivity: Reflections on the role and nature of Rasch models. *Nordic Psychology, 59*(3), 268-298.

Kreiner, S. (2012). The Rasch model for dichotomous items. *Rasch models in health*, 5-26.

Kreiner, S., & Christensen, K. B. (2002). Graphical rasch models. In *Statistical methods for quality of life studies* (pp. 187-203): Springer.

Kreiner, S., & Christensen, K. B. (2004). Analysis of local dependence and multidimensionality in graphical loglinear Rasch models. *Communications in Statistics-Theory and Methods, 33*(6), 1239-1276.

Kreiner, S., & Christensen, K. B. (2007). Validity and objectivity in health-related scales: Analysis by graphical loglinear Rasch models. In *Multivariate and mixture distribution Rasch models* (pp. 329-346): Springer.

Kreiner, S., & Christensen, K. B. (2012). Person parameter estimation and measurement in Rasch models. *Rasch models in health*, 63-78.

Kreiner, S., & Nielsen, T. (2013). Item analysis in DIGRAM 3.04: Part I: Guided tours.

Lachman, M. E., & Weaver, S. L. (1998). The sense of control as a moderator of social class differences in health and well-being. *Journal of Personality and Social Psychology, 74*(3), 763.

Lee, J., Lim, N., Yang, E., & Lee, S. M. (2011). Antecedents and consequences of three dimensions of burnout in psychotherapists: A meta-analysis. *Professional Psychology: Research and Practice, 42*(3), 252.

Linacre, J. (2000). Redundant items, overfit and measure bias. *Rasch Measurement Transactions, 14*(3), 755.

Lundgren, O., Garvin, P., Andersson, G., Jonasson, L., & Kristenson, M. (2018). Inverted items and validity: A psychobiological evaluation of two measures of psychological resources and one depression scale. *Health psychology open, 5*(1), 2055102918755045.

Marais, I. (2013). Local Dependence. In S. K. K. B. Christensen, & M. Mesbah (Ed.), *Rasch models in health* (pp. 111-130). London, England: Wiley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Mesbah, M., & Kreiner, S. (2012). Rasch models for ordered polytomous items. *Rasch models in health*, 27-42.

Nielsen, T. & Kreiner S. (2013). Improving items that do not fit the Rasch model: exemplified with the physical functioning scale of the SF-36. *Annales de L'I.S.U.P. Publications de L'Institut de Statistique de L'Université de Paris, Numero Special, 57*(1-2), 91-108.

Paradies, Y. C., & Cunningham, J. (2012). The DRUID study: Exploring mediating pathways between racism and depressive symptoms among Indigenous Australians. *Social Psychiatry and Psychiatric Epidemiology, 47*(2), 165-173.

Parker, E. J., Misan, G., Chong, A., Mills, H., Roberts-Thomson, K., Horowitz, A. M., & Jamieson, L. M. (2012). An oral health literacy intervention for Indigenous adults in a rural setting in Australia. *BMC Public Health, 12*(1), 461.

Pearlin, L. I., Menaghan, E. G., Lieberman, M. A., & Mullan, J. T. (1981). The stress process. *Journal of Health and Social Behavior*, 337-356.

Presson, P. K., & Benassi, V. A. (1996). Locus of control orientation and depressive symptomatology: A meta-analysis. *Journal of Social Behavior and Personality, 11*(1), 201.

R Core Team. (2013). R: A language and environment for statistical computing.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

Reilly, R. E., Doyle, J., Bretherton, D., Rowley, K. G., Harvey, J. L., Briggs, P., . . . Atkinson, V. (2008). Identifying psychosocial mediators of health amongst Indigenous Australians for the Heart Health Project. *Ethnicity and Health, 13*(4), 351-373.

Rosenbaum, P. R. (1989). Criterion-related construct validity. *Psychometrika, 54*(4), 625-633.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied, 80*(1), 1.

Sanders, A. E., & Slade, G. D. (2011). Gender modifies effect of perceived stress on orofacial pain symptoms: National Survey of Adult Oral Health. *Journal of Orofacial Pain, 25*(4).

Santiago, P. H. R., Roberts, R., Smithers, L. G., & Jamieson, L. (2019). Stress beyond coping? A Rasch analysis of the Perceived Stress Scale (PSS-14) in an Aboriginal population. *PloS One, 14*(5), e0216333.

Seligman, M. E., & Maier, S. F. (1967). Failure to escape traumatic shock. *Journal of Experimental Psychology, 74*(1), 1.

Shateri, Z. E., Lavasani, F. F., Farid, A.-A. A., Dehghani, M., Attar, S., & Barzegar, N. (2018). Assessing mastery through psychotherapy: Psychometric properties of the persian version of the mastery scale. *Iranian Journal of Psychiatry and Behavioral Sciences, 12*(1).

Slade, G. D., Spencer, A., & Roberts-Thomson, K. F. (2007). *Australia's Dental Generations: The National Survey of Adult Oral Health, 2004-06*: Australian Institute of Health and Welfare.

Spector, P. E., Sanchez, J. I., Siu, O. L., Salgado, J., & Ma, J. (2004). Eastern versus Western control beliefs at work: An investigation of secondary control, socioinstrumental control, and work locus of control in China and the US. *Applied psychology, 53*(1), 38-60.

Togari, T., & Yonekura, Y. (2015). A Japanese version of the Pearlin and Schooler’s Sense of Mastery Scale. *SpringerPlus, 4*(1), 399.

Wang, Q., Bowling, N. A., & Eschleman, K. J. (2010). A meta-analytic examination of work and general locus of control. *Journal of Applied Psychology, 95*(4), 761.

Wells, C. S., & Wollack, J. A. (2003). An instructor’s guide to understanding test reliability. *Testing & Evaluation Services. University of Wisconsin*.

Ziersch, A. M., Gallaher, G., Baum, F., & Bentley, M. (2011). Responding to racism: Insights on how racism can damage health from an urban study of Australian Aboriginal people. *Social Science and Medicine, 73*(7), 1045-1053.

Supplementary Materials

Supplementary Table 1. The SPCS items.

Item number	Item content	Item present in Pearlin's (1981) Mastery Scale	Item present in Lachman and Weaver (1998) SPCS
1	I can do just about anything I really set my mind to	X	X
2	Other people decide most of what I can and cannot do		X
3	When I really want to do something I usually find a way to do it		X
4	Whether or not I am able to get what I want is in my own hands		X
5	There is little I can do to change many of the important things in my life	X	X
6	I often feel helpless in dealing with life's problems	X	X
7	There are many things that interfere with what I want to do		X
8	I have little control over the things that happen to me	X	X
9	There is really no way I can solve all the problems I have	X	X
10	I sometimes feel I am being pushed around in my life	X	X
11	What happens to me in the future mostly depends on me	X	X
12	What happens in my life is often beyond my control		X

Supplementary Table 2. Characteristic of the study participants.

	Teeth Talk Study (Aboriginal Sample)		National Survey of Adult Oral Health 2004-2006 (Non-Aboriginal Sample)	
	n	%	n	%
Age				
Mean	36.4		50.3	
SD	14.0		14.8	
Min/Max	18/82		18/82	
Missing	0	0%	0	0%
Sex				
Female	214	76%	2388	61.9%
Male	103	24%	1469	38.1%
Missing	0	0%	0	0%
Education				
High school or less	236	74%	1252	32.5%
TAFE or university	81	26%	2605	67.5%
Missing	0	0%	0	0%
Income				
Job	80	25%	2947	76.4%
Benefits	237	75%	668	17.3%
Missing	0	0%	242	6.3%

Note. Mean values, minimum, maximum and standard deviations; numbers and percentages.

TAFE, Technical and Further Education (trade school/college).

Supplementary Table 3. Overall tests of fit to the Rasch model for the PC and MA subscales§

Overall Tests	PC (Aboriginal Australians)			MA (Aboriginal Australians)			MA (Non-Aboriginal Australians)			PC (Non-Aboriginal Australians)		
	CLR	*df*	*p*	CLR	*df*	*p*	CLR	*df*	*p*	CLR	*df*	*p*
Homogeneity	129.4	31	<0.001	110.5	15	<0.001	233.4	15	<0.001	867.2	31	<0.001
DIF relative to:												
Sex	27.7	31	0.635	13.6	15	0.553	33.3	15	0.004	81.5	31	<0.001
Age	31.1	31	0.462	21.7	15	0.116	60.4	15	<0.001	120.5	31	<0.001
Education	53.2	31	0.008	16.2	15	0.372	17.5	15	0.288	327.2	31	<0.001
Income	48.9	31	0.021	18.0	15	0.262	36.3	15	0.002	443.7	31	<0.001

Notes. PC: Personal Constraints Scale. MA: Mastery Scale. CLR: Conditional likelihood ratio. df: degrees of freedom. p: p-value. DIF: differential item function. Overall homogeneity test compares item parameters in approximately equal-sized groups of high and low scoring persons, while the global DIF test for DIF across the entire set of items. The critical limits for the p-values after adjusting for false discovery rate in the RM were: 5% limit p = .01 and 1% limit p = .002 for the Aboriginal sample; 5% = 0.05 and 1% limit p = .002 for the non-Aboriginal sample. § The results displayed in this table refer to the original subscales with all items included.

Supplementary Table 4. Item fit statistics for the Rasch model M of the PC and MA subscales

	Conditional Infit			Conditional Outfit			Item-restscore association		
Items[§]	Observed	SE	*p*	Observed	SE	*p*	Observed γ	Expected γ	*p*
Aboriginal Australians									
PC									
2	1.345	0.075	<0.001	1.276	0.075	<0.001	0.395	0.493	0.017
5	1.095	0.076	0.209	1.116	0.075	0.121	0.461	0.493	0.430
6	0.966	0.075	0.650	0.974	0.075	0.725	0.544	0.488	0.182
7	0.997	0.076	0.965	1.006	0.076	0.936	0.508	0.486	0.614
8	0.785	0.075	0.004	0.775	0.076	0.003	0.645	0.488	<0.001
9	0.805	0.076	0.010*	0.801	0.076	0.008	0.633	0.483	<0.001
10	1.073	0.077	0.345	0.962	0.076	0.612	0.546	0.497	0.226
12	1.113	0.078	0.149	1.107	0.075	0.157	0.478	0.498	0.630
MA									
1	1.106	0.088	0.228	1.030	0.095	0.754	0.657	0.616	0.339
3	0.676	0.088	<0.001	0.658	0.097	<0.001	0.794	0.621	<0.001
4	0.971	0.082	0.728	0.934	0.090	0.465	0.670	0.620	0.231
11	1.589	0.091	<0.001	1.455	0.099	<0.001	0.504	0.614	0.012
Non-Aboriginal Australians									
PC									
2	1.297	0.027	<0.001	1.231	0.026	<0.001	0.538	0.600	<0.001
5	1.296	0.026	<0.001	1.176	0.023	<0.001	0.569	0.617	<0.001
6	0.876	0.026	<0.001	0.853	0.025	<0.001	0.709	0.607	<0.001
7	1.249	0.024	<0.001	1.130	0.022	<0.001	0.560	0.601	<0.001
8	0.765	0.027	<0.001	0.761	0.025	<0.001	0.747	0.606	<0.001
9	0.876	0.027	<0.001	0.884	0.025	<0.001	0.690	0.612	<0.001
10	1.003	0.025	0.914	0.987	0.023	0.590	0.639	0.609	0.006
12	1.075	0.025	0.003	1.021	0.023	0.367	0.631	0.608	0.041
MA									
1	0.926	0.032	0.022	1.000	0.027	0.997	0.701	0.711	0.490
3	0.804	0.036	<0.001	0.882	0.029	<0.001	0.786	0.754	0.021
4	0.873	0.032	<0.001	0.903	0.025	<0.001	0.596	0.579	0.334
11	1.303	0.035	<0.001	1.240	0.029	<0.001	0.596	0.579	0.334

Notes. γ = Goodman & Kruskal's gamma coefficients. PC: Perceived Constraints Scale. MA: Mastery Scale. § The results displayed in this table refer to the original subscales with all items. After the Benjamini-Hochberg procedure, statistical significance was adjusted as $p < 0.02$ for a 5% FDR in the Mastery subscale for Aboriginal Australians, $p < 0.02$ in the Perceived Constraints subscale for Aboriginal Australians, $p < 0.04$ in the Mastery subscale for non-Aboriginal Australians and $p < 0.02$ in the Perceived Constraints subscale for Aboriginal Australians, $p < 0.04$ in the Perceived Constraints subscale for non-Aboriginal Australians.

Supplementary Table 5. Kelderman's likelihood ratio tests of no DIF for the GLLRM of the PC subscale.

	Conditional Likelihood Ratio test			Obs γ
Item 2 & Sex:	lr = 0.21	df = 4	p = 0.99	
Item 5 & Sex:	lr = 10.07	df = 4	p = 0.04	0.10
Item 7 & Sex:	lr = 2.35	df = 4	p = 0.67	
Item 10 & Sex:	lr = 4.00	df = 4	p = 0.41	
Item 12 & Sex:	lr = 3.05	df = 4	p = 0.55	
Item 2 & Education:	lr = 8.27	df = 4	p = 0.08	
Item 5 & Education:	lr = 3.27	df = 4	p = 0.51	
Item 7 & Education:	lr = 10.21	df = 4	p = 0.04	-0.03
Item 10 & Education:	lr = 4.25	df = 4	p = 0.37	
Item 12 & Education:	lr = 0.66	df = 4	p = 0.95	
Item 2 & Income:	lr = 5.40	df = 4	p = 0.25	
Item 5 & Income:	lr = 4.77	df = 4	p = 0.31	
Item 7 & Income:	lr = 4.28	df = 4	p = 0.37	
Item 10 & Income:	lr = 1.58	df = 4	p = 0.81	
Item 12 & Income:	lr = 9.54	df = 4	p = 0.05	0.09
Item 2 & Age:	lr = 2.92	df = 4	p = 0.57	
Item 5 & Age:	lr = 5.21	df = 4	p = 0.27	
Item 7 & Age:	lr = 4.58	df = 4	p = 0.33	
Item 10 & Age:	lr = 4.05	df = 4	p = 0.40	
Item 12 & Age:	lr = 9.26	df = 4	p = 0.05	

Note. After the Benjamini-Hochberg procedure, statistical significance was adjusted as $p < 0.003$ for a 5% FDR.

Supplementary Table 6. Kelderman's likelihood ratio tests no DIF for the GLLRM of the MA subscale.

	Conditional Likelihood Ratio test			Obs γ
Item 1 & Sex:	lr = 4.39	df = 4	p = 0.36	
Item 2 & Sex:	lr = 0.31	df = 4	p = 0.96	
Item 1 & Education:	lr = 4.41	df = 4	p = 0.35	
Item 2 & Education:	lr = 3.97	df = 4	p = 0.27	
Item 1 & Income:	lr = 9.81	df = 4	p = 0.04	
Item 2 & Income:	lr = 5.47	df = 4	p = 0.14	
Item 1 & Age:	lr = 3.36	df = 4	p = 0.50	
Item 2 & Age:	lr = 5.21	df = 4	p = 0.16	

Note. After the Benjamini-Hochberg procedure, statistical significance was adjusted as $p < 0.006$ for a 5% FDR.

Chapter 7 – R function for Item Characteristic Curves

Rationale and examples

In this section, I discuss the rationale behind the development of the R function to plot Item Characteristic Curves and its relevance to the (1) work conducted in this PhD and to (2) future Rasch modelling by other researchers.

The Item Characteristic Curve is the graphical representation of the regression model that calculates expected item responses (outcome) as a function of the latent trait (exposure) according to the Rasch model (link function). Thus, the visual inspection of the Item Characteristic Curve is used for the evaluation of model fit since it is possible to examine whether the observed item responses diverge from the expected item responses (Lord, 1977). The visual inspection of Item Characteristic Curve is a *standard* procedure in item response theory and Rasch analysis, being widely employed in several fields, including psychology and education (Phillip & Ojo, 2017).

During my candidature, I developed the R function to plot Item Characteristic Curves since the available functions from mainstream item response theory and Rasch analysis R packages had notable limitations. For example, the Item Characteristic Curve function from R package *eRm* functioned only for dichotomous items (e.g. 0=Disagree, 1=Agee), while the three scales validated in this PhD were composed of polytomous items (1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly Agree). Furthermore, the R package *ltm* does plot Item Characteristic Curves for polytomous items but does not display observed item responses, so visual investigation of model fit was not possible. Hence, the available R functions could not be used to evaluate the fit of polytomous items.

The most complete Item Characteristic Curves at that time could be found in the private software RUMM2030 (Rasch Unimensional Measurement Models); however, it was impossible to integrate this function with the broader, free and open-source statistical framework R software. This is troublesome since the biggest advantaged of R is that it integrates functions created by independent researchers and research groups so these functions can "talk to each other". For example, it is possible to use a R package to calculate missingness of individual items, one package to conduct multiple imputation by chained equations and one package to perform Rasch analysis and plot Item Characteristic Curves. The R Project for Statistical Computing is increasingly accepted by researchers across all fieds and certain research groups dedicate exclusive time to Rasch analysis in R. For example, one of these groups is the ERRTG – European Rasch Research and Teaching Group. For an in-depth discussion of Rasch analysis of polytomous items using R software, please see Robison et al. (2019).

Hence, I developed the function initially for plotting graphs in the papers, but the function received attention from researchers in the area such as A/Prof. Tine Nielsen and Prof. Mariane Muller, evolving into a broader project to incorporate the function into R package *iarm*. The function will soon be available on CRAN as part of the next *iarm* update (mail correspondence with Prof. Mariane Muller). In summary, the development of the Item Characteristic Curve function (1) enabled the evaluation of the fit of polytomous items analysed in this PhD and (2) will allow other researchers to use utilities that were previously only available in private software.

In the next section, I briefly compare the developed Item Characteristic Curve function to other Item Characteristic Curve functions previously available in R packages for Rasch modelling.

Functions available from other R packages

The widely used package **eRM (**Extended Rasch Modeling) developed by A/Prof Patrick Mair from Harvard University offers the function **plotICC**.

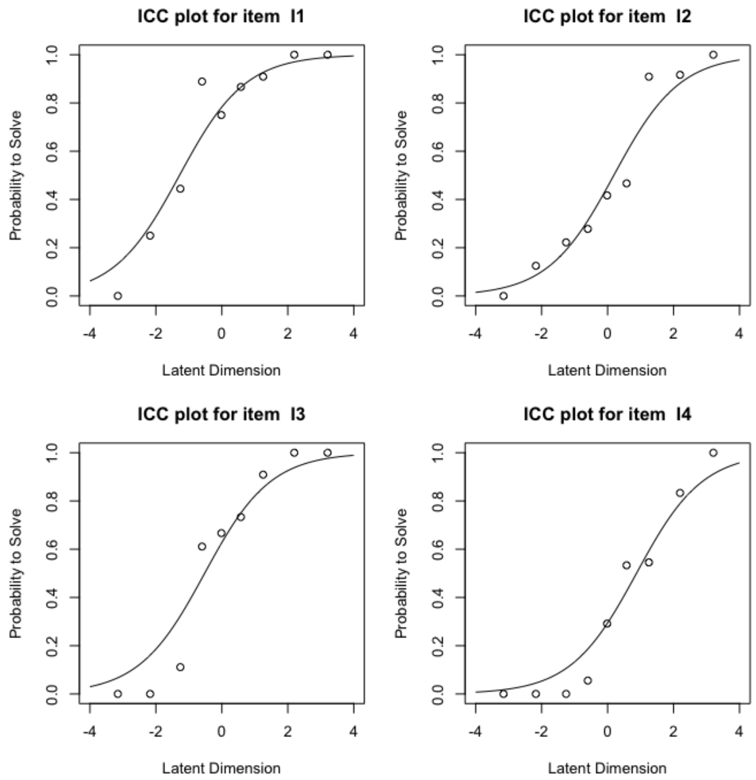


Figure 1. Item Characteristic Curve function for dichotomous items available in the R package *eRm*

The limitation of this function is that it is only available for dichotomous items. Notice on the plot above that the scores range from 0 to 1 (y-axis). Hence, this function would not be useful for this PhD project since all three psychological instruments (Perceived Stress Scale, Sense of Personal Control Scale and Social Support Scale) had polytomous items.

Another option was the package **ltm** (Latent Trait Models under IRT) developed by Prof. Dimitris Rizopoulos from Erasmus Medical Center Rotterdam offers the function **plot IRT**.

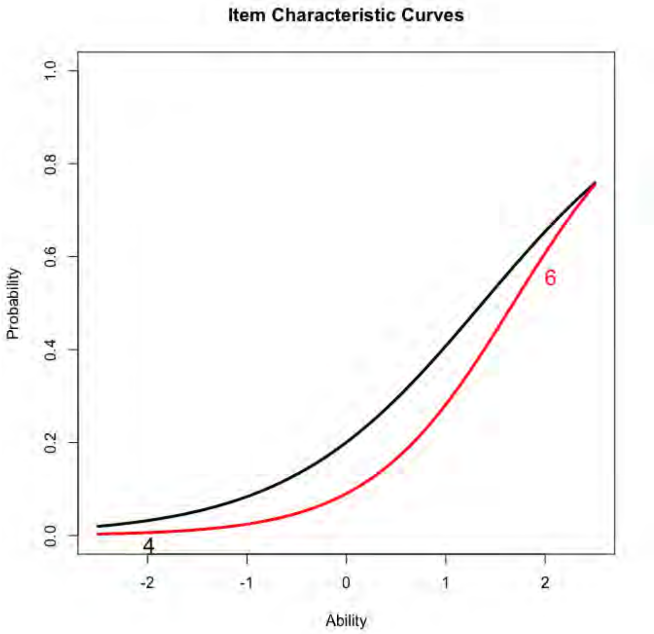


Figure 2. Item Characteristic Curve function for polytomous items available in the R package *ltm*

The limitation of this function is that it displays only the expected values but not the observed values. Notice on Figure 2 that the lines correspond to expected item responses from two items but the observed item responses (dots) could not be displayed. Thus, this function was not useful for this PhD project since it could not be used to visually inspect item fit.

In our opinion, the best Item Characteristic Curve function available at the time, from private software RUMM2030 created by Prof. David Andrich from The University of Western Australia.

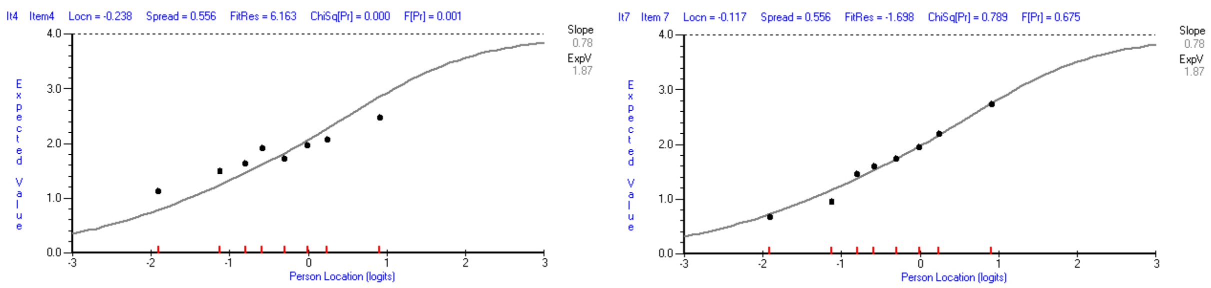


Figure 3. Item Characteristic Curve function for polytomous items available in the private software RUMM2030

The limitation of this function is that it is restricted to private software RUMM2030 and, therefore, cannot interact with other packages from the broader statistical framework R. These limitations led to the development of a new function. The function was initially developed as a private function to plot graphics required for the PhD papers but the project evolved. The function is now part of the package IARM (Marianne Mueller (2019). iarm: Item Analysis in Rasch Models. R package version 0.2.0. https://CRAN.R-project.org/package=iarm) and will be released over the next months on CRAN.

The function developed

The **ICCplot** (Item Characteristic Curve) was developed by me as part of the R package **iarm** (Item Analysis with Rasch Models) created by Prof. Marianne Mueller from Zurich University of Applied Sciences. The function can be used with polytomous items:

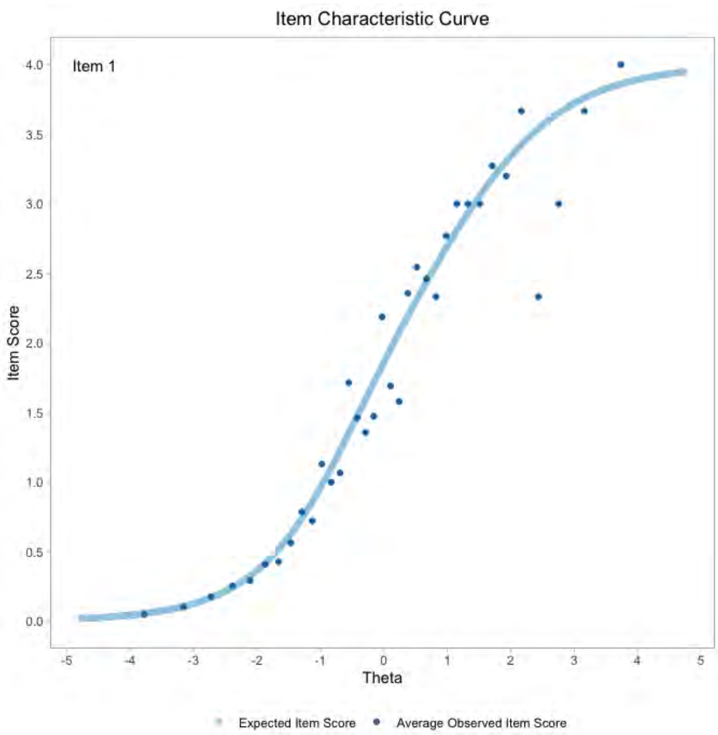


Figure 4. Item Characteristic Curve function for polytomous items using total scores

Or using class intervals instead of total scores:

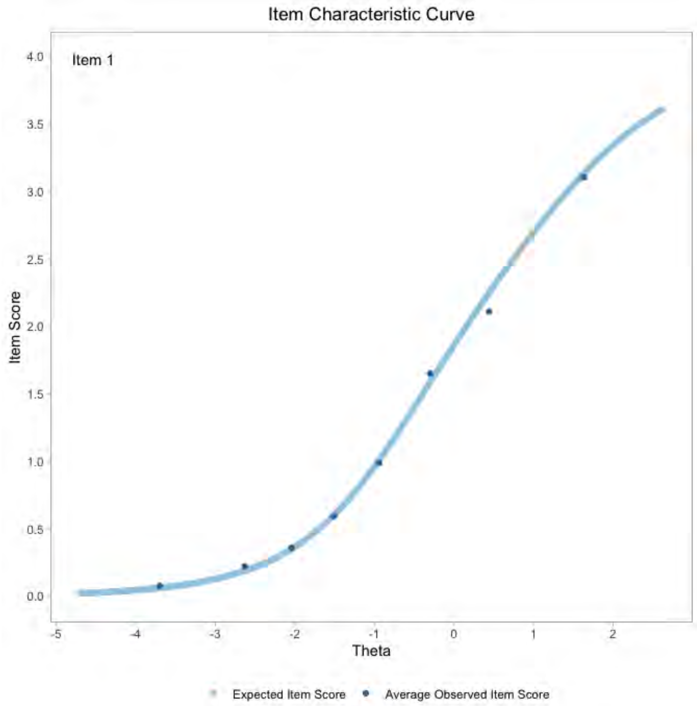


Figure 5. Item Characteristic Curve function for polytomous items using class intervals

Notice in Figure 5 that this is the same items as Figure 4 but now plotted with class intervals instead of total scores. The function also includes the option to plot observed scores according to values of an exogenous variable to evaluate differential item function:

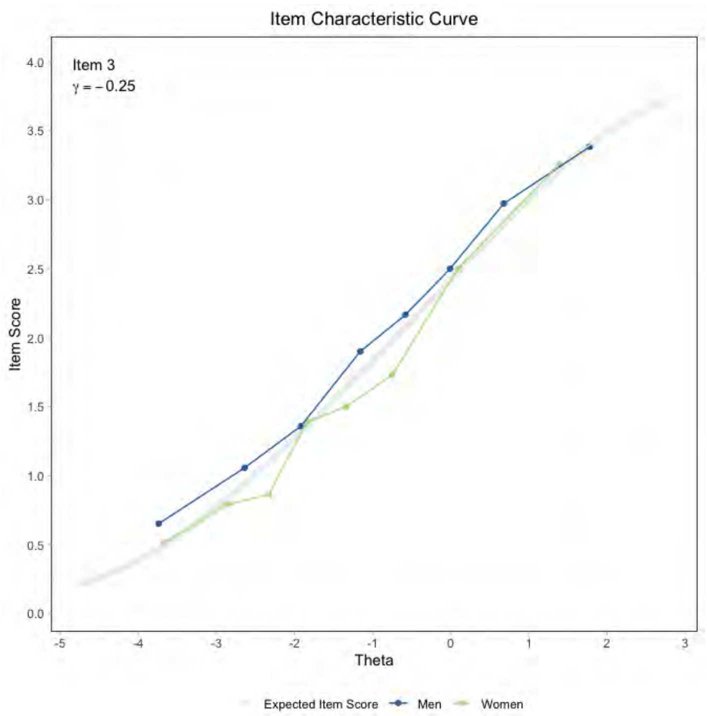


Figure 6. Item Characteristic Curve function for polytomous items with scores stratified according to values of an exogenous variable.

And the creation of multiple plots:

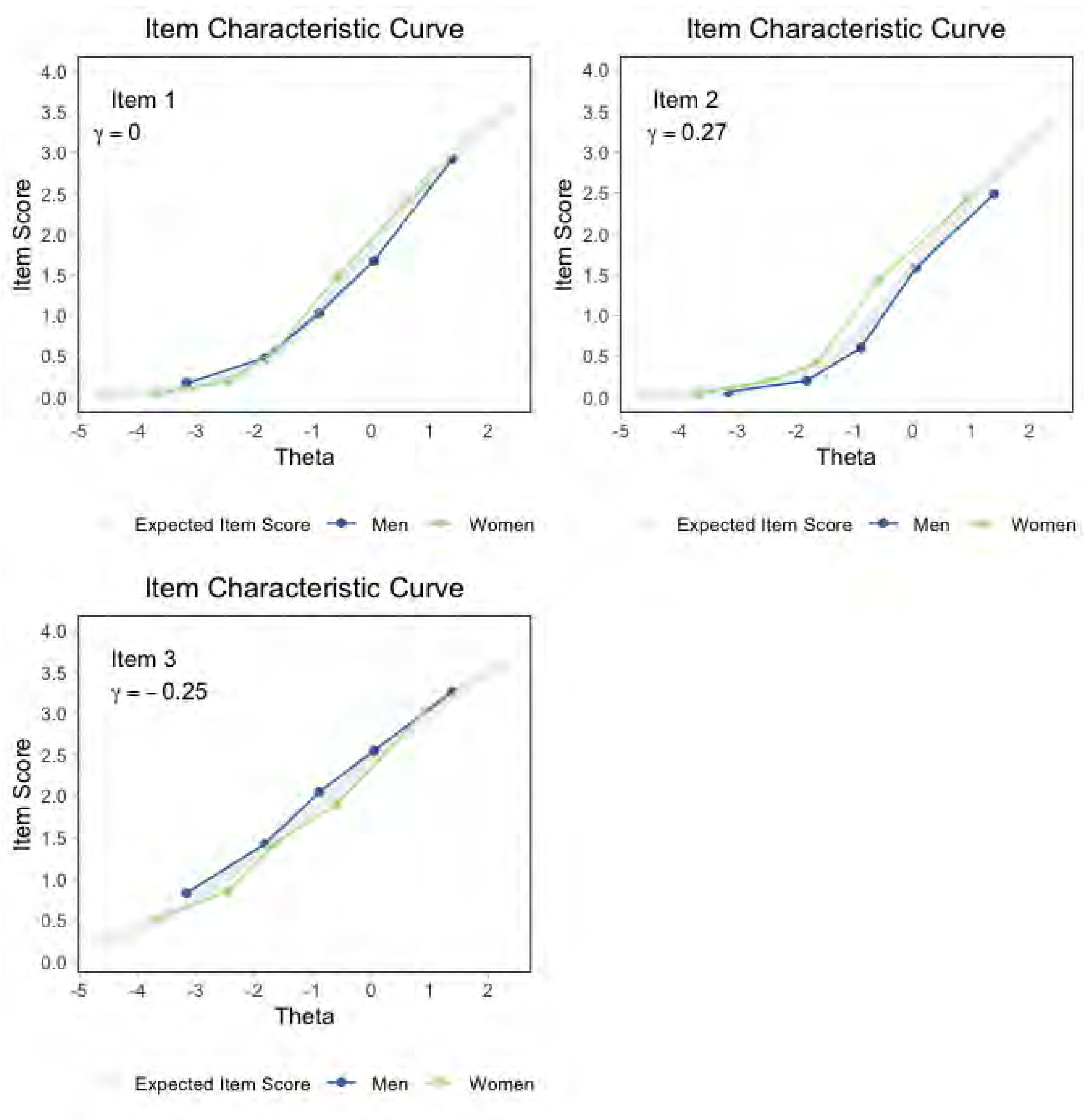
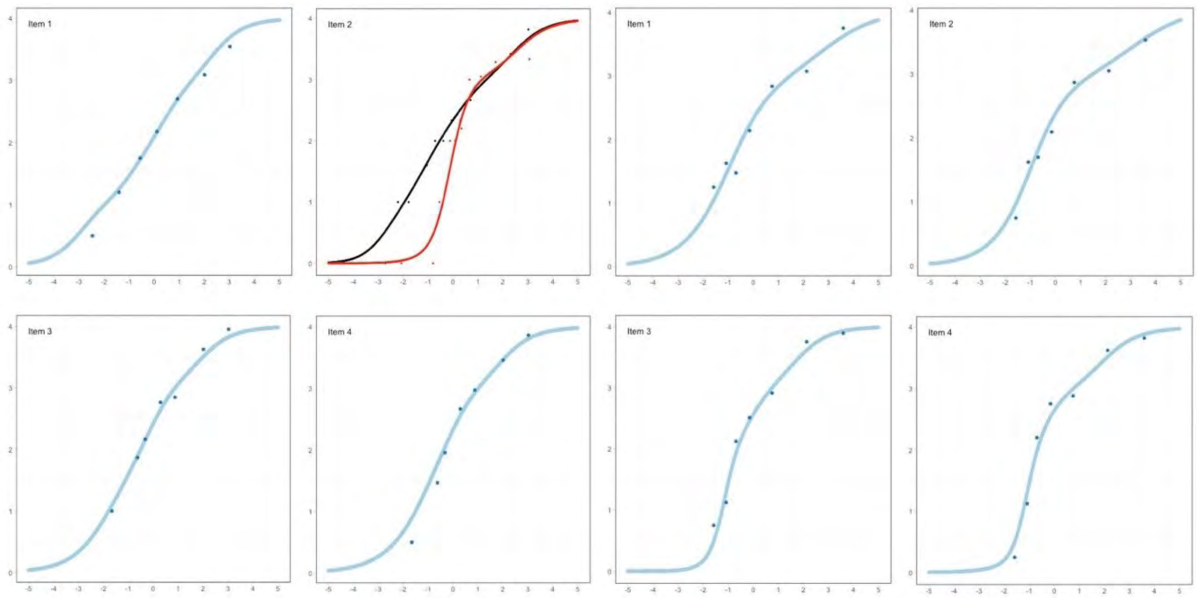


Figure 7. Plots of Item Characteristic Curve functions from three distinct items.

After development, the function was applied to generate the plots of the paper “Networks of support: psychometric properties of the Social Support Scale (SSS) in two Aboriginal samples”. We display one of the figures below.

Figure 2. Item characteristic curves of the GLLRM for Sample 1 (left) and Sample 2 (right).



Note. The x-axis indicates the latent trait ("Social support") and the y-axis indicates the item score. The points represent the *average* observed item responses in each of the seven class intervals. Since Item 2 had DIF by employment, the Item Characteristic Curves (ICCs) for participants employed (black) and unemployed (red) are displayed.

Help file

ICCplot {iarm2} R Documentation

Item Characteristic Curves

Description

Plots Item Characterist Curves for dichotomous and polytomous data. The plot can display observed scores as total scores (method="score") or as average scores within adjacent class intervals (method="cut"). Class intervals can be useful when the sample size is not large enough to contain an adequate number of respondents with the same total score for each possible total score. It is also possible to plot observed scores according to values of an exogenous variable to evaluate differential item function (dif="yes").

Usage

```
ICCplot(
  data,
  itemnumber,
  pallete = "Paired",
  xticks = 1,
  yticks = 0.5,
  thetain = -6,
  thetaend = 6,
  method = "score",
  grid = "yes",
  cinumber = 6,
  itemdescrip = "",
  axis.rumm = "yes",
  dif = "no",
  difvar = NA,
  diflabels = c("Group1", "Group 2", "Group 3", "Group 4", "Group5"),
  difstats = "yes",
  title = "Item Characteristic Curve",
  icclabel = "yes"
)
```

Arguments

`data`	An object of class "dataframe" containing item responses. The variables need to be numeric.
`itemnumber`	Subset of items to be plotted. An numeric vector indicating the columns in the data. Maximum of four items per plot.
`pallete`	An object of class "character". Pre-made color pallete from package RColorBrewer. Only available for dif="no".
`xticks`	Specify x-axis tick values.
`yticks`	Specify y-axis tick values.
`thetain`	Specify minimum theta values for person parameters.
`thetaend`	Specify maximum theta values for person parameters.
`method`	The method for displaying observed scores. "score" refers to plots with total scores. "cut" refers to plots with class intervals.

`grid`	Whether the background grid should be displayed. Options are "yes" or "no".
`cinumber`	The number of adjacent class intervals in which participants will be divided. The number of class intervals cannot be higher than the number of total scores.
`itemdescrip`	An character vector indicating the description of the plotted items.
`axis.rumm`	Configure whether the plot should display the entire trait range or solely the trait range close to the observed scores (similar to private software RUMM2030). Options are "yes" or "no".
`dif`	Configure whether the observed scores will be plotted according to values of an exogenous variable to evaluate differential item function. Options are "yes" or "no".
`difvar`	Choose the variable which will be used to evaluate differential item functioning. Only necessarwy when dif="yes".
`diflabels`	Choose the labels to values of the variable choosen to evaluate differential item functioning. Only necessary when dif="yes".
`difstats`	Display the partial gamma coefficients to indicate the magnitude of differential item functioning. Only necessary when dif="yes".
`title`	Title of the plot.
`icclabel`	Display the label indicating Expected Item Score and Observed Item Score. Options are "yes" or "no".

Author(s)

Pedro Henrique Ribeiro Santiago, (The University of Adelaide, pedro.ribeirosantiago@adelaide.edu.au)

Examples

```
# Creates a plot for Item 1 using total scores
ICCplotR(desc2[,5:13], itemnumber=1, method="score", itemdescrip="Item 1")

# Creates a plot for Item 1 using 8 class intervals
ICCplotR(desc2[,5:13], itemnumber=1, method="cut", cinumber=8,
itemdescrip="Item 1")

# Creates a plot for Item 1 using 8 class intervals without axis on RUMM
style
ICCplotR(desc2[,5:13], itemnumber=1, method="cut", cinumber=8,
itemdescrip="Item 1", axis.rumm="no")

# Creates a plot for Item 3 using 8 class intervals and evaluating DIF
according to gender
ICCplotR(desc2[,5:13], itemnumber=3, method="cut", cinumber=8,
itemdescrip="Item 3",
dif="yes", difvar=desc2$gender, diflabels=c("Men", "Women"),
difstats="yes")

# Creates a plot with three items using 5 class intervals and evaluated DIF
according to gender
ICCplotR(desc2[,5:13], itemnumber=c(1,2,3), method="cut", cinumber=5,
itemdescrip=c("Item 1","Item 2","Item 3"),
dif="yes", difvar=desc2$gender, diflabels=c("Men", "Women"),
difstats="yes")
```

R Syntax

```
### THE FUNCTION STARTS HERE ###

ICCplot <- function(data, itemnumber, pallete='Paired', xticks=1.0, yticks=0.5,
                 thetain=-6.000, thetaend=6.000, method="score", grid="yes", cinumber=6,
itemdescrip="",
                 axis.rumm="yes", dif="no", difvar=NA, diflabels=c("Group1", "Group 2",
"Group 3", "Group 4", "Group5"),
                 difstats="yes", title="Item Characteristic Curve", icclabel="yes") {

  pltC <- function() {

    if(grid=="no") {
      background <- element_blank()
      gridy <- element_blank()
      gridx <- element_blank()
      panelgrid <- element_blank()}
    else {background <- element_rect(fill = "white", colour="black")
    gridy <- element_blank()
    gridx <- element_blank()
    panelgrid <- element_line(colour="grey87", size=0.25)}

    if(icclabel=="yes") {
      icclabels <- "bottom"}
    else {icclabels <- "none"}

    annotations <- data.frame(
      xpos = c(-Inf,-Inf,Inf,Inf),
      ypos =  c(-Inf, Inf,-Inf,Inf),
      annotateText = itmn,
      hjustvar = c(-0.5) ,
      vjustvar = c(3.0))

    if (any(lapply(data,class)=="character")==TRUE)
    { stop(' "Input variables must be numeric" ')}
    else if (any(lapply(data,class)=="factor")==TRUE)
    { stop(' "Input variables must be numeric" ')}
    else {}

    for (i in 1:ncol(data))
    {
      seqres[i] <- all(abs(diff(sort(unique(data[,i])))) == 1)
      seqres
    }
    if (any(seqres=="FALSE")==TRUE)
    { stop(' "You need to provide the number of responses for all items categories.
          If there were zero responses to one category, please include this information in the
data" ')} else {}

    maxr <- max(data, na.rm=TRUE)
    minr <- min(data, na.rm=TRUE)
```

```
if(minr>0) {data2 <- (data - minr)}
if(minr==0) {data2 <- data}
else {}
adat <- as.data.frame(data2)
adat <- adat[complete.cases(adat),]
ttsc <- rowSums(adat, na.rm=TRUE)
rtsc <- rowSums(adat[,-itmc], na.rm=TRUE)
adat <- cbind(adat, ttsc, rtsc)
mxsc <- max(adat$ttsc)
adat <- adat[order(ttsc),]
itms <- ncol(adat)-2
pop <- seq(from=thetain, to=thetaend, by=0.01)
if ((maxr-minr)>1) {
  pcmo <- pcmodel(adat[,1:itms])
  prcu <- predict(pcmo, newdata=pop, type="cumprobability")
}
else {
  ramo <- raschmodel(adat[,1:itms])
  prcu <- predict(ramo, newdata=pop, type="cumprobability")
}
type <- 1
cure <- cbind(prcu, pop, type)
cure <- as.data.frame(cure)
cat <- table(adat[,itmc])
ncat <- length(cat)
pron=c()
for (i in 1:(ncol(cure)-2)) {
  pron[i]=(sum(cure[,i])==nrow(cure))
}
rcat <- which(pron == TRUE)
nrem <- length(rcat)
ini <- rcat[itmc] + 1
fin <- rcat[itmc] + (ncat-1)
mxcl <- ncat*itmc
mncl <- (ncat*(itmc-1))+2
if(ncat>2) {
  cend <- cbind(rowSums(cure[,ini:fin]), cure$pop, cure$type,
                rowSums(cure[,1:(length(cure)-2)])-nrem)
}
else {cend <- cbind(cure[,mxcl], cure$pop, cure$type, rowSums(cure[,1:(length(cure)-2)])-
nrem)}
cend <- as.data.frame(cend)
names(cend) <- c("score", "theta", "type", "ttsc")
if ((maxr-minr)>1) {
  ppar <- person_estimates(pcmo)
  ppar <- ppar[2:(nrow(ppar)-1),3]
}
else {
  ppar <- personpar(ramo)
}
mxpp <- length(ppar)
ppar <- ppar[1:mxpp]

if (method=="cut"&cinumber >= max(adat$ttsc))
```

```
  { stop('Number of class intervals need to be smaller than the total scores')}
  else {}

  if (method=="cut"){
    pcrv <- adat
    pmsc <- which(pcrv$ttsc==0)
    if (length(pmsc)==0) { pcrv <- pcrv}
    else {pcrv <- pcrv[-which(pcrv$ttsc==0),] }
    pxsc <- which(pcrv$ttsc==mxsc)
    if (length(pxsc)==0) { pcrv <- pcrv}
    else {pcrv <- pcrv[-which(pcrv$ttsc==mxsc),] }
    pcrv$class <- cut2(pcrv$ttsc, g=cinumber, oneval=TRUE, levels.mean = TRUE)}
  else {}

  if (method=="cut")
  {if (cinumber > length(unique(pcrv$class)))
  { stop('There are not enough subjects in each total score to produce this number of class
intervals')}
    else{}
  }
  else {}

  if(method=="cut"&dif!="yes") {
    obs <- aggregate(pcrv[,itmc], by=list(pcrv$class), FUN=mean)
    obs[,1] <- as.numeric(levels(obs[,1]))
    type <- 2
    x <- rep(NA, cinumber)
    for (i in 1:cinumber) {
      x[i] = cend$theta[which(abs(cend$ttsc-obs[i,1])==min(abs(cend$ttsc-obs[i,1])))]
    }
    x <- x[1:cinumber]
    obs <- cbind(obs[,2], x, type)
    obs <- as.data.frame(obs)
    names(obs) <- c("score", "theta", "type")
    cend <- cend[,1:3]
    cend <- as.data.frame(cend)
    names(cend) <- c("score", "theta", "type")
    cend <- rbind(cend, obs)
    cend[,1:3] <- as.numeric(unlist(cend[,1:3]))

    if(axis.rumm=="yes") {
      classbreak=c(
        max(min(subset(cend$theta, cend$type==1)),
            min((subset(cend$theta, cend$type!=1))-
                 1.0)),
        min(max((subset(cend$theta, cend$type==1))),
            max((subset(cend$theta, cend$type!=1)))+
              1.0))
    }
    else {classbreak=c(NA,NA)}

    myICCplot <<- ggplot(cend, aes(x=cend$theta, y=cend$score, col=as.factor(type))) +
#The graph itself
      ggtitle(title) +
```

```
    theme_light() +
    theme(plot.title = element_text(hjust = 0.5), legend.position=icclabels,
        panel.grid = panelgrid,
        panel.grid.major.y = gridy,
        panel.grid.major.x = gridx,
        panel.border = element_rect(colour="black", size=0.25, fill=NA),
        panel.background = background) +
    scale_x_continuous(breaks = round(seq(min(cend$theta), max(cend$theta), by =
xticks),1),
                limits=classbreak) +
    scale_y_continuous(breaks = round(seq(min(0), max(cend$score)+0.5, by = yticks),1)) +
    scale_color_brewer(palette = pallete, name="", labels=c("Expected Item Score",
"Average Observed Item Score")) +
    labs(y = "Item Score", x = "Theta") +
    geom_point(na.rm=TRUE) +

geom_text(data=annotations,aes(x=xpos,y=ypos,hjust=hjustvar,vjust=vjustvar,label=annotate
Text), color="black")

   assign(paste("plot", itmc, sep=""), myICCplot, envir = .GlobalEnv)
  }

  else if (method!="cut"&dif!="yes") {

   obs <- aggregate(adat[,itmc], by=list(adat$ttsc), FUN=mean)
   zero <- which(obs[,1]==0)
   if (length(zero)==0) { obs <- obs}
   else {obs <- obs[-which(obs[,1]==0),] }
   ext <- which(obs==mxsc)
   if (length(ext)==0) { obs <- obs}
   else {obs <- obs[-which(obs==mxsc),] }
   rest <- aggregate(adat[,itmc], by=list(adat$rtsc), FUN=mean)
   type <- 2
   tlsc <- seq(1, length(ppar), by=1)
   names(obs) <- c("score", "obsmean")
   ptmp <- cbind(ppar, tlsc)
   ptmp <- as.data.frame(ptmp)
   names(ptmp) <- c("theta", "score")
   obs <- merge(obs, ptmp, by="score")
   obs <- cbind(obs[,2:3], type)
   obs <- as.data.frame(obs)
   names(obs) <- c("score", "theta", "type")
   cend <- cend[,1:3]
   cend <- rbind(cend, obs)

   if(axis.rumm=="yes") {
    classbreak=c(
     max(min(subset(cend$theta, cend$type==1)),
       min((subset(cend$theta, cend$type!=1))-
          1.0)),
     min(max((subset(cend$theta, cend$type==1))),
       max((subset(cend$theta, cend$type!=1)))+
        1.0))
   }
```

```
    else {classbreak=c(NA,NA)}

    myICCplot <<- ggplot(cend, aes(x=cend$theta, y=cend$score, col=as.factor(type))) +
      ggtitle(title) + #Choose title
      theme_light() +
      theme(plot.title = element_text(hjust = 0.5), legend.position=icclabels,
            panel.grid = panelgrid,
            panel.grid.major.y = gridy,
            panel.grid.major.x = gridx,
            panel.border = element_rect(colour="black", size=0.25, fill=NA),
            panel.background = background) +
      scale_x_continuous(breaks = round(seq(min(cend$theta), max(cend$theta), by =
xticks),1),
                         limits=classbreak) +
      scale_y_continuous(breaks = round(seq(min(0), max(cend$score)+0.5, by = yticks),1)) +
      scale_color_brewer(palette = pallete, name="", labels=c("Expected Item Score",
"Average Observed Item Score")) +
      labs(y = "Item Score", x = "Theta") +
      geom_point(na.rm=TRUE) +

geom_text(data=annotations,aes(x=xpos,y=ypos,hjust=hjustvar,vjust=vjustvar,label=annotate
Text), color="black")

    assign(paste("plot", itmc, sep=""), myICCplot, envir = .GlobalEnv)
  }

  else if (method=="cut"&dif=="yes") {

    difd <- cbind(data2, difvar)
    difd <- difd[complete.cases(difd),]
    difd[,1:ncol(difd)] <- as.numeric(unlist(difd[,1:ncol(difd)]))
    difs <- min(difd$difvar)
    if (difs>1) {difd$difvar <- ((difd$difvar)-(difs-1))}
    else if (difs==1) {}
    else if (difs<1) {difd$difvar <- ((difd$difvar)+1)}
    else {}
    ldif <- list()
    ttsc <- list()
    obs <- list()
    zero <- list()
    ext <- list()
    class <- list()
    x <- list()
    for (i in 1:length(unique(difd$difvar))){
      x[[i]]<- vector()
    }

    for (i in 1:length(unique(difd$difvar))) {
      ldif[[i]] <- subset(difd, difd$difvar==i)
      ldif[[i]] <- ldif[[i]][complete.cases(ldif[[i]]),]
      ttsc[[i]] <- rowSums(ldif[[i]][,1:(ncol(ldif[[i]])-1)], na.rm=TRUE)
      ldif[[i]] <- cbind(ldif[[i]][,1:(ncol(ldif[[i]])-1)], ttsc[[i]])
      ldif[[i]] <- ldif[[i]][order(ldif[[i]][,ncol(ldif[[i]])]),]
```

```
    zero[[i]] <- which(ldif[[i]]$`ttsc[[i]]`==0)
    if (length(zero[[i]])==0) { ldif[[i]] <- ldif[[i]]}
    else {ldif[[i]] <- ldif[[i]][-which(ldif[[i]]$`ttsc[[i]]`==0),] }
    ext[[i]] <- which(ldif[[i]]==mxsc)
    if (length(ext[[i]])==0) { ldif[[i]] <- ldif[[i]]}
    else {ldif[[i]] <- ldif[[i]][-which(ldif[[i]]$`ttsc[[i]]`==mxsc),] }
    ldif[[i]]$class <- cut2(ldif[[i]][,ncol(ldif[[i]])], g=cinumber, oneval=TRUE, levels.mean
= TRUE)
    obs[[i]] <- aggregate(ldif[[i]][,itmc], by=list(ldif[[i]][,ncol(ldif[[i]])]), FUN=mean)
    type[i] <- i+1
    obs[[i]][,1] <- as.numeric(levels(obs[[i]][,1]))
  }

  for (i in 1:length(unique(difd$difvar))) {
    for (j in 1:nrow(obs[[i]])) {
      x[[i]][j] = cend$theta[which(abs(cend$ttsc-obs[[i]][j,1])==min(abs(cend$ttsc-
obs[[i]][j,1])))]
    }}

  for (i in 1:length(unique(difd$difvar))) {
    obs[[i]] <- cbind(obs[[i]][,2], x[[i]], type[i])
    obs[[i]] <- as.data.frame(obs[[i]])
    names(obs[[i]]) <- c("score", "theta", "type")}

  cend <- cend[,1:3]
  cend <- as.data.frame(cend)
  names(cend) <- c("score", "theta", "type")
  big_data = do.call(rbind, obs)
  cend <- rbind(cend, big_data)

  allg <- partgam_DIF(data2, difvar)
  pgmm <- allg[itmc,3]
  pgmm <- format(round(pgmm, digits=2), nsmall=2)

  if(difstats=="yes") {

    annotateDIF <- data.frame(
      xpos = c(-Inf,-Inf,Inf,Inf),
      ypos =  c(-Inf, Inf,-Inf,Inf),
      annotateText = paste("gamma == ", pgmm),
      hjustvar = c(-0.35) ,
      vjustvar = c(4.0)) }

  else {annotateDIF <- data.frame(
    xpos = c(-Inf,-Inf,Inf,Inf),
    ypos =  c(-Inf, Inf,-Inf,Inf),
    annotateText = c(""),
    hjustvar = c(-0.35) ,
    vjustvar = c(4.0)) }

  if(axis.rumm=="yes") {
    classbreak=c(
      max(min(subset(cend$theta, cend$type==1)),
          min((subset(cend$theta, cend$type!=1))-
```

```
            1.0)),
        min(max((subset(cend$theta, cend$type==1))),
           max((subset(cend$theta, cend$type!=1)))+
            1.0))
      }
      else {classbreak=c(NA,NA)}

      myICCplot <<- ggplot(cend, aes(x=cend$theta, y=cend$score, col=as.factor(type))) +
        ggtitle(title) + #Choose title
        theme_light() +
        theme(plot.title = element_text(hjust = 0.5), legend.position=icclabels,
           panel.grid = panelgrid,
           panel.grid.major.y = gridy,
           panel.grid.major.x = gridx,
           panel.border = element_rect(colour="black", size=0.25, fill=NA),
           panel.background = background) +
        scale_x_continuous(breaks = round(seq(min(cend$theta), max(cend$theta), by =
xticks),1),
                  limits=classbreak) +
        scale_y_continuous(breaks = round(seq(min(0), max(cend$score)+0.5, by = yticks),1)) +

scale_color_manual(values=c("#F0F0F0","#1F78B4","#B2DF8A","#33A02C","#FB9A99",

"#E31A1C","#FDBF6F","#FF7F00","#CAB2D6","#6A3D9A","#FFFF99",
                   "#B15928"), name="", labels=c("Expected Item Score", diflabels)) +
        labs(y = "Item Score", x = "Theta") +
        geom_point(na.rm=TRUE) +
        geom_line(na.rm=TRUE) +

geom_text(data=annotations,aes(x=xpos,y=ypos,hjust=hjustvar,vjust=vjustvar,label=annotate
Text), color="black") +

geom_text(data=annotateDIF,aes(x=xpos,y=ypos,hjust=hjustvar,vjust=vjustvar,label=annotat
eText), color="black", parse=TRUE)

      assign(paste("plot", itmc, sep=""), myICCplot, envir = .GlobalEnv)
    }

    else if (method!="cut"&dif=="yes") {

      difd <- cbind(data2, difvar)
      difd <- difd[complete.cases(difd),]
      difd[,1:ncol(difd)] <- as.numeric(unlist(difd[,1:ncol(difd)]))
      difs <- min(difd$difvar)
      if (difs>1) {difd$difvar <- ((difd$difvar)-(difs-1))}
      else if (difs==1) {}
      else if (difs<1) {difd$difvar <- ((difd$difvar)+1)}
      else {}

      ldif <- list()
      ttsc <- list()
      obs <- list()
      zero <- list()
      ext <- list()
```

```
for (i in 1:length(unique(difd$difvar))) {
  ldif[[i]] <- subset(difd, difd$difvar==i)
  ldif[[i]] <- ldif[[i]][complete.cases(ldif[[i]]),]
  ttsc[[i]] <- rowSums(ldif[[i]][,1:(ncol(ldif[[i]])-1)], na.rm=TRUE)
  ldif[[i]] <- cbind(ldif[[i]][,1:(ncol(ldif[[i]])-1)], ttsc[[i]])
  ldif[[i]] <- ldif[[i]][order(ldif[[i]][,ncol(ldif[[i]])]),]
  obs[[i]] <- aggregate(ldif[[i]][,itmc], by=list(ldif[[i]][,ncol(ldif[[i]])]), FUN=mean)
  zero[[i]] <- which(obs[[i]][,1]==0)
  if (length(zero[[i]])==0) { obs[[i]] <- obs[[i]]}
  else {obs[[i]] <- obs[[i]][-which(obs[[i]][,1]==0),] }
  ext[[i]] <- which(obs[[i]]==mxsc)
  if (length(ext[[i]])==0) { obs[[i]] <- obs[[i]]}
  else {obs[[i]] <- obs[[i]][-which(obs[[i]]==mxsc),] }
  type[i] <- i+1
  names(obs[[i]]) <- c("score", "obsmean")
  tlsc <- seq(1, length(ppar), by=1)
  ptmp <- cbind(ppar, tlsc)
  ptmp <- as.data.frame(ptmp)
  names(ptmp) <- c("theta", "score")
  obs[[i]] <- merge(obs[[i]], ptmp, by="score")
  obs[[i]] <- cbind(obs[[i]][,2:3], type[i])
  obs[[i]] <- as.data.frame(obs[[i]])
  names(obs[[i]]) <- c("score", "theta", "type")
}

cend <- cend[,1:3]
cend <- as.data.frame(cend)
names(cend) <- c("score", "theta", "type")
big_data = do.call(rbind, obs)
cend <- rbind(cend, big_data)

allg <- partgam_DIF(data2, difvar)
pgmm <- allg[itmc,3]
pgmm <- format(round(pgmm, digits=2), nsmall=2)

if(difstats=="yes") {

  annotateDIF <- data.frame(
    xpos = c(-Inf,-Inf,Inf,Inf),
    ypos =  c(-Inf, Inf,-Inf,Inf),
    annotateText = paste("gamma == ", pgmm),
    hjustvar = c(-0.35) ,
    vjustvar = c(4.0)) }

else {annotateDIF <- data.frame(
  xpos = c(-Inf,-Inf,Inf,Inf),
  ypos =  c(-Inf, Inf,-Inf,Inf),
  annotateText = c(""),
  hjustvar = c(-0.35) ,
  vjustvar = c(4.0)) }

if(axis.rumm=="yes") {
  classbreak=c(
```

```
        max(min(subset(cend$theta, cend$type==1)),
            min((subset(cend$theta, cend$type!=1))-
                1.0)),
        min(max((subset(cend$theta, cend$type==1))),
            max((subset(cend$theta, cend$type!=1)))+
              1.0))
    }
    else {classbreak=c(NA,NA)}

    myICCplot <<- ggplot(cend, aes(x=cend$theta, y=cend$score, col=as.factor(type))) +
      ggtitle(title) + #Choose title
      theme_light() +
      theme(plot.title = element_text(hjust = 0.5), legend.position=icclabels,
          panel.grid = panelgrid,
          panel.grid.major.y = gridy,
          panel.grid.major.x = gridx,
          panel.border = element_rect(colour="black", size=0.25, fill=NA),
          panel.background = background) +
      scale_x_continuous(breaks = round(seq(min(cend$theta), max(cend$theta), by =
xticks),1),
                    limits= classbreak) +
      scale_y_continuous(breaks = round(seq(min(0), max(cend$score)+0.5, by = yticks),1)) +

scale_color_manual(values=c("#F0F0F0","#1F78B4","#B2DF8A","#33A02C","#FB9A99",

"#E31A1C","#FDBF6F","#FF7F00","#CAB2D6","#6A3D9A","#FFFF99",
                          "#B15928"), name="", labels=c("Expected Item Score", diflabels)) +
      labs(y = "Item Score", x = "Theta") +
      geom_point(na.rm=TRUE) +
      geom_line(na.rm=TRUE) +

geom_text(data=annotations,aes(x=xpos,y=ypos,hjust=hjustvar,vjust=vjustvar,label=annotate
Text), color="black") +

geom_text(data=annotateDIF,aes(x=xpos,y=ypos,hjust=hjustvar,vjust=vjustvar,label=annotat
eText), color="black", parse=TRUE)

    assign(paste("plot", itmc, sep=""), myICCplot, envir = .GlobalEnv)
   }

  }

  if (length(itemnumber)>4)

  { stop(' "The function plots only a maximum of 4 items simultaneously" ')}

  else if (length(itemnumber)==1) {
    itmc=itemnumber
    itmn=itemdescrip[1]
    pltC()
    get(paste("plot", itmc, sep=""))
  }

  else {
```

```
    plst <- list()
    for (i in 1:length(itemnumber)) {
      itmc=itemnumber[i]
      itmn=itemdescrip[i]
      plst[[i]] <- pltC()
    }

    do.call(grid.arrange, args=(c(plst, nrow=2, ncol=2)))
    paste("Please press Zoom on the Plots window to see the plot")
  }
}
```

THE FUNCTION ENDS HERE

Chapter 8 – Final Considerations

Conclusion

Beyond Porteus and onward with Ewert: 21st-century psychological assessment of Aboriginal Australians

In 1922, after moving from Australia to Hawaii, Stanley Porteus founded the Psychological and Psychopathic Clinic at the University of Hawaii, where he worked for years as a lecturer and professor of clinical psychology. More than half a century later and after two-decades of student activism against racism, in the spring of 1998, the University of Hawaii Board of Regents decided to remove an honorable mention to Stanley Porteus at the Social Science Building. During the removal, the students and academics from the University of Hawaii argued that Porteus did not simply reflect “the professional opinion of his time”; he was a “professional racist”, due to anti-immigrant activism, advocacy of eugenics and creation of pseudo-scientific evidence to promote race ‘improvement’ throughout his entire career (Stannard, 1999, pp. 1, 9). For example, when Porteus conducted his maze testing with Aboriginal Australians, in one application there was “among his subjects one convicted murderer whose test performance was complicated by the presence of a chain on his leg and a police constable standing over him with a gun” (Klineberg, 1935, p. 156). These “assessments” led Porteus to remark that, although “imbecility can occur in both Australian Aborigines and Whites”, his research provided opposite evidence “until, of course, there appears an Aboriginal Shakespeare or Einstein or even a few Edisons” (Porteus, 1961, p.327).

The atrocious past of psychological assessment with Aboriginal Australians, which brought devastating consequences to Aboriginal people, remains unknown to many researchers and Australian citizens. The removal of any honorable mention to Stanley Porteus by the University of Hawaii in 1998 stated a clear message: racism will not be tolerated anymore in the 21st century. The second message came twenty years later, in 2018, when the Supreme Court of Canada (2018, p. 168) ruled that Canadian institutes (such as Correctional Services Canada) need to validate psychological instruments specifically for Canadian Indigenous people due to the risk of Western-developed tools "exhibiting cultural bias".

In this PhD project, I employed modern psychometric techniques to answer the research question: are the Perceived Stress Scale (PSS), Social Support Scale (SSS) and Sense of Personal Control Scale (SPCS) valid and reliable measures for an Aboriginal Australian population? The *empirical* results showed unambiguously that validity of Western-developed instruments for Aboriginal Australians is not a given. For example, the Perceived Stress Scale (aPSS-13) displayed initial evidence of validity only after one misfitting item was excluded. Furthermore, the weak association between the Perceived Stress and Perceived Coping was different from the moderate association constantly reported in Western populations (Lee, 2012). Additionally, the Sense of Personal Control Scale also exhibited problems and five items had to be removed. While Western-developed instruments *can* function properly, such as the Social Support Scale which displayed good psychometric properties in two independent Aboriginal samples, validation is still necessary to *ensure* that the instrument can be applied to Aboriginal Australians without further modifications (e.g. item deletion).

I joined previous efforts by researchers such as Kowal, Gunthorpe, and Bailie (2007) and Brown et al. (2013) to validate instruments specifically for Aboriginal Australians. Furthermore, I employed cutting-edge psychometric methodology to advance the field of psychological assessment in Aboriginal Australians and to provide the highest quality evidence regarding the validity of the Perceived Stress Scale (PSS), Social Support Scale (SSS) and Sense of Personal Control Scale (SPCS). Finally, we hope that the findings of this PhD project will be part of a large body of scientific evidence that one day will guide policymakers to fully guarantee fair assessment for Aboriginal Australians. That is, that assessment will not be conducted in Aboriginal Australians without using instruments validated specifically for this population. That is our hope for the future.

From little research, big things grow.

Appendix

Appendix 1 – Ethnic-racial identity affirmation: Validation in Aboriginal Australian children

PLOS ONE

RESEARCH ARTICLE

Ethnic-racial identity affirmation: Validation in Aboriginal Australian children

Davi Manzini Macedo[1]*, Pedro Ribeiro Santiago[1], Rachel M. Roberts[2], Lisa G. Smithers[3], Yin Paradies[4], Lisa M. Jamieson[1]

1 Indigenous Oral Health Unit, Adelaide Dental School, The University of Adelaide, Adelaide, Australia, 2 School of Psychology, The University of Adelaide, Adelaide, Australia, 3 BetterStart Child Health and Development Research Group, School of Public Health, The University of Adelaide, Adelaide, Australia, 4 Alfred Deakin Institute for Citizenship and Globalisation, Deakin University, Melbourne, Australia

* davi.manzinimacedo@adelaide.edu.au

OPEN ACCESS

Citation: Macedo DM, Santiago PR, Roberts RM, Smithers LG, Paradies Y, Jamieson LM (2019) Ethnic-racial identity affirmation: Validation in Aboriginal Australian children. PLoS ONE 14(11): e0224736. https://doi.org/10.1371/journal.pone.0224736

Editor: Geilson Lima Santana, University of Sao Paulo Medical School, BRAZIL

Received: June 5, 2019

Accepted: October 20, 2019

Published: November 7, 2019

Data Availability Statement: The data used in this study is property of the Australian Government Department of Social Services. Thus it cannot be shared publicly by the authors. There are security and confidentiality protocols for accessing LSIC data. Interested parts must submit an application and sign a deed of license. Information can be found on the LSIC webpage: http://www.dss.gov.au/lsic. The authors did not receive special access privileges to the data that others would not have. Interested researchers will be able to access the data in the same manner as the authors.

Abstract

Introduction

Positive attitudes towards ethnic-racial identity (ERI) is a key factor in Aboriginal Australian children's development. The present study aims to offer evidence of construct and criterion validity, reliability, and measurement invariance of a brief measure of Aboriginal children's ERI affirmation.

Methods

Data was from 424 children aged 10–12 years (mean 10.5 years; SD 0.56) participating in the 8^{th} wave of the Longitudinal Study of Indigenous Children (LSIC). Information on ERI was obtained from 4 child-reported items. Sociodemographic characteristics and child social and emotional outcomes were caregiver-reported. A factorial structure was tested by Confirmatory Factor Analysis. The estimation method was weighted least squares with mean and variance adjusted test statistic (WLSMV). For reliability verification, the ordinal α and Ω hierarchical α were assessed. For construct validity, a generalized linear model with log-Poisson link estimated the association between ERI and children's social and emotional outcomes. We hypothesized that children with positive ERI would have lower behavioural and emotional difficulties.

Results

We found evidence of excellent fit for a unidimensional model of ERI affirmation after adjusting for correlated uniqueness between items 1 and 3 ($\chi^2(2) = 0.06$, $p = 0.80$; RMSEA = 0.000 [90% CI 0.000–0.080], $p = 0.088$; CFI = 1.000). Internal consistency reliability was considered adequate (ordinal $\alpha = 0.83$; Ω hierarchical $\alpha = 0.72$). The unidimensional model was shown to be invariant among boys and girls ($\Delta\chi^2$ (4) = 6.20, $p = 0.18$; ΔCFI = 0.000). Higher ERI was associated with lower risk of problematic scores (>17) on the SDQ (Risk Ratio$_a$ = 0.91, 95% CI 0.64, 1.29).

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Discussion

The four LSIC items perform as a brief measure of Aboriginal children ERI affirmation among boys and girls. Results contribute much needed evidence for LSIC's ongoing success and to future research on Aboriginal children's development and wellbeing.

Introduction

Ethnic and racial minorities' feelings and beliefs about their ethnic and racial memberships have been a topic of interest in the social sciences for decades [1]. A recent effort to unify this field of research–the Ethnic and Racial Identity in the 21st Century Study Group- has proposed the adoption of a meta-construct to refer to this process of self-understanding and self-categorization. The ethnic-racial identity (ERI) concept was proposed to represent the perception of belonging to a social group across ethnic and racial groups from different heritages, nationalities, cultural backgrounds, and socialization experiences [1].

The process of identification with one's ethnic-racial group starts early in development. Children as young as five-years possess a sense of ERI and demonstrate that they explore, commit, and consolidate attitudes and preferences based on ethnic-racial membership [2–4]. During childhood ERI is associated with higher self-esteem, better adaptive behaviour and fewer externalizing and internalizing problems [3, 5]. Furthermore, positive attitudes and a sense of commitment to ERI are shown to exert a protective role against the effects of racism on racial minority children and adolescents' wellbeing [6, 7]. The attitude component of ERI has been referred to as ERI affirmation and can be observed from an early age [8]. Attitudes towards ERI are central to wellbeing and mental health, as feelings towards social identities (e.g., gender, race-ethnicity, nationality) are decisive in perceptions of self-esteem and global self-worth [9, 10].

Aboriginal Australians are the descendants of the occupants of the Australian continent prior to European colonisation [11]. There are diverse communities of Aboriginal Australians with unique traditions, political systems, cultural characteristics, and languages, living all across the Australian territory (from metropolitan centres to remote communities) [11]. It is estimated that Aboriginal Australians comprise approximately 3.3% of the Australian population, corresponding to 798,365 inhabitants accordingly to the last Australian Bureau of Statistics census, dated from 2016 [12]. As a disadvantaged group in Australia, due to a history of colonization and dispossession [13], Aboriginal Australians face a range of social inequalities (e.g. lower educational attainment and income, poor access to health services) [11] and can experience discrimination from early ages [14, 15]. Discrepancies in their mental health and wellbeing are also documented. A recent report on Aboriginal youth wellbeing suggested one third of participants (33%) indicated experiencing high to very high levels of psychological distress, against 13% of their non-Indigenous counterparts. Alarmingly, suicide was identified as one of the leading causes of death among Aboriginal Australians aged 10–24 between 2011 and 2015. [16].

Nonetheless, approaches have been proposed emphasizing the resilience of Aboriginal people in facing the adversities that affects this population. The importance of culture, spirituality, connection to land, ancestry, kinship, and a sense of pride about being Aboriginal have been consistently reported as a central determinant of Aboriginal Australians' health and social and emotional wellbeing across the lifespan [13, 17]. Research on Aboriginal perspectives of positive child development highlights the importance of a strong sense of attachment to culture and pride about their Aboriginal identity [14, 18]. Despite the relevance of ERI to both developmental psychology and the Aboriginal holistic perspective of health and development, limited data measuring this construct among Aboriginal children is currently available.

Footprints in Time: The Longitudinal Study of Indigenous Children (LSIC) is one of the few initiatives that have assessed ERI among Aboriginal children [19]. LSIC collects information on determinants of Aboriginal children's development across a wide range of communities and environments, including more than 80 Aboriginal clans and tribal groups across Australia. [20]Nonetheless, there is no published evidence regarding the validity and reliability of the ERI items used in data collection. Therefore, the present study aims to evaluate the construct validity and reliability of the ERI items as a measure of content/attitudinal ERI. Our hypothesis is that the 4 items used provide a brief unidimensional psychological instrument of how Aboriginal children perceive and feel about their ethnic-racial membership. A unidimensional instrument is one which the responses to all items (or, alternatively, the covariance between items) can be explained by a single latent variable [21]. That is, all the items measure a single underlying construct. In practical terms, methods such as factor analysis can show whether an instrument is unidimensional by evaluating if a one-factor model is a good fit for the data (compared to other models such as two or three-factor models, for example) and checking if "all items have substantial factor loadings on a single factor" [22]. The psychometric analysis will evaluate: a) the factorial structure of the items; b) measurement invariance by gender; c) reliability; and d) criterion validity.

To the best of our knowledge, this is the first study to assess the efficacy of a brief instrument targeting an affective component of Aboriginal children ERI (ERI affirmation). Additionally, LSIC is a pioneer study due not only to its longevity, but the diversity of children and families represented and the integration of Aboriginal cultural values and perspectives in its design and data collection [19]. Therefore, the verification of the validity of the measures applied, especially when concerning an aspect of central importance for Aboriginal Australians, may aid in its continued success.

Methods

Study design

LSIC employs an accelerated cross-sequential design aimed to collect information on the first nine to ten years of Aboriginal children's development in a six-year period. The study involves two cohorts. The B cohort includes children who were aged 0.5 to 2 years at wave 1. The K cohort consists of children aged 3.5 to 5 years at the beginning of the study. The content of the questionnaires is selected through consultation with working reference groups, community stakeholders from urban, regional and rural Indigenous communities, as well as academic institutions and government agencies [20]. Ethical approval for the content selection and data collection processes is obtained from the Human Research Ethics Committee of the Australian Institute of Aboriginal and Torres Strait Islander Studies [20]. LSIC waves occur annually between February and December. Data from waves 1 to 9 (2008–2016) is currently available upon application and a signed deed of license from the Australian Government Department of Social Services (DSS), the party responsible for conducting LSIC [23].

Data collection procedures

A non-random purposive sample was recruited from records of Centrelink and Medicare Australia, welfare and health-assistance programs, respectively [24, 25]. Signed consent was obtained from the eligible families who agreed to participate. Participants were also recruited through informal means of communication such as local study promotion and personal communication among community members. Interviews were conducted by Department of Social Services Aboriginal and Torres Strait Islander Research Administration Officers [20]. In wave 1, over 1,680 interviews were conducted with children's primary caregivers. A total of 1,255 interviews were conducted in wave 8 (2016), corresponding to an 87.2 retention rate from the previous wave [20]. Authors received permission to access de-identified data upon DSS's authorization [23]

Participants

Children in the K-Cohort participating in Wave 8 of LSIC were included in the analysis. Between both cohorts, there were 1,240 participating children. However, ERI was only assessed among the children in the K cohort (n = 496). Of those, 47 were excluded as caregivers did not authorize the research administration officers to administer the ERI affirmation items. Among the 449 children that responded to the measure, 9 were excluded due to missing values in at least one of the 4 items. Since our aim was to evaluate the validity and reliability of the 4-items for a specific age range, we focused on children aged 10–12 years (n = 435). Children aged 9 years (n = 5) or who had already turned 12 (n = 11) were removed due to small sample sizes. Our final sample thus comprised 424 Aboriginal children (51.3% males; mean age: 10.5 (SD 0.5) years).

Measures

Ethnic-racial identity affirmation measure. A set of four child self-report items was used to assess participant's ERI affirmation. All items had a 6-point Likert Scale response option, ranging from "Yes (Always)", "Yes (Most of the time)", "Sometimes (Fair bit)", "Sometimes (Little bit)", "No (Not much)", "No (Never)". Values from 1 to 6 were assigned to responses and reverse-coded so higher values would suggest higher ERI affirmation. Two other alternative response options were "Don't know" and "Refused", coded as missing. The 4-items were: 1) "I feel good about being Aboriginal and/or Torres Strait Islander in class"; 2) "I want to share (tell others) things about being Aboriginal and/or Torres Strait Islander in class"; 3) "I feel safe about being Aboriginal and/or Torres Strait Islander in class"; and 4) "I like people to know I am Aboriginal and/or Torres Strait Islander in class".

The measure was selected by the LSIC team after consultation with the LSIC steering committee and community stakeholders, as a standard procedure adopted to guarantee community participation and the integration of Aboriginal cultural values and perspectives [20]. The original items are part of a measure to assess cultural and Aboriginal educational strategies [26]. The items were originally presented as the factor "Strength of Cultural Identity". Two of the items were modified for use in LSIC. Item 3 was originally worded "I feel comfortable about my culture in class" and item 4 was "I am proud of my culture when I am in class".

Socio-demographic characteristics. Information on participant's age and sex was collected at wave 1 through an open and caregiver-reported question. For confounding adjustment in the criterion validity analysis, information on the family Level of Relative Isolation (LORI), and the index for Indigenous Socio-Economic Outcomes (IRISEO) were also used. The LORI is based on the Accessibility/Remoteness index of Australia and is a measure of remoteness that reflects distance to service centers. The LORI index is an area level indicator and it ranges from 1 to 5, from "no isolation", which corresponds to metropolitan areas, to "extreme isolation" [27]. The IRISEO is calculated specifically for Aboriginal Australians and is an area-level measure of community socioeconomic disadvantage based on education, employment, income, and housing. It ranges from (1) disadvantaged to (10) advantaged [28].

Strengths and difficulties questionnaire (SDQ). Child social and emotional outcomes were assessed by the caregiver's version of the SDQ. The instrument is validated for use among 4 to 17 years old [29]. The SQD has been recently validated for Aboriginal children of this age range (4–17 years), displaying good psychometric properties and excellent overall reliability [30]. It assesses levels of emotional and behavioral difficulties in four domains: emotional difficulties, conduct problems, hyperactivity, and peer problems. Each domain is composed of five items with responses ranging from 0 "Not true" to 2 "Certainly true". Examples of items are "often unhappy, depressed, or tearful" (emotional difficulties), "steals from home, school, or elsewhere" (conduct problems) and "restless, overactive, cannot stay sill for long" (hyperactivity). A score-range from 0 to 10 is obtained for each domain. A total score for emotional and behavioral difficulties is computed by summing the scores on the four domains (0–40). Higher scores indicate higher levels of difficulties that might represent risk for future clinical symptomatology [29].

Statistical analysis

The first step of the analysis was a Confirmatory Factor Analysis (CFA) to evaluate the fit of the hypothesized one-factor model. The estimation method was weighted least squares (WLSMV) with mean and variance adjusted test statistic [31]. WLSMV estimation is recommended for use with non-normal distributions [32], such as the four *ordinal* ERI items, and skewed data [33]. Considering that the percentage of missing data in individual items was below 1%, multiple imputation would not be likely to change the results and listwise deletion was employed. Furthermore, WLSMV estimation with listwise deletion can be used when the amount of missing data is unsubstantial, producing unbiased estimates for the parameters and their standard errors [34].

The sample size used (n = 424) was considered adequate for our analytical purposes. In general, there are two guidelines for sample size requirements in CFA models: (1) the absolute sample size (N), in which N ^{3} 300 guarantees accurate parameters and fit statistics in WLSMV estimation [35]; and the relative sample size to number of estimated parameters (q), namely the N:q ratio, which should have a value above 10:1 [36]. In our study, considering that the most complex model had 25 estimated parameters (q = 25), the sample size requirements were achieved both in an absolute (n = 435) and relative (N:q = 17.4) sense. Model fit was evaluated with the scaled χ^2, in addition to the scaled Comparative Fit Index (CFI) and the scaled Root Mean Square Error of Approximation (RMSEA). Values of CFI $\geq$ 0.96 and RMSEA $\leq$ 0.5 indicated good fit [37]. Values of RMSEA > 1.0 were considered to be indicative of poor fit [38] and the hypothesis of close-fit (RMSEA $\leq$ 0.5) was evaluated [39].

In case of a poor fitting model, model re-specifications were conducted by the evaluation of standardized residual correlations, modification indices (MI) and standardized expected parameter change (SEPC) [40]. After a model was established, we proceeded to evaluate measurement invariance by gender to check whether the items functioned differently between boys and girls. Testing invariance by gender intends to account for possible differences among boys and girls regarding transmission of cultural practices and racial socialization, which can influence children's attitudes towards ERI [41, 42]. Configural, metric and scalar invariance were evaluated with $\chi 2$ [43]. In the event $\chi 2$ was statistically significant, the ΔCFI [44] was used, with invariance being assumed when the CFI values do not vary above 0.002 points between models. Finally, reliability was evaluated with the ordinal $\boldsymbol{\alpha}$ [45]. The use of the ordinal $\boldsymbol{\alpha}$ is required since Cronbach's $\boldsymbol{\alpha}$ [46] underestimates reliability in ordinal items, such as Likert scales [45]. Reliability above 0.80 is usually deemed acceptable for validation studies such as ours [47]. Analyses were conducted in R software [48], R packages lavaan 0.6–2 [49] and semTools [50].

For the criterion validity analysis, the association between ERI affirmation and child total emotional and behavioural difficulties was tested. Our hypothesis was that children with high ERI affirmation would be at decreased risk for the onset of emotional and behavioural difficulties, as per the associations of ERI and positive developmental outcomes among ethnic-racial minority children [51]. Generalized linear models were preferred as the specification of the link function allows accommodation of non-normal distributions and skewed data [52]. We estimated risk ratios as a measure of the effect by testing a generalized linear model with a log-Poisson link and robust errors (model 2). The log Poisson link was chosen as we aimed for risk ratios as effect-measures and robust errors were specified to generate unbiased effect estimates in case of model misspecification [53]. The exposure and outcome variables were dichotomized. ERI affirmation was divided into "high" and "low". The high ERI affirmation category was composed of the children who endorsed "Yes (Always)" and "Yes (Most of the time)" to all four items of ERI affirmation (≥20). The SDQ total score was dichotomized in "high difficulties" (scores ≥17), and "low difficulties" (scores ≤ [illegible] 54]. The two models were adjusted for child age and sex, and family LORI, and IRISEO, confounding selected as per associations reported in the literature among these sociodemographics and both ERI and wellbeing [42, 55, 56]. The models were tested with 419 children, as five children had no information on the SDQ total score, LORI and IRISEO variables.

Results

The first model tested was the one-factor model and the fit indices provided mixed evidence regarding model fit (Table 1). The sample of the fitted model was 424 participants. Although the CFI was above the threshold of 0.96, a statistically significant $\chi 2$ and a RMSEA of 0.128 were observed. Additionally, the p-value of 0.01 indicates that the hypothesis of close-fit (RMSEA<0.5) was rejected. We explored possible adjustments to improve the model by carrying specification searches. The examination of the standardized expected parameter changes (SEPC) showed that items 1 and 3 residuals had a correlation of 0.65. Therefore, we observed that correlated uniqueness between two items could be limiting the fit of the data to the confirmatory structure tested.

We proceeded to test a second model that accounted for correlated uniqueness between items 1 and 3. The model had an excellent fit to the data (χ^2 (1) = 0.121, p = 0.72). The RMSEA value of 0.000–90% CI [0.000, 0.091] for this second model suggested that the covariance of item responses was sufficiently explained by the underlying one-factor model specified. The 90% CI shows that the range of compatible values with the model are mostly below the threshold of 0.05, although values above it could also be compatible. Nonetheless, testing for the alternative hypothesis that the RMSEA value falls below the value of 0.05 resulted in a p-value of 0.83, suggesting that the hypothesis of close-fit should not be rejected. The CFI value was above the 0.96 threshold, suggesting excellent model fit. The interpretation of the fit indices and the available CI and significance tests suggest there is evidence of construct validity for this one-factor model of ERI affirmation. Comparisons between the fit indices from the two

Table 1. Fit indices for the two unidimensional models of ERI affirmation.

Model	χ^2	*df*	*p*-value	RMSEA	90% CI	p-close	CFI
Model 1	15.89	2	0.000	0.128	0.075–0.190	0.01	0.983
Model 2	0.121	1	0.72	0.000	0.000–0.091	0.837	1.000

https://doi.org/10.1371/journal.pone.0224736.t001

Table 2. Item loading estimates of the two unidimensional models of ERI affirmation.

Items	Model 1		Model 2	
	Estimate	95% CI	Estimate	95% CI
1."I feel good about being Aboriginal and/or Torres Strait Islander in class"	0.77	(0.70, 0.85)	0.66	(0.56, 0.76)
2."I want to share (tell others) things about being Aboriginal and/or Torres Strait Islander in class"	0.63	(0.56, 0.73)	0.65	(0.57, 0.73)
3."I feel safe about being Aboriginal and/or Torres Strait Islander in class"	0.82	(0.74. 0.88)	0.72	(0.64, 0.81)
4."I like people to know I am Aboriginal and/or Torres Strait Islander in class"	0.79	(0.73, 0.87)	0.86	(0.79, 0.94)

https://doi.org/10.1371/journal.pone.0224736.t002

CFA models tested are showed in Table 1. Table 2 presents the second model items' loadings on the underlying ERI affirmation factor.

The unidmensional structural model of ethnic-racial identity affirmation is illustrated in Fig 1, accounting for correlated uniqueness between items 1 and 3. Internal consistency reliability for the measure was considered adequate, as was the ordinal α of 0.83. The Ω hierarchical α, a reliability index that accounts for correlated uniqueness among items, was also assessed and its value of 0.72 was considered satisfactory [57]. Finally, the analysis of measurement invariance indicated scalar invariance ($\Delta\chi^2$ (4) = 8.86, p = 0.78; ΔCFI = 0.000), demonstrating that the unidimensional model is invariant among boys and girls (Table 3).

The next step of the analysis was the criterion validity. The results of the generalized linear model (n = 419) confirmed the hypothesis tested. Results showed that children with high ERI affirmation had a 9% decreased risk for presenting high SDQ scores (RR_a = 0.91, 95% CI 0.64, 1.29). The CI, however, showed that values above 1 could be compatible with the model, affecting the precision of the effect-estimate. Nonetheless, the results indicated that ERI affirmation had a protective effect for the onset of emotional and behavioural difficulties over and above levels of age, sex, geographical location, and socio-economic status. This suggests evidence of criterion validity of the ERI affirmation measure tested. Table 4 includes the frequency distribution of the exposure, outcomes, and confounding variables included in the analysis.

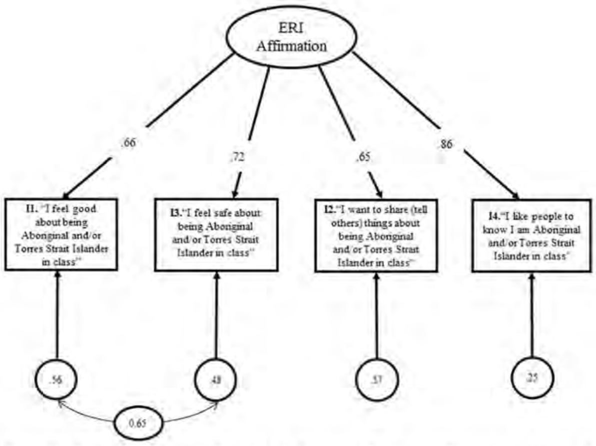


Fig 1. Unidmensional structural model of ethnic-racial identity affirmation accouting for correlated uniqueness.

https://doi.org/10.1371/journal.pone.0224736.g001

Table 3. Fit statistics for measurement invariance according to gender.

Model	χ^2	*df*	*p*-value	RMSEA	90% CI	CFI	$\Delta\chi2$ (df)	*p*-value	Δ CFI
Configural	0.25	2	0.88	0.000	[0.000, 0.066]	1.000	-	-	-
Metric	2.16	5	0.83	0.000	[0.000, 0.057]	1.000	2.10 (3)	0.55	0.000
Scalar	8.86	13	0.78	0.000	[0.000, 0.046]	1.000	8.26 (8)	0.41	0.000

Note. $\chi2$ = chi-square; df = degrees of freedom; RMSEA = root mean square error of approximation; CI = confidence interval; CFI = comparative fit index; $\Delta\chi2$ (df) = chi-square difference and degrees of freedom; Δ CFI = CFI difference. The $\chi2$ column reports scaled $\chi2$. $\Delta\chi2$ (df) is a function of standard (not the scaled) $\chi2$ statistics.

https://doi.org/10.1371/journal.pone.0224736.t003

Discussion

The CFA analysis provided evidence of construct validity that the brief measure of ERI affirmation works as a unidimensional scale among Australian Aboriginal children aged 10–12 years. The ordinal **α** and Ω hierarchical provided evidence that internal consistency reliability was adequate [47, 57]. In addition, the results of the generalized linear model tested contributes to evidence of criterion validity. The association between ERI affirmation and children's emotional and behavioural difficulties reflect literature on the protective effect of positive attitudes towards ERI on the wellbeing of ethnic-racial minorities, including Indigenous youth from the U.S., Canada, and New Zealand [5, 58–60]. It also reflects Aboriginal Australians' perspective on the importance of pride (positive attitudes) over ERI for positive Aboriginal children's health and development [14, 18]. Such results contributes to the necessary evidence for research based on ERI data from LSIC, as it demonstrates that ERI affirmation is being assessed with a valid and reliable measure.

Table 4. Participant's characteristics (n = 419).

Characteristic	Prevalence (95%CI)	n =
Child Age (years)		
10	47.0 (42.7, 52.3)	199
11	52.5 (47.7, 57.2)	220
Mean(SD)	10.5 (0.5)	
Gender		
Male	51.3 (46.5, 56.0)	215
Female	48.7 (44.0, 53.5)	204
ERI affirmation		
High ERI affirmation (≥20)	49.4 (44.6, 54.2)	207
Low ERI affirmation (<20)	50.5 (45.8, 55.4)	212
Mean (SD)	20.6 (3.6)	
Emotional and Behavioural difficulties		
Low difficulties (<17)	76.8 (72.5, 80.6)	322
High difficulties (≥17)	23.1 (19.3, 27.4)	97
Mean (SD)	10.4 (5.9)	
Level of Relative Isolation (LORI)		
None	28.9 (24.7, 32.4)	121
Low	55.1 (50.3, 59.8)	231
Moderate	7.6 (5.4, 10.6)	32
High/Extreme	8.3 (6.0, 11.4)	35
Indigenous Index of Socioeconomic Outcomes (IRISEO)		
Mean (SD)	5.8 (2.1)	

https://doi.org/10.1371/journal.pone.0224736.t004

The initial model tested was a unidimensional model and the evidence regarding model fit was mixed. The results indicated strong correlated uniqueness between items 1 and 3. MacCallum, Roznowski [61] have discussed additional parameters due to specification searches, such as correlated uniqueness, and recommend that these parameters should be included only when justified by the theoretical background of the construct [62]. In our study, such a theoretical justification exists. The strong correlation between items 1 and 3 reflects the association between positive in-group attitudes ("I feel good about being Aboriginal and/or Torres Strait Islander in class") and the levels of cultural safety perceived by respondents ("I feel safe about being Aboriginal or Torres-Strait Islander in class"). The perception of one's social environment as accepting of cultural diversity might be linked, for example, to reduced experiences of racial discrimination and more positive experiences of ERI expression. Promoting cultural safety features as a key factor in improving Aboriginal health, education, and community wellness [63, 64]. For example, perceptions of cultural respect, peer acceptance of ERI, community involvement, and teacher's cultural sensitivity–all contribute to a culturally safe environment—and are associated with less school absenteeism, higher classroom participation, and importance placed on school among Aboriginal students [26, 65].

After the inclusion of the correlated uniqueness between items 1 and 3, the model was correctly specified and achieved excellent fit. Therefore, the final measurement model was a unidimensional model *with* correlated uniqueness between items 1 and 3. The unidimensionality of the ERI items (i.e. the four items constitute a one-factor model) after the inclusion of the correlated uniqueness indicates that, although feelings of safety and positive experiences regarding ERI are more highly correlated with each other than others aspects of the construct (e.g., "I want to share (tell others) things about being Aboriginal and/or Torres Strait Islander in class"), the four items measure a single construct. That is, although the four items measure four distinct attitudes towards ERI ((1) feeling good about being Aboriginal and Torres Strait Islander; (2) wanting to share things about being Aboriginal and Torres Strait Islander; (3) feeling safe about being Aboriginal and Torres Strait Islander; and (4) liking people to know that they are Aboriginal and Torres Strait Islander), the CFA indicated that these attitudes constitute the broader construct of ERI affirmation. These findings are consistent with previous psychometric studies of ERI measures showing ERI affirmation as a distinct construct that encompasses several attitudes towards ERI (e.g. "I feel negatively about my ethnicity") [8]. Finally, one practical implication of the four items measuring a common underlying construct is that item scores can be summated to create a total score [22] and this total score provides a measure of ERI affirmation.

To the best of our knowledge, there is just one validated scale to assess Aboriginal Australian children's ERI [66]. The scale, however, focuses on exploration of cultural practices (knowledge of Aboriginal culture) and salience of racial identity, with no specific assessment of attitudes towards ERI [66]. The 40-item-length of the scale might also limit its applicability to large-scale studies such as LSIC. Here we provide evidence for a measure that's strength resides in its brevity and specificity of content. This permits its inclusion in surveys desiring a holistic and comprehensive perspective of Aboriginal children's wellbeing. The specificity of the measure reflects the debate on the importance of clearly defining which ERI component is being assessed [1]. Such accuracy might permit researchers to investigate how affective components of ERI relate, for example, to exploration of cultural practices and levels of commitment to one's ERI (ERI processes) later in development [1, 8, 67, 68]. These distinctions might shed light onto how ERI develops among Aboriginal children and how the interplay between ERI processes and content relates to racial socialization processes and discrimination [67, 69].

Finally, investing on promotion of positive development from early age might assist in reducing health inequalities among Aboriginal children and youth and their non-Indigenous counterparts [16]. Due to its centrality to positive development and wellbeing, promotion of positive ERI attitudes might protect Aboriginal children against adversity and increase wellbeing [58, 70]. Research on evaluation of programs whose purpose is to increase the social and emotional wellbeing of Aboriginal children is still limited [71]. However, there is evidence of the efficacy of school-based interventions designed to increase affirmative ERI among other ethnic-minority youth (e.g., Latin, African, and Native-Americans), with reported effects on wellbeing and learning outcomes [72, 73]. Valid and reliable measurement of ERI can assist at baseline measurement and monitoring of outcomes for future interventions targeting cultural socialization and promotion of ERI in the Aboriginal Australian context [73].

We conclude that the LSIC items tested work as a brief measure of ERI affirmation by providing evidence of construct and criterion validity on a sample of 424 Aboriginal children aged 10 to 12 years. We recognise the limitation of not using a representative sample of the Aboriginal Australian children population. Nonetheless, the sample size used was considered sufficient for the analytical purposes of this study [35]. It is also noticeable that the LSIC is possibly the largest currently available source of information on ERI and other determinants of this population health and wellbeing [20]. The LSIC team has been committed to involving community stakeholders and field specialists in the selection of content and data collection procedures. The ERI items used were based on previous work on Aboriginal perspectives of wellbeing, which further contributes to the content validity of the measure. This is one of the few empirical demonstrations of the psychometric properties of a measure assessing components of Aboriginal children ERI. As such, it contributes to the development of this area of research in the Aboriginal Australian context.

Author Contributions

Conceptualization: Davi Manzini Macedo, Lisa M. Jamieson.

Formal analysis: Davi Manzini Macedo, Pedro Ribeiro Santiago.

Methodology: Davi Manzini Macedo, Pedro Ribeiro Santiago.

Software: Pedro Ribeiro Santiago.

Supervision: Rachel M. Roberts, Lisa G. Smithers, Lisa M. Jamieson.

Visualization: Rachel M. Roberts, Lisa G. Smithers, Yin Paradies, Lisa M. Jamieson.

Writing – original draft: Davi Manzini Macedo, Pedro Ribeiro Santiago.

Writing – review & editing: Davi Manzini Macedo, Rachel M. Roberts, Lisa G. Smithers, Yin Paradies, Lisa M. Jamieson.

References

1. Schwartz SJ, Syed M, Yip T, Knight GP, Umana-Taylor AJ, Rivas-Drake D, et al. Methodological issues in ethnic and racial identity research with ethnic minority populations: theoretical precision, measurement issues, and research designs. Child Dev. 2014; 85(1):58–76. https://doi.org/10.1111/cdev.12201 PMID: 24490892
2. Corenblum B. Development of racial-ethnic identity among First Nation children. J Youth Adolesc. 2014; 43(3):356–74. https://doi.org/10.1007/s10964-013-0007-5 PMID: 23982921
3. Serrano-Villar M, Calzada EJ. Ethnic identity: evidence of protective effects for young, Latino children. J Appl Dev Psychol. 2016; 42:21–30. https://doi.org/10.1016/j.appdev.2015.11.002 PMID: 26778873
4. Derlan CL, Umana-Taylor AJ, Updegraff KA, Jahromi LB. Longitudinal relations among Mexican-origin mothers' cultural characteristics, cultural socialization, and 5-year-old children's ethnic-racial identification. Dev Psychol. 2017; 53(11):2078–91. https://doi.org/10.1037/dev0000386 PMID: 29094971

5. Corenblum B. Relationships between racial-ethnic identity, self-esteem and in-group attitudes among First Nation children. J Youth Adolesc. 2014; 43(3):387–404. https://doi.org/10.1007/s10964-013-0081-8 PMID: 24343260

6. Marcelo AK, Yates TM. Young children's ethnic-racial identity moderates the impact of early discrimination experiences on child behavior problems. Cultur Divers Ethnic Minor Psychol. 2019; 25(2):253–65. https://doi.org/10.1037/cdp0000220 PMID: 30058832

7. Romero AJ, Edwards LM, Fryberg SA, Orduña M. Resilience to discrimination stress across ethnic identity stages of development. J Appl Soc Psychol. 2014; 44(1):1–11. https://doi.org/10.1111/jasp.12192

8. Umaña-Taylor AJ, Yazedjian A, Bámaca-Gómez M. Developing the Ethnic Identity Scale using Eriksonian and Social Identity perspectives. Identity. 2004; 4(1):9–38. https://doi.org/10.1207/s1532706xid0401_2

9. Tajfel H, C. Turner J. The Social Identity Theory of intergroup behavior. In: Worchel S, Austin WG, editors. Psychology of intergroup relations. Chicago: Nelson-Hall; 1986. pp. 7–24.

10. Ashmore RD, Kay D, McLaughlin-Volpe T. An organizing framework for collective identity: articulation and significance of multidimensionality. Psychol Bull. 2004; 130(1):80–114. https://doi.org/10.1037/0033-2909.130.1.80 PMID: 14717651

11. Australia Institute of Health and Welfare. The health and welfare of Australia's Aboriginal and Torres Istrait Islander peoples 2015. Canberra: Australia Institute of Health and Welfare; 2015. 200 p. Report No. IHW 147

12. Australian Bureau of Statistics [Internet]. Estimates of Aboriginal and Torres Strait Islander Australians [cited 2019 September 25]. Available from: https://www.abs.gov.au/ausstats/abs@.nsf/mf/3238.0.55.001.

13. Zubrick SR, Shepherd C, Dudgeon P, Gee G, Paradies Y, Scrine C, et al. Social determinants of social and emotional wellbeing for Aboriginal Australians. In: Dudgeon P, Milroy H, Walker R, editors. Working together: Aboriginal and Torres Strait Islander mental health and wellbeing principles and practice. Canberra: Australian Government, Department of The Prime Minister and Cabinet; 2014. pp. 93–112.

14. Priest N, Thompson L, Mackean T, Baker A, Waters E. 'Yarning up with Koori kids'—hearing the voices of Australian urban Indigenous children about their health and well-being. Ethn Health. 2017; 22(6):631–47. https://doi.org/10.1080/13557858.2016.1246418 PMID: 27764969

15. Priest N, Paradies Y, Stewart P, Luke J. Racism and health among urban Aboriginal young people. BMC Public Health. 2011; 11(568):1–9. https://doi.org/10.1186/1471-2458-11-568 PMID: 21756369

16. Australian Institute of Health and Welfare. Aboriginal and Torres Strait Islander adolescent and youth health and wellbeing. Canberra: Australian Institute of Health and Welfare; 2018. 38 p. Report No. IHW 198.

17. Tonkinson R. Australian Aboriginal society and culture: an overview. International Encyclopedia of the Social & Behavioral Sciences. 2nd ed: Elsevier; 2015. p. 234–8.

18. Priest N, Mackean T, Davis E, Briggs L, Waters E. Aboriginal perspectives of child health and wellbeing in an urban setting: developing a conceptual framework. Health Soc Rev. 2014; 21(2):180–95. https://doi.org/10.5172/hesr.2012.21.2.180

19. Thurber KA, Banks E, Banwell C, Team L. Cohort profile: Footprints in Time, the Australian Longitudinal Study of Indigenous Children. Int J Epidemiol. 2015; 44(3):789–800. https://doi.org/10.1093/ije/dyu122 PMID: 25011454

20. Footprints in Time: The Longitudinal Study of Indigenous Children—Data user guide, Release 9.0. Department of Social Services, 2018.

21. Gessaroli ME, Champlain AF. Test Dimensionality: Assessment of. Wiley StatsRef: Statistics Reference Online: John Wiley & Sons, Ltd; 2014.

22. Gardner PL. Measuring attitudes to science: Unidimensionality and internal consistency revisited. Res Sci Educ. 1995; 25(3):283–9.

23. Footprints in Time—The Longitudinal Study of Indigenous Children [Internet]. Department of Social Services; 2019 [2019 May 14]. Available from: https://www.dss.gov.au/about-the-department/publications-articles/research-publications/longitudinal-data-initiatives/footprints-in-time-the-longitudinal-study-of-indigenous-children-lsic#1.

24. Payments and support for Indigenous Australians [Internet]. Australian Government Department of Human Services; 2019 [14 May 2019]. Available from: https://www.humanservices.gov.au/individuals/subjects/payments-and-support-for-indigenous-australians.

25. Medicare services for Indigenous Australians [Internet]. Department of Human Services; 2019 [14 May 2019]. Available from: https://www.humanservices.gov.au/individuals/subjects/whats-covered-medicare/medicare-services-indigenous-australians.

26. Craven RG, Alex Y, Munns G, Bodkin-Andrews G, Nida D, O'Rourke V. Seeding success for Aboriginal primary students. Centre for Positive Psychology and Education and NSW Aboriginal Education Consultative Group, 2016.
27. Zubrick SR, Lawrence DM, Silburn SR, Blair E, Milroy H, Wilkes T, et al. The Western Australian Aboriginal Child Health Study: the health of Aboriginal children and young people. Perth: Telethon Institute for Child Health Research, 2004.
28. Biddle N. Ranking regions: revisiting an Index of Relative Indigenous Socioeconomic Outcomes. Canberra: Centre for Aboriginal Economic Policy Research, 2009.
29. Goodman R. The Strengths and Difficulties Questionnaire: a research note. The Journal of Child Psychology and Psychiatry. 1997; 38:581–6. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x PMID: 9255702
30. Williamson A, McElduff P, Dadds M, D'Este C, Redman S, Raphael B, et al. The construct validity of the S trengths and D ifficulties Questionnaire for A boriginal children living in urban New South Wales, Australia. Aust Psychol. 2014; 49(3):163–70. https://doi.org/10.1111/ap.12045
31. Asparouhov T, Muthén B. Simple second order chi-square correction. Mplus technical appendix. 2010.
32. Beauducel A, Herzberg PY. On the performance of Maximum Likelihood versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. Structural Equation Modeling: A Multidisciplinary Journal. 2006; 13(2):186–203. https://doi.org/10.1207/s15328007sem1302_2
33. Muthén B, du Toit S, Spisic D. Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. 1997.
34. Asparouhov T, Muthén B. Weighted least squares estimation with missing data. Mplus Technical Appendix. 2010; 2010:1–10.
35. Moshagen M, Musch J. Sample size requirements of the robust weighted least squares estimator. Meth Eur J Res Meth Behav Soc Sci. 2014; 10(2):60–70. https://doi.org/10.1027/1614-2241/a000068
36. Jackson DL. Revisiting sample size and number of parameter estimates: Some support for the n:q hypothesis. Struct Equ modeling. 2003; 10(1):128–41. https://doi.org/10.1207/s15328007sem1001_6
37. Y Yu C. Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. Los Angeles: University of California; 2002.
38. Kline RB. Principles and practice of structural equation modeling. New York, NY: Guilford publications; 2015.
39. Hayduk LA, Pazderka-Robinson H, Cummings GG, Levers M-JD, Beres MA. Structural equation model testing and the quality of natural killer cell activity measurements. BMC Med Res Methodol. 2005; 5 (1):1.
40. Saris WE, Satorra A, Van der Veld WM. Testing structural equation models or detection of misspecifications? Structural Equation Modeling. 2009; 16(4):561–82.
41. Lohoar S, Butera N, Kennedy E. Strength of Australian Aboriginal cultural practices in family life and child rearing. Canberra: Australian Institute of Families Studies. 2014. 20p. Paper No. 25.
42. Rogers LO, Meltzoff AN. Is gender more important and meaningful than race? An analysis of racial and gender identity among Black, White, and mixed-race children. Cultur Divers Ethnic Minor Psychol. 2017; 23(3):323–34. https://doi.org/10.1037/cdp0000125 PMID: 27736104
43. Satorra A. Scaled and adjusted restricted tests in multi-sample analysis of moment structures. Innovations in multivariate statistical analysis: Springer; 2000. pp. 233–47.
44. Meade A, C Johnson E, Braddy P. Power and sensitivity of alternative fit indices in tests of measurement invariance. J Appl Psychol. 2008; 93(3):568–92. https://doi.org/10.1037/0021-9010.93.3.568 PMID: 18457487
45. Zumbo BD, Gadermann AM, Zeisser C. Ordinal versions of coefficients alpha and theta for Likert rating scales. Journal of modern applied statistical methods. 2007; 6(1):4.
46. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951; 16(3):297–334.
47. Wells CS, Wollack JA. An instructor's guide to understanding test reliability. Testing & Evaluation Services University of Wisconsin. 2003.
48. R Core Team. R: A language and environment for statistical computing. 2013.
49. Rosseel Y. Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). Journal of statistical software. 2012; 48(2):1–36.
50. Jorgensen T, Pornprasertmanit S, Schoemann A, Rosseel Y. semTools: Useful tools for structural equation modeling. R Package Version 0.5–1. 2018.

51. Priest N, Paradies Y, Trenerry B, Truong M, Karlsen S, Kelly Y. A systematic review of studies examining the relationship between reported racism and health and wellbeing for children and young people. Soc Sci Med. 2013; 95:115–27. https://doi.org/10.1016/j.socscimed.2012.11.031 PMID: 23312306

52. Fahrmeir L, Tutz G. Modelling and analysis of cross-sectional data: a review of univariate. Multivariate statistical modelling based on generalized linear models. 2nd ed. New York: Springer-Verlag New York; 2001. pp. 15–68.

53. Chen W, Qian L, Shi J, Franklin M. Comparing performance between log-binomial and robust Poisson regression models for estimating risk ratios under model misspecification. BMC Med Res Methodol. 2018; 18(1):63. https://doi.org/10.1186/s12874-018-0519-5 PMID: 29929477

54. Meltzer H, Gatward R, Goodman R, Ford T. Mental health of children and adolescents in Great Britain. Int Rev Psychiatry. 2003; 15(1–2):185–7. https://doi.org/10.1080/0954026021000046155 PMID: 12745331

55. Shepherd C, Li J, Mitrou F, Zubrick SR. Socioeconomic disparities in the mental health of Indigenous children in Western Australia. BMC Public Health. 2012; 12:756. https://doi.org/10.1186/1471-2458-12-756 PMID: 22958495

56. Cooper SM, Banerjee M, Griffin CB, Glover CS, Shaheed J, McBride M, et al. African American fathers' race-related discussions with their children: Associations with racial identity beliefs and sociodemographic factors. J Child Fam Stud. 2019; 28(10):2800–12. https://doi.org/10.1007/s10826-019-01460-2

57. Dunn TJ, Baguley T, Brunsden V. From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. Br J Psychol. 2014; 105(3):399–412. https://doi.org/10.1111/bjop.12046 PMID: 24844115

58. Snowshoe A, Crooks C, F. Tremblay P, E. Hinson R. Cultural connectedness and its relation to mental wellness for First Nations youth. J Prim Prev. 2017; 38:67–86. https://doi.org/10.1007/s10935-016-0454-3 PMID: 27807659

59. Smokowski PR, Evans CB, Cotter KL, Webber KC. Ethnic identity and mental health in American Indian youth: examining mediation pathways through self-esteem, and future optimism. J Youth Adolesc. 2014; 43(3):343–55. https://doi.org/10.1007/s10964-013-9992-7 PMID: 23929530

60. Williams AD, Clark TC, Lewycka S. The associations between cultural identity and mental health outcomes for Indigenous Maori youth in New Zealand. Front Public Health. 2018; 6:319. https://doi.org/10.3389/fpubh.2018.00319 PMID: 30483490

61. MacCallum RC, Roznowski M, Necowitz LB. Model modifications in covariance structure analysis: The problem of capitalization on chance. Psychol Bull. 1992; 111(3):490. https://doi.org/10.1037/0033-2909.111.3.490 PMID: 16250105

62. Brown TA. Confirmatory factor analysis for applied research. New York, NY: Guilford Publications; 2014.

63. Brascoupé S, Waters C. Cultural Safety: exploring the applicability of the concept of cultural safety to Aboriginal health and community wellness. Journal of Aboriginal Health. 2009; 5(2):6–41. https://doi.org/doi.org/10.3138/ijih.v5i2.28981

64. Rigby W, Duffy E, Manners J, Latham H, Lyons L, Crawford L, et al. Closing the gap: cultural safety in Indigenous health education. Contemp Nurse. 2010; 37(1):21–30. https://doi.org/10.5172/conu.2011.37.1.021 PMID: 21591823

65. Bodkin-Andrews G, O'Rourke V, Grant R, Denson N, Craven RG. Validating racism and cultural respect: testing the psychometric properties and educational impact of perceived discrimination and multiculturation for Indigenous and non-Indigenous students. Educational Research and Evaluation. 2011; 16(6):471–93. https://doi.org/10.1080/13803611.2010.550497

66. Kickett-Tucker CS, Christensen D, Lawrence D, Zubrick SR, Johnson DJ, Stanley F. Development and validation of the Australian Aboriginal racial identity and self-esteem survey for 8–12 year old children (IRISE_C). Int J Equity Health. 2015; 14:103. https://doi.org/10.1186/s12939-015-0234-3 PMID: 26499852

67. Umana-Taylor AJ, Quintana SM, Lee RM, Cross WE Jr., Rivas-Drake D, Schwartz SJ, et al. Ethnic and racial identity during adolescence and into young adulthood: an integrated conceptualization. Child Dev. 2014; 85(1):21–39. https://doi.org/10.1111/cdev.12196 PMID: 24490890

68. Phinney JS. Ethnic identity in adolescents and adults: review of research. Psychol Bull. 1990; 108(3):499–514. https://doi.org/10.1037/0033-2909.108.3.499 PMID: 2270238

69. Hughes DL, Watford JA, Del Toro J. A transactional/ecological perspective on ethnic-racial identity, socialization, and discrimination. In: SS H, MD H, LS L, editors. Equity and justice in developmental science: implications for young people, families, and communities: Academic Press; 2016. pp. 1–41.

70. Macedo DM, Smithers LG, Roberts RM, Haag DG, Paradies Y, Jamieson LM. Does ethnic-racial identity modify the effects of racism on the social and emotional wellbeing of Aboriginal Australian children? PLoS One. 2019; 14(8):e0220744. https://doi.org/10.1371/journal.pone.0220744 PMID: 31390371

71. Australian Institute of Health and Welfare. Strategies and practices for promoting the social and emotional wellbeing of Aboriginal and Torres Strait Islander people. Canberra: Australian Institute of Health and Welfare; 2013. 15 p. Report No. IHW 82.

72. Zamora AL, Curtis H, Lancaster L. Promoting racial and ethnic identity: A school-based intervention to support Latino youth. Journal of Latinos and Education. 2017; 18(3):215–27. https://doi.org/10.1080/15348431.2017.1386104

73. Umana-Taylor AJ, Kornienko O, Douglass Bayless S, Updegraff KA. A universal intervention program increases ethnic-racial identity exploration and resolution to predict adolescent psychosocial functioning one year later. J Youth Adolesc. 2018; 47(1):1–15. https://doi.org/10.1007/s10964-017-0766-5 PMID: 29030792

Appendix 2 – Development and initial psychometric assessment of the Race-related Attitudes and Multiculturalism Scale in Australia

Development and initial psychometric assessment of the Race-related Attitudes and Multiculturalism Scale in Australia

Pedro Henrique Ribeiro Santiago[1&], Dandara Gabriela Haag[1&], Davi Manzini Macedo[1], Joao Luiz Dornelles Bastos[2], Yin Paradies[3], Lisa Jamieson[1]

1. Australian Research Centre for Population Oral Health (ARCPOH), Adelaide Dental School, University of Adelaide.

2. Department of Public Health, Federal University of Santa Catarina, Brazil

3 Alfred Deakin Institute for Citizenship and Globalisation, Deakin University, Australia

*Corresponding author

E-mail: dandara.haag@adelaide.edu.au

[&]These authors contributed equally to this work.

Abstract

Aim: The present study aims to develop the Race-related Attitudes and Multiculturalism Scale (RRAMS) as a measure of multiculturalist attitudes, as well as to assess its psychometric properties in a national sample of Australian adults.

Methods: The sample comprised 2,714 Australian adults who took part in the 2013 National Dental Telephone Interview Survey (NDTIS), which includes a telephone-based interview and a follow-up postal questionnaire. We employed Exploratory Factor Analysis to evaluate the RRAMS' factorial structure (n=271) and then proceeded with Confirmatory Factor Analysis to confirm the proposed structure in an independent sample (n=2,443). Measurement invariance was evaluated according to sex, age, and educational attainment. Criterion validity was assessed through known-groups comparisons. Internal consistency was assessed with McDonald's Ω_H and ordinal α. Multiple imputation by chained equations was adopted to handle missing data.

Results: EFA indicated a two-factor structure would best fit the data following the exclusion of 4 out of 12 items, which was then confirmed in an independent sample ($\chi^2(19) = 341.070$, $p<0.001$, CFI = 0.974, RMSEA = 0.083; 90% CI [0.076, 0.091]). Measurement invariance analyses indicated that the RRAMS items can be used to compare men/women, participants with/without tertiary education and young/older participants. The "Anglo-centric/Assimilationist attitudes" ($\Omega_H = 0.83$, $\alpha_{ORDINAL} = 0.85$) and "Inclusive/Pluralistic attitudes" subscales ($\Omega_H = 0.77$, $\alpha_{ORDINAL} = 0.79$) showed adequate reliability. Men and participants without tertiary education had higher Anglo-centric/assimilationist attitudes and lower inclusive/pluralistic attitudes, suggesting criterion-related validity.

Conclusions: The RRAMS appears to be a valid and reliable measure to evaluate multiculturalist attitudes in the Australian context. The instrument may be useful to

interventions aiming to promote multiculturalist inclusive attitudes and to increase social cohesion in Australia.

Introduction

Racism emerges whenever social and individual values, norms and practices of a given group are considered superior to others. Racism occurs with the particular aim of creating, maintaining or reinforcing power imbalances, as well as the corresponding inequalities in opportunities and resources along racial lines (1). Similar to most contemporary societies, Australia is characterized by co-existing expressions of cultural diversity on the one hand, and negative impacts of racism on social cohesion, on the other (1). In Australia, the mental health costs directly attributable to racism have been estimated at 235,452 disability-adjusted life years lost, which is equivalent to an average $37.9 billion in productivity loss per annum, equivalent to 3% of the Australian annual Gross Domestic Product (GDP) over 2001–2011 (2). Such a strong relationship is an indication that racism may erode the very social fabric of Australian society by producing mental disorders and suffering, which unevenly impacts upon racially marginalized groups.

Social conceptions that shape intergroup relations form the common ground upon which intergroup attitudes and discriminatory behaviour take place (3). On an empirical level, findings suggest that racist attitudes are associated with racist behaviours and racial-ethnic minorities' experiences of discrimination (4). Positive attitudes towards diversity, however, are negatively associated with discriminatory behaviour (5). In this study, we propose to explore attitudes in relation to multiculturalism, a construct of special relevance to the social, economic and political fabric of contemporary Australia (6). We focus on multiculturalism as an ideology of acknowledging and celebrating ethnic and cultural differences, in which the need for preserving cultural identities is recognized (7). It reflects a "sensibility and disposition towards cultural differences among large sections of the population"(8). Data from the 2016 Australian Census revealed that one in three Australians were born overseas, and a similar proportion of individuals speak a language other than English at home. Nevertheless, assimilationist attitudes – expectations of conformation to the dominant culture – often prevail, as opposed to multiculturalist perspectives that accept and praise racial and ethnic-cultural diversity (9). Understanding attitudes to multiculturalism can contribute to unveil the dynamics of racism and discrimination against minorities in the country, fostering public debate and policy formulation aimed to promote positive intergroup relations (10).

Research on ethnic-racial intergroup attitudes benefitted from the inputs of theories on ideological attitudes that explain group-based dominance and social cohesion (11-13). Social Dominance Orientation (SDO), for example, reflects the degree to which respondents believe that hierarchy-based dominance between social groups is natural (14). Discrimination against minorities, therefore, can be explained by the degree of endorsement of the notion that group-based hierarchies are natural and inevitable (14). Endorsement of group-based dominance and out-group prejudice tends to increase among those with highly identify with the dominant group, as they represent a mechanism of maintaining the in-group *status quo* (12).

Research on ethnic-racial intergroup relations in contemporary society has also explored the Right-wing Authoritarianism (RWA) concept (15-17). RWA is characterized by the endorsement of social conservative values, morality, collective security, group-based social cohesion, and strict obedience to social authorities (15, 17). Those who endorse RWA values can be more sensitive to threats to social stability, being prone to conservative values to increase their perception of control and collective security (18). Perception of threat has been shown to mediate the association between group identification and attitudes towards multiculturalism (11). Those that consider immigrants or ethnic-racial minorities as a threat to the control of resources or maintenance of the dominant social values tend to endorse more conservative/assimilationist attitudes towards multiculturalism (11, 19).

Sustaining the dominant group status quo can also be achieved by the avoidance of acknowledging and approaching ethnic-racial inequalities in the population. The so-called colour-blind racial ideology denies the existence of racism and justifies racial inequalities as a result of personal decisions, meritocratic achievements, and market forces (20, 21). By denying racist practices and racial inequalities, it provides the discursive tools to downplay policy proposals to promote racial justice and therefore maintains the power imbalance between ethnic-racial groups (20). Following this perspective, public denial of racism has been pointed as an obstacle to a deeper commitment to multiculturalism in Australia (13, 22). Although the existence of racism is acknowledged, most Australians fail to recognise the existence of Anglo-privilege, a step necessary in reducing the imbalance in resource distribution and political representation among ethnic-racial groups (13).

Taken together, the results mentioned above point to the centrality of properly assessing the different facets of intergroup attitudes towards multiculturalism to inform public debate and contribute to prevent and counteract discrimination. It is important to note that the majority of the available scales used to assess race-related attitudes have been developed and psychometrically examined among U.S. populations (7). These tools may not be relevant or provide valid/reliable estimates of race-related attitudes in non-US contexts, though, given the considerable contextual dependency of racism. Historiographic and sociological accounts of racial dynamics usually emphasize specificities in terms of colonization, past and contemporary immigration policies, and patterns of cultural diversity as key aspects.

Australia is a settler society that started with a policy of Anglo-celtic migration only, later expanded to include migrants from other European-backgrounds (e.g., Greeks, Italians), having only in the 1980’s opened its borders to migrants from Asian and Middle-Eastern descent. That and other differences in relation to other settler colonies (e.g., limited involvement on the Atlantic slavery trade) produce specific effects on social integration and justify the limitations of transposing tools validated in other populations to the Australian context. Just like other multiculturalist societies like Canada and New Zealand, multiculturalism was debated at a national level as a state-policy in the 1970’s. Backlashes from conservative sectors, nonetheless, contributed to prioritise an assimilationist perspective on the implementation of multiculturalism values in society. Australia has also historically dispossessed and oppressed the native Aboriginal Australians since the beginning of colonization with ongoing effects until present (23). Our study does not contemplate this agenda as to respect the pledge that the effects of colonisation and racism faced by Aboriginal Australians has unique features and can be diminished when contemplated under the umbrella of multiculturalism (24).

To the best of our knowledge, two measurement instruments that provide information on racial, ethnic, and cultural acceptance (i.e. race-related and multiculturalist attitudes) have been previously developed and assessed in Australia (7, 25). While the first has focused on intercultural understanding among teachers and students in schools (25), psychometric evaluation of the second was carried out in relatively young and convenience samples of primary and secondary school students (all younger than 15 years-old residing in Victoria) and community members (mean age of 23 years-old with 70% residing mainly in Victoria), which limits their applicability at a national level and among older age groups. Therefore, neither an integrated picture of attitudes towards multiculturalism across the country has yet been delineated, nor a range of strategies to advance racial equity based on this knowledge have been proposed.

The present study proposes the Race-related Attitudes and Multiculturalism Scale (RRAMS) as a measure of attitudes towards multiculturalism. The items were formulated to reflect social ideologies and collective beliefs identified to influence ethnic-racial intergroup attitudes. The aim of this study was to verify its applicability to the Australian context by assessing the extent to which the RRAMS provides valid and reliable measurement in a sample of Australian adults across all states and territories. In particular, the internal validity of the RRAMS was assessed in terms of its configural structure (i.e., the number of underlying factors), metric properties – the magnitude of item loadings and thresholds –, as well as measurement invariance (i.e., whether it allowed meaningful comparisons across sociodemographic characteristics). External validity of the RRAMS was then assessed in terms of its criterion-related validity.

Methods

Study design and participants

This was an Australian population-based study, with data obtained from the 2013 National Dental Telephone Interview Survey (NDTIS), which includes a telephone-based interview and a follow-up postal questionnaire. The NDTIS has been carried out periodically by the University of Adelaide since 1994, and comprises a large national sample of Australian residents aged 5 years and over. The NDTIS is a random sample survey that collects information on the dental health and use of dental services of Australians in all states and territories. The survey also collects data on social determinants of oral health and wellbeing, which include detailed information on sociodemographic factors, such as household income, education, country of birth, remoteness of location and main language spoken at home. For the 2013 survey, an overlapping dual sampling frame design was adopted. The first sampling frame was created from the electronic product ‘Australia on Disc 2012 Residential; an annually updated electronic listing of people/households listed in the White Pages across Australia. Both landline and mobile telephone numbers were provided on records where applicable.

A stratified two-stage sampling design was used to select a sample of people from this sampling frame. Records listed on the frame were stratified by state/territory and region, where region was defined as Capital City/Rest of State. A systematic sample of records was selected from each stratum using specified sampling fractions (26). To include households that were not listed in the White Pages, a second sampling frame comprising 20,000 randomly generated mobile telephone numbers was used. This sampling frame was supplied by *Sampleworx* and the mobile telephone numbers were created by appending randomly generated suffix numbers to all known Australian mobile prefix numbers. As the mobile numbers did not contain address information, the sampling frame could not be stratified by geographic region. A random sample of mobile numbers was selected from the frame and contacted to establish the main user of the mobile phone. This person was asked to participate in the telephone interview, provided that they were aged 18 years or over (26).

Following the completion of the telephone interview survey, participants were invited to respond to the postal questionnaire component. Those who agreed were sent a covering letter with the questionnaire and reply-paid envelope enclosed. A reminder postcard was sent two weeks later, with, if necessary, two additional follow-up letters/questionnaires sent subsequent to the postcard. A total of 6,340 Australian adults aged 18+ years took part in the 2013 NDTIS, with 2,935 (46.3%) completing the follow-up postal questionnaire. Sample characteristics are displayed in Table 1. Two-thirds of the sample were 45 to 98 years-old and had Technical and Further Education (TAFE) or went to university. Women corresponded to 60.3% of the sample. The majority of participants were born in Australia (76.7%), 12.8% were originally from Europe and 10.5% from the other continents (Asia, Africa, and Americas).

Table 1. Characteristics of study participants (n=2,714).

	Total sample		**EFA sample**		**CFA sample**	
Sample characteristics	**n**	**%**	**n**	**%**	**n**	**%**
Age						
18 to 45 years old	809	29.8	101	37.3	708	29.0
46 to 98 years old	1818	67.0	162	59.8	1656	67.8
Missing	87	3.2	8	3.0	79	3.2
Sex						
Female	1637	60.3	176	64.9	1461	59.8
Male	990	36.5	87	32.1	903	37.0
Missing	87	3.2	8	3.0	79	3.2
Education						
High school or less	548	20.2	60	22.1	1876	20.0
TAFE* or university	2079	76.6	203	74.9	488	76.8

Missing	87	3.2	8	3.0	79	3.2
Country of birth						
Australia	2079	76.7	209	77.1	1870	76.5
Rest of Oceania	72	2.7	6	2.2	66	2.7
Europe	347	12.8	36	13.3	311	12.7
Africa & Middle East	43	1.6	1	0.4	42	1.7
Asia	56	2.1	5	1.8	51	2.1
Americas	30	1.1	6	2.2	24	1.0
Missing	87	3.2	8	3.0	79	3.2

*TAFE, Technical and Further Education (trade school/college).

Ethical approval and consent

Ethical approval for the study was granted by the University of Adelaide's Human Research Ethics Committee (approval number HS-2013-036). All

Statistical Analysis

Statistical analyses were conducted with R software (27) and R packages lavaan (28), and semTools (29).

Phase 1: Item development

The RRAMS was developed by a group of researchers with expertise on the topics of racism, multiculturalism, and race-related attitudes in Australia. To ensure content validity (30) in the Australian context, the scale was based on large surveys carried out in the country that was co-designed by the abovementioned group of researchers. These include the 2015-16 Challenging Racism Project (31) and the 2013 survey of Victorians' attitudes to race and cultural diversity (32). The RRAMS was proposed as comprised of two subscales. The first subscale included six items reflecting theories and social ideologies in agreement with "Anglo-centric/Assimilationist attitudes" It included items measuring compliance to RWA (e.g, 'We need to stop spreading dangerous ideas and stick to the way things have always been done in Australia'), agreement with SDO ('It is okay if some racial or ethnic groups have better opportunities in life than others'), endorsement of colour-blind racial ideology (e.g., 'We shouldn't talk about racial or ethnic differences'), zero-sum racist thinking (e.g., 'Racial or ethnic minority groups take away jobs from other Australians'), and endorsement of assimilationist ideology (e.g., 'People from racial or ethnic minority groups should behave more like mainstream Australians').

The second subscale comprised six items assessing agreement with "Inclusive/Pluralistic attitudes". It included low compliance to RWA (e.g., 'Some of the best people in our country are those who are challenging our government and ignoring the 'normal' way things are supposed to be done'), low SDO (e.g., 'We should do what we can to create equal conditions for different racial or ethnic groups'), acknowledgment of racism (e.g., People from racial or ethnic minority groups experience discrimination in Australia), acknowledgment of white privilege (e.g., 'Australians from an Anglo background (that is, of British descent) enjoy an advantaged position in our society'), and embracement of multiculturalism (e.g. "People from racial or ethnic minority groups benefit Australian society"). Besides their theoretical relevance, these constructs have been found in previous national studies in Australia to be acceptable and appropriate for assessing population race-related attitudes (31) (32). Response options for each item ranged from ‘strongly disagree’ (0), ‘disagree’ (1), ‘neither agree nor disagree’ (2), and ‘agree’ (3) to ‘strongly agree’ (4).

Phase 2: Identification of a potential factorial structure

Since the RRAMS was conceptualized to measure agreement with both issues of conformity to the dominant ethnoculture (“Anglo-centric/Assimilationist attitudes”) and agreement with promotion of ethnic diversity (“Inclusive/Pluralistic attitudes”), an Exploratory Factor Analysis (EFA) was initially run to *empirically* test this assumption (i.e., that a two-factor solution would underlie the set of items). The factorial solution suggested by the EFA was then confirmed by means of a Confirmatory Factor Analysis (33) in an *independent sample* to avoid capitalization on chance (34, 35). We randomly divided the NDTIS sample into one group for the EFA and another group for the CFA. Considering that a sample size with at least 200 participants is sufficient for EFA under normal conditions (medium communalities and at least three measured variables loading on each factor) (36) and CFA has higher sample requirements, 271 participants from the original survey were randomly selected for the EFA.

Factor retention relied on the Scree Plot (37) criteria and Parallel Analysis (PA) (38). In the PA, 1,000 random and resampled datasets with the same number of RRAMS items and respondents were generated. The rationale of the PA is that meaningful factors extracted in the current study should account for more variance than factors extracted from random data (36). Factor extraction was conducted with maximum likelihood (39) and oblique rotation (“direct oblimin”) (40). Items with non-salient factor loadings (.<40) were deleted. Additionally, 100 bootstrapped samples were used to generate factor loadings’ 95% confidence intervals (41).

Phase 3: Confirmation of the factorial structure in an independent sample

After a factorial structure was derived from the EFA, this was assessed using CFA in an independent sample (n = 2,443). The estimation method was weighted least squares (42), with a mean- and variance-adjusted (WLSMV) test statistic (43). Missingness of individual item responses ranged from 0.9% to 2.2%, and this was handled with multiple imputation of 20 datasets using the fully conditional specification method (44). We imputed information for individuals who responded at least one item of the RRAMSs (n = 2,714). Rubin's rules (45) were used to pool point estimates and standard errors (SE). To evaluate model fit, the scaled χ^2 was used to test the hypothesis of *exact-fit.* Additionally, we evaluated *approximate fit* indices, such as the scaled Comparative Fit Index (CFI) and scaled[5] Root Mean Squared Error of Approximation (RMSEA). Values of CFI $\geq$ 0.96 and RMSEA $\leq$ 0.5 indicate good model fit (46), while 0.5 < RMSEA $\leq$ 1.0 indicates acceptable fit (35).

Since factorial structures derived from EFA do not necessarily imply good fitting CFA models (e.g. due to cross-loadings or error correlations) (47), in case the factorial structure had a poor fit, model re-specifications were informed by standardized residuals, modification indices (MI) and the standardized expected parameter change (SEPC) (48). Completely standardized solutions were reported in the present paper.

Phase 4: Analysis of measurement invariance

An initial Multigroup CFA (49) was conducted to check if the same factorial structure would hold for all sex, age, and education-based groups – i.e., to whether *configural invariance* could be confirmed with the data at hand. The χ^2, CFI and RMSEA and their previously described cut-off points were used to evaluate configural invariance. The second level of measurement invariance, *metric invariance*, was assessed to ascertain whether factor loadings were similar across the same groups. The final test, *scalar invariance*, was used to determine whether item thresholds were equal across sex, age and education. Since scalar models are nested within metric models, and metric models are nested within configural models, metric and scalar invariance were evaluated through a Likelihood Ratio Test (LRT),

[5] For simplicity, the term 'scaled' will be omitted from now on.

namely the $\Delta\chi^2$ (50). The $\Delta\chi^2$ statistic was computed in each imputed dataset and pooled according to Li, Meng (51) recommendations (i.e. D2 statistic). When the $\Delta\chi^2$ was *statistically* significant, the ΔCFI (52) was employed to evaluate the *magnitude* of the difference. Models with ΔCFI ≤ -.002 indicated lack of invariance (53). In instances when measurement invariance was not achieved, tests of partial invariance were conducted (54).

Phase 5: Reliability

Internal consistency was calculated with McDonald's Ω_H (55) and ordinal α (56). The McDonald's Ω_H has two advantages over the traditional and widely used Cronbach's α. It does not assume (1) tau-equivalence and a (2) congeneric model without correlated errors (i.e. locally independent items) (57). Furthermore, the ordinal α is reported given that Cronbach's α underestimates reliability in ordinal Likert scales. Adequate methods for calculating ordinal α confidence intervals are not available (58).

Phase 6: Item reduction analysis

In the item reduction analysis, we evaluated inter-item correlations, corrected item-total correlations (CITC) and item difficulties. Inter-item correlations indicate the extent to which all items on a scale are examining the same construct without redundancy. Thus, inter-item correlations should be moderate (i.e. items measure the same construct but also have unique variances) and items with correlations lower than .20 were considered for deletion (59).

The next step was the evaluation of CITC. One important aspect in instrument development is achieving a good balance between a small number of items (lengthy questionnaires can induce lower response rates (60)) and adequate reliability. A recent study by Zijlmans, Tijmstra (61) showed that the CITC (62) performed better than other methods at identifying which items can be removed while maximizing reliability. Therefore, items with the lowest CITC should be the first to be considered for removal. The corrected item-*total* correlation needs to be calculated *within subscale* since items can only be summed into a *total* score when they measure the same construct (63). For this reason, CITCs were calculated *after* the factorial structure was established (i.e. we had no prior information about which item belonged to which subscale to calculate corrected total scores). Given the ordinal nature of the data, the inter-item correlations and CITCs were investigated with non-parametric Kendall's τ (64).

Finally, due to the limitations of classical difficulty indices such as the p-value (i.e. the proportion of correct responses given the total score) (65), we evaluated item difficulty with the LI_{IRF}, the location index based on the item-response function (66). The LI_{IRF} is calculated based on the *item locations* (β_i), which are a well-known reparameterization of *item thresholds* (τ_i) of adjacent *i* and *i* +1 response categories (67). The LI_{IRF} indicates the value of the latent trait in which respondents have an *average* score of half the maximum item score. For example, in a 5-point rating scale (items ranging from 0 = Strongly Disagree to 4 = Strongly Agree), the LI_{IRF} indicates the level of inclusive/pluralistic attitudes required for participants to score *on average* 2 (2 = Neutral). In our study, the LI_{IRF} was chosen over item thresholds (τ_i) to convey item difficulty due to two advantages: the interpretation of the LI_{IRF} is (a) easier, since it is a single index compared to four thresholds per item; and (b) more substantive, since it is based on the *latent trait* ("Anglo-centric/Assimilationist attitudes" or "Inclusive/Pluralistic attitudes") rather than on the *latent response variables* (68). Nonetheless, for the sake of completeness, we also reported the item thresholds (τ_i).

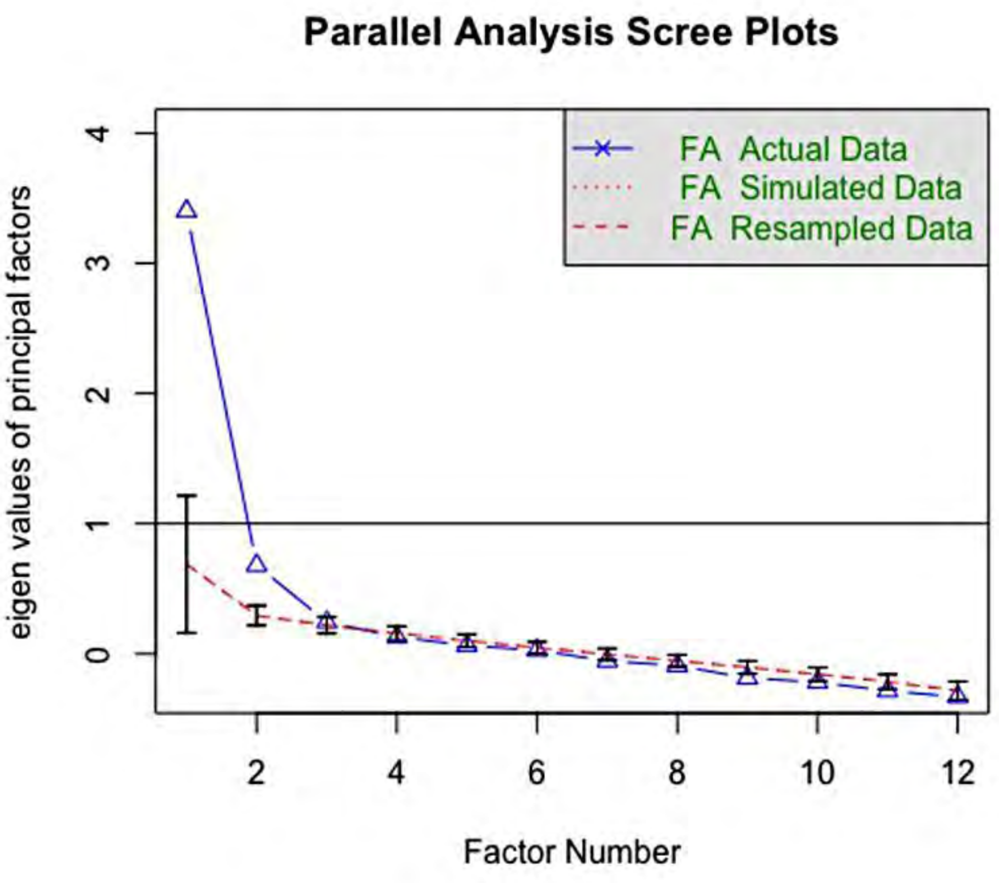
Phase 7: Criterion-related validity

To evaluate the RRAMS' criterion-related validity, we investigated known-groups validity according to sex, education, and age. Known-groups validity compares the levels of the constructs in different groups (e.g. men compared to women) and should be applied when it is known, theoretically or due to previous empirical research, that these groups differ on the variable of interest. Therefore, known-groups validity can inform whether the instrument is able to *discriminate* between two groups that are *known* to be different regarding the construct (e.g. individuals with more education have more inclusive attitudes). The investigation of known-groups validity is important in many instances, such as when there is no “gold standard” method of measurement to which the instrument can be compared (69). That is, since there is no “gold standard” or established (based on robust psychometric evidence) instrument to measure racial related attitudes and multiculturalism in Australia, it is hard to define what would constitute a good measure for the RRAMS to display convergent validity with. Furthermore, in our case, there is previous evidence of groups that are known to differ according to multiculturalism and race-related attitudes. For example, as multiculturalism can be perceived as identity-threatening by dominant group members (11, 19), we expected men to have more conservative attitudes towards multiculturalism when compared to women (22, 70). The same pattern was expected for older participants (>45 years old) when compared to younger respondents (22, 70, 71). Participants with a university degree, in turn, were expected to be more supportive of multiculturalism than those with lower educational attainment. This hypothesis is in accordance with previous findings showing that sense of economic security (economic, personal, and cultural), higher education and younger age were associated with more positive attitudes towards multiculturalism and lesser exclusionary attitudes (22, 70, 71). Therefore, sex, age, and education were chosen as the exogenous variables for the evaluation of known-groups validity. To assess known-groups validity, latent mean differences were calculated by constraining the latent means in one of the groups (i.e. women and participants with higher education) to zero, so this group would function as a reference group. Considering that latent variances were constrained to one in the completely standardized solution, latent mean differences are interpreted as effect sizes analogous to Cohen’s (72) *d* (73). Finally, we employed the Empirical Bayes model (74) to estimate factor scores, which were plotted using Kernel density (75) to inform not only the *average* but also the distribution of the latent trait according to groups.

Results

Identification of a potential factorial structure

Investigation of the Scree Plot and PA indicated that 2 factors substantially explained more variance than factors extracted from randomly generated data (Figure 1).



Note. The triangles indicate the factors' eigenvalues extracted from the study data. The dashed lines and 95% CI indicate the factors' eigenvalues extracted from the 1,000 simulated and resampled datasets. FA stands for factor analysis.

Figure 1. Parallel Analysis and Scree Plots of the Race-related Attitudes and Multiculturalism Scale.

It should be noted that, although the third factor accounted for more variance than the third factor extracted from the random datasets, the difference was trivial. For this reason, only two factors were retained. The next step was the evaluation of the factor loadings (Table 2). Results showed that Item 2 ("Some of the best people in our country are those who are challenging our government and ignoring the 'normal' way things are supposed to be done"), Item 3 ("It is okay if some racial or ethnic groups have better opportunities in life than others") and Item 6 ("We shouldn't talk about racial or ethnic differences") did not have substantial factor loadings (>.40) and were therefore excluded. Item 5 had the smallest factor loadings ($\lambda_2 = 0.440$ 95% CI [0.220, 0.610]).

Table 2. Exploratory Factor Analysis: Factor Loadings (λs) and Bootstrapped 95% CI (n = 271).

	Factor 1		Factor 2	
Item	**Estimate**	**95% CI**	**Estimate**	**95% CI**
1. We need to stop people spreading dangerous ideas and stick to the way things have always been done in Australia.	0.59	[0.40, 0.77]	-0.10	[-0.26, 0.03]
2. Some of the best people in our country are those who are challenging our government and ignoring the 'normal' way things are supposed to be done.	**0.08**	**[-0.16, 0.26]**	**0.38**	**[0.15, 0.57]**
3. It is okay if some racial or ethnic groups have better opportunities in life than others.	**0.27**	**[0.00, 0.47]**	**0.10**	**[-0.16, 0.30]**
4. We should do what we can to create equal conditions for different racial or ethnic groups.	-0.12	[-0.28, 0.02]	0.57	[0.39, 0.74]
5. Australians from an Anglo background (that is, of British descent) enjoy an advantaged position in our society.	-0.03	[-0.25, 0.15]	0.44	[0.22, 0.61]
6. We shouldn't talk about racial or ethnic differences.	**0.23**	**[-0.02, 0.44]**	**-0.06**	**[-0.28, 0.13]**
7. People from racial or ethnic minority groups benefit Australian society.	-0.06	[-0.27, 0.11]	0.47	[0.24, 0.64]
8. People from racial and ethnic minority groups experience discrimination in Australia.	0.01	[-0.16, 0.11]	0.74	[0.57, 0.88]
9. Something more should be done to reduce discrimination experienced by people from racial or ethnic minority groups in Australia.	0.02	[-0.14, 0.14]	0.88	[0.73, 1.00]
10.Racial or ethnic minority groups take away jobs from other Australians.	0.65	[0.44, 0.81]	-0.07	[-0.27, 0.07]
11.The Australian way of life is weakened by people from minority racial or ethnic backgrounds maintaining their cultural beliefs and values.	0.65	[0.46, 0.83]	0.04	[-0.11, 0.13]
12.People from racial and ethnic minority groups should behave more like mainstream Australians.	0.81	[0.63, 0.95]	0.01	[-0.18, 0.13]

Note. Deleted items highlighted in bold.

After the deletion of these four items and EFA re-analysis, the two-factor solution achieved simple structure. This time, however, Item 5 did not achieve a substantial factor loading ($\lambda_2 = 0.390$; 95% CI [0.180, 0.590]) (Supplementary Table 1); that is, the factors explained only 19% of the variance of item responses ("communality"), while 81% of the variance was explained by other sources ("uniqueness"), such as measurement error. For this reason, Item 5 was also excluded from the analysis.

Confirmation of the factorial structure in an independent sample

The 2-factor model was then selected and its fit, examined ($\chi^2(19) = 341.070$, $p<0.001$, CFI = 0.974, RMSEA = 0.083; 90% CI [0.076, 0.091]). Since the null hypothesis of *exact-fit* was rejected ($\chi^2(19) = 341.070$, $p<0.001$), we proceeded to evaluate the indices of *approximate-fit*. The CFI indicated a good fit to the data (>.960), while the RMSEA was adequate ($0.5 < \text{RMSEA} \leq 1.0$). Residual correlations are displayed in Supplementary Table 2. Considering the overall good fit of the model and that all items exhibited substantial factor loadings (Table 3), the two-factor model with 8 items was accepted.

"Anglo-centric/Assimilationist attitudes" (e.g. "Racial or ethnic minority groups take away jobs from other Australians"), whereas the second subscale comprised six items assessing agreement with "Inclusive/Pluralistic attitudes"

Table 3. Confirmatory Factor Analysis: Factor Loadings (λs) and Factor correlations (n = 2,443).

Item	Estimate (SE)	*p*-value	95% C.I.	CITC	LI_{IRF}
Subscale 1: Anglo-centric/Assimilationist attitudes					
1. We need to stop people spreading dangerous ideas and stick to the way things have always been done in Australia.	0.629 (0.014)	<0.001	[0.601, 0.656]	0.43	0.00
10. Racial or ethnic minority groups take away jobs from other Australians.	0.784 (0.010)	<0.001	[0.764, 0.804]	0.50	0.72
11.The Australian way of life is weakened by people from minority racial or ethnic backgrounds maintaining their cultural beliefs and values.	0.856 (0.009)	<0.001	[0.838, 0.874]	0.58	0.44
12.People from racial and ethnic minority groups should behave more like mainstream Australians.	0.814 (0.010)	<0.001	[0.794, 0.834]	0.57	0.01
Subscale 2: Inclusive/Pluralistic attitudes					
4. We should do what we can to create equal conditions for different racial or ethnic groups.	0.652 (0.016)	<0.001	[0.620, 0.684]	0.41	-1.58
7. People from racial or ethnic minority groups benefit Australian society.	0.627 (0.016)	<0.001	[0.595, 0.658]	0.39	-1.16
8. People from racial and ethnic minority groups experience discrimination in Australia.	0.680 (0.013)	<0.001	[0.655, 0.706]	0.43	-0.80
9. Something more should be done to reduce discrimination experienced by people from racial or ethnic minority groups in Australia.	0.835 (0.012)	<0.001	[0.813, 0.858]	0.54	-0.86
Factor correlation (anglo-centric/assimilationist attitudes x inclusive/pluralistic attitudes)	-0.638 (0.016)	<0.001	[-0.669, -0.608]	-	-

Note. CITC = Corrected Item-Total Correlations. LI_{IRF} = Location Index based on the Item Response Function. Standardized factor loadings are displayed. Point estimates and SE were pooled across 20 imputed datasets according to Rubin's rules. LI_{IRF} was calculated based on pooled item thresholds and factor loadings.

Analysis of measurement invariance

Next, measurement invariance by sex, education, and age was evaluated. Regarding sex, the LRT indicated that the metric model was not statistically different from the configural model ($\Delta\ \chi 2$ (6) = 11.86; p = 0.065) and that the scalar model was not statistically different from the metric model ($\Delta\ \chi 2$ (16) = 24.26; p = 0.083). In other words, factor loadings and thresholds were invariant across men and women. Regarding education, although the configural model and scalar model were statistically different (Δ $\chi2$ (6) = 19.14; p = 0.004), the fit of the (constrained) metric model improved (ΔCFI =0.002) providing evidence of metric invariance between those with and without higher education. The same happened when metric invariance was evaluated by age; although the configural model and scalar model were statistically different (Δ $\chi2$ (6) = 15.15; p = 0.019), the fit of the metric model (ΔCFI = 0.005) was better. When scalar invariance was evaluated, the pooled Δ $\chi2$ was negative for both education and age-based groups. Although a negative Δ $\chi2$ is not interpretable (and, therefore, values were set to zero), these negative values can occur when the difference between models is small (76). For this reason, the threshold constraints were regarded as tenable (77) and provided indirect support for scalar invariance.

Table 4. Measurement invariance according to sex and education.

Model	χ^2	*df*	*p*-value	RMSEA	90% CI	CFI	Δ $\chi2$ (df)	*p*-value	Δ CFI
Sex									
Configural	381.703	38	<0.001	0.086	[0.078, 0.094]	0.973	-	-	-
Metric	340.310	44	<0.001	0.074	[0.067, 0.082]	0.976	11.86 (6)	0.065	0.003
Scalar	428.058	60	<0.001	0.074	[0.065, 0.077]	0.971	24.26 (16)	0.083	0.005
Education									
Configural	363.867	38	<0.001	0.084	[0.076, 0.092]	0.974	**-**	**-**	**-**
Metric	339.008	44	<0.001	0.074	[0.067, 0.082]	0.976	19.14 (6)	0.004	0.002
Scalar	422.999	60	<0.001	0.070	[0.064, 0.077]	0.971	0 (6)*	1.000	-0.005
Age									
Configural	385.254	38	<0.001	0.087	[0.079, 0.094]	0.973	**-**	**-**	**-**
Metric	332.751	44	<0.001	0.073	[0.066, 0.081]	0.978	15.15 (6)	0.019	0.005
Scalar	386.834	60	<0.001	0.067	[0.061, 0.073]	0.975	0 (6)*	1.000	-0.003

Note. $\chi2$ = chi-square; df = degrees of freedom; RMSEA = root mean square error of approximation; CFI = comparative fit index; $\Delta\chi2$ (df) = chi-square difference and degrees of freedom; ΔCFI = CFI difference. * Negative pooled test statistic was set to zero.

Reliability

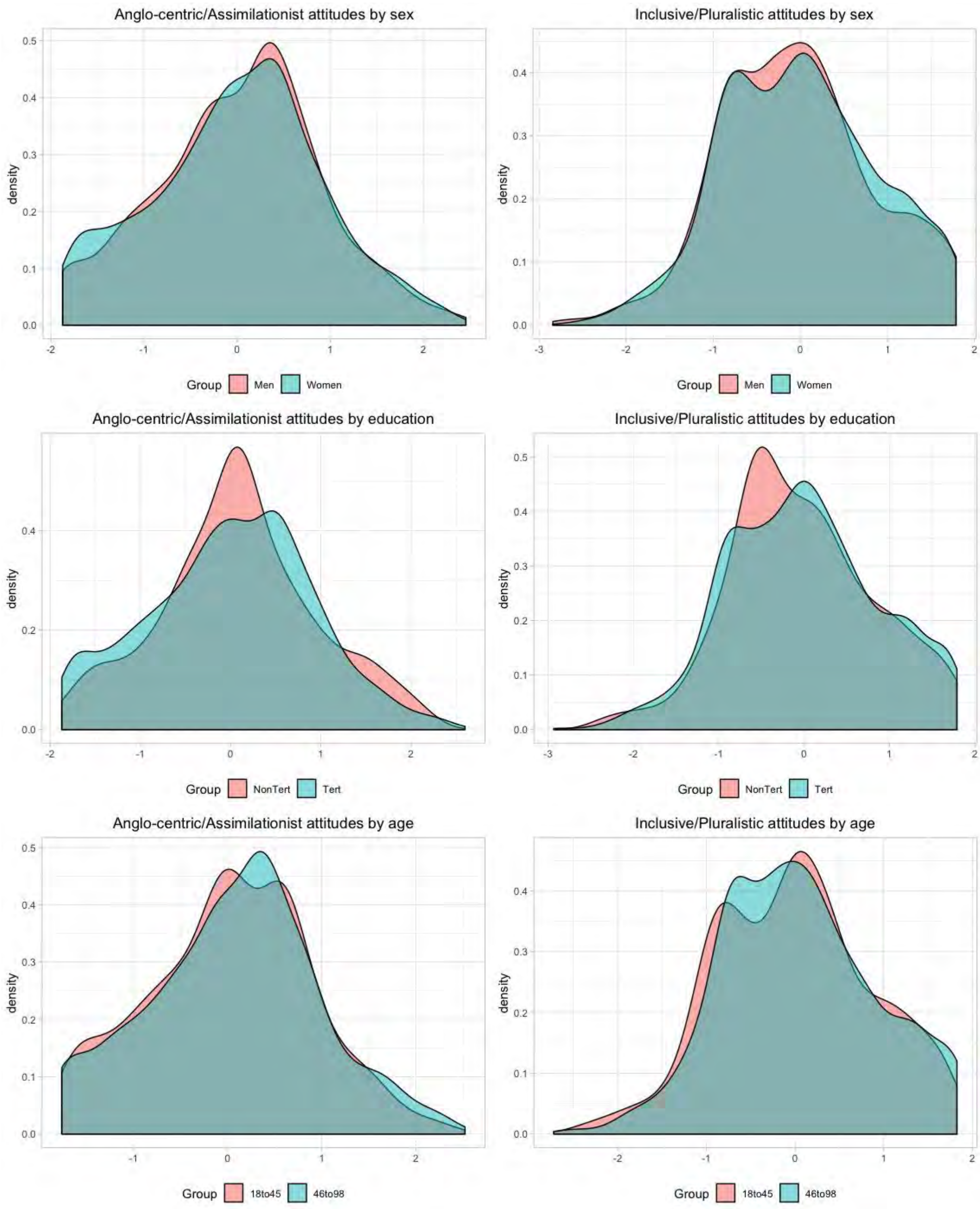
The first subscale "Anglo-centric/Assimilationist attitudes" ($\Omega_H = 0.83$, $\alpha_{ORDINAL} = 0.85$, α = 0.85; 95% CI [0.84, 0.86]) showed good reliability, while the "Inclusive/Pluralistic attitudes" subscale (Ω_H = 0.77, $\alpha_{ORDINAL}$ = 0.79, α = 0.72; 95% CI [0.70, 0.73]) exhibited adequate reliability.

Item Reduction Analysis

Inter-item correlations ranged from 0.29 to 0.56 (Supplementary 3) and no correlations were lower than 0.20. The CITCs ranged from 0.39 to 0.58. Within the "Anglo-centric/Assimilationist attitudes" dimension, the easiest item was "We need to stop people spreading dangerous ideas and stick to the way things have always been done in Australia" (LI_{IRF} = 0.00), while the hardest item was "Racial or ethnic minority groups take away jobs from other Australians" (LI_{IRF} = 0.72) (Table 3). That is, with respect to Item 10, respondents needed to have 0.72 standard deviations more anglo-centric/assimilationist attitudes than the average Australian to produce an expected score of 2 out of 4. Item 10 was the *hardest* item in the "Anglo-centric/Assimilationist attitudes" subscale since its endorsement required more anglo-centric/assimilationist attitudes than the other items. Within the "Inclusive/Pluralistic attitudes" subscale, the easiest item was "We should do what we can to create equal conditions for different racial or ethnic groups" (LI_{IRF} = -1.58), while the hardest item was "People from racial and ethnic minority groups experience discrimination in Australia." (LI_{IRF} = -0.80). The hierarchy of item difficulties was identical when average item thresholds ($\overline{\tau}$) were inspected (Supplementary Table 4).

Criterion-related validity

Examination of criterion-related validity indicated that men (M = 0.105; 95% CI [0.014, 0.197]), participants without tertiary education (M = 0.585; 95% CI [0.474, 0.696]) and those aged 45 years and over (M = 0.373; 95% CI [0.275, 0.470]) had higher Anglo-centric/assimilationist attitudes. Furthermore, men (M = -0.116; 95% CI [-0.213, -0.020]) and participants without tertiary education (M = -0.304; 95% CI [-0.420, -0.188]) also presented lower inclusive/pluralistic attitudes. The difference in inclusive/pluralistic attitudes between participants aged 45 years and over (M = -0.045; 95% CI [-0.148, 0.057]) and their peers was close to zero. The distribution of factor scores is displayed in Figure 2.



Note. The Kernel density plots indicate the distribution of factor scores.

Figure 2. Factor scores Kernel density plots of assimilationist and pluralistic attitudes.

Discussion

The current study aimed to present the RRAMS as a measure of attitudes towards multiculturalism in Australia and to examine its psychometric properties using data from a nationwide sample. The two-factor solution proposed in exploratory stages of the analysis was thereafter confirmed by means of a CFA in an independent sample. Results showed that the two subscales of “Anglocentric/Assimilationist attitudes” and “Inclusive/Pluralistic attitudes” are initially valid and reliable for the Australian population.

In the initial stage of psychometric assessment, we identified poorly performing items, and these were excluded. One of these was Item 2 (“Some of the best people in our country are those who are challenging our government and ignoring the ‘normal’ way things are supposed to be done”), an item originally designed to reflect RWA in relation to multiculturalism. Despite its original purpose, Item 2 might not reflect the cultural and race-related topic in question. This is one possible explanation why the responses to this item were not strongly influenced by respondent’s Inclusive/Pluralistic attitudes towards multiculturalism (only 12% of the variance was explained by the factor). For instance, the wording “challenging our government” can be interpreted as referring to a general debate not reflecting ethnic-racial differences on political representation and resource distribution, for example. Future studies might test the item fit by emphasizing ‘challenging our government’ as pressuring for a political agenda that prioritize reducing social inequalities among ethnic-racial groups and promotion of a pluralistic society.

Items 3 (“It is okay if some racial or ethnic groups have better opportunities in life than others”) and 6 (“We shouldn’t talk about racial or ethnic differences”) also performed poorly and failed to capture assimilationist views. Item 3 was designed to reflect the respondent’s SDO. It was hypothesized that participants with high SDO, and thus assimilationist views of multiculturalism, would endorse the item. Contrarily to expected, these respondents might have interpreted the phrasing ‘some racial or ethnic groups’ as a reference to ethnic-racial minorities. Conservatives might perceive affirmative action and social assistance policies as privileges and can endorse the notion that minorities ‘have it easy’. Conservative attitudes such as that of RWA and SDO have been linked to social and economic conservatism, reflecting ideologies of competition and meritocracy (78). The ambiguity left by the item wording can thus explain its failure in discriminating assimilationist attitudes. Item 6, in turn, might have not worked in its subdomain because, again contrarily to our hypothesis, respondents with high assimilationist views might be *willing* to discuss racial and ethnic differences with the intent of promoting assimilationist and racist views (79). Therefore, the item performed poorly as respondents in the different strata of assimilationist attitudes could be prone do endorse the item for different reasons.

The last deleted item was Item 5 (“Australians from an Anglo background [that is, of British descent] enjoy an advantaged position in our society”). One possible explanation for the item’s poor performance is that the recognition of privilege *does not necessarily* inform on inclusive/pluralistic attitudes. For example, a previous study in the Australian states of Queensland and New South Wales showed these as two independent dimensions (9). The item poor loading in the inclusive attitudes domain suggests respondents might not link acknowledgment of white privilege to their notion of a pluralistic society. Taken together, these results potentially indicate that debates over multiculturalism in Australia need to promote awareness of the connection between Anglo-privilege and racism. Scholars advocate that challenging racism and privilege is a necessary step to promote the abandonment of assimilationist views in favour of more inclusive perspectives (9, 13).

The subscales “Anglo-centric/Assimilationist attitudes” and “Inclusive/Pluralistic attitudes” achieved *metric invariance* and *scalar invariance* according to sex. Furthermore, the two subscales achieved *metric invariance* according education and the results also (indirectly) supported *scalar invariance*. That is, “Anglo-centric/Assimilationist attitudes” and "Inclusive/Pluralistic attitudes" influenced the item responses the same way in each group (*metric invariance*) and the items were not more difficult for one group compared to another (*scalar invariance*). The RRAMS items can thus be used to compare men/women, participants with/without tertiary education and young/older participants, and the scores will reflect *true differences* regarding "Anglo-centric/Assimilationist attitudes" and "Inclusive/Pluralistic attitudes" rather than measurement bias (35).

After ensuring measurement invariance between subgroups, we compared the factor scores between men and women, participants with and without tertiary education, and participants up to and over 45 years of age. The stronger predictor of assimilationist *and* inclusive attitudes was education status, while sex also influenced both constructs. Furthermore, older individuals were more likely to have higher assimilationist attitudes. The role of education in promoting inclusive/pluralistic has been previously established (22, 70) and suggests education as an important target for future interventions aimed at promoting multiculturalism in Australia. The results also indicated that men and older individuals had stronger assimilationist attitudes in comparison with women and younger counterparts (71). In general, the associations of the two subscales with sex, education, and age conformed to the theoretical expectations and provide further evidence of the RRAMS' construct validity.

With regards to reliability, the "Anglo-centric/Assimilationist attitudes" and "Inclusive/Pluralistic attitudes" subscales showed adequate reliability (>.70) (80), since values between .70 and .80 are considered appropriate for research purposes (81). In case the RRAMS is used in the future in high-stakes scenarios (i.e. where decisions need to be made based on scale scores) (82), new items should be developed to increase reliability.

In the item reduction analysis, all items displayed moderate inter-item correlations and CITC, so no items were required to be removed. The item with the smallest CITC was Item 7 ("People from racial or ethnic minority groups benefit Australian society"), followed by Item 4 ("We should do what we can to create equal conditions for different racial or ethnic groups."). Since reliability was only modest, we considered that further shortening the questionnaire would be more detrimental in terms of reliability and content validity than beneficial as a means of creating a briefer measure. In addition, with the exception of Item 1 ("We need to stop people spreading dangerous ideas and stick to the way things have always been done in Australia.") and Item 12 ("People from racial and ethnic minority groups should behave more like mainstream Australians."), the items difficulties were spread across the latent trait. Once again, although Item 1 or Item 2 could potentially be removed due to similar difficulties, we believe removing additional items would be detrimental to content validity and the psychometric properties of the scale.

One limitation of the current study was that we were not able to evaluate convergent and discriminant validity. The RRAMS was originally applied at the 2013 NDTIS, a study that focused on collecting information on the use of dental services in Australia and did not include other psychosocial measures. For this reason, we considered known-groups validity to be the best strategy to investigate the RRAMS' criterion-related validity. While the results from known-groups validity were in accordance with theoretical expectations (e.g. inclusive attitudes were more present in individuals with more education), future studies need also to investigate other forms of validity, such as convergent/discriminant and predictive validity. For example, future studies should evaluate whether the scores from the "Inclusive/Pluralistic attitudes" subscale are positively correlated (i.e. convergent validity) with scores from other instruments evaluating multiculturalist and inclusive attitudes. Our analyses did not account for sampling weights, meaning that our sample is not representative of the Australian population. It is important to highlight, however, that our study included Australians from all age groups and socioeconomic backgrounds across all states and territories of the country. Furthermore, to the best of our knowledge, this is the largest sample in which a measure of attitudes towards multiculturalism has been employed in Australia. Lack of representativeness and its implications to the validity of scientific findings are central to longstanding discussions in the literature (ref.). Because the purpose of the current analysis was to assess the psychometric properties of the RRAMS, as opposed to purely describe prevalence estimates, we do not believe that the lack of representativeness of our sample limits the validity of inferences made here. The fact that a study sample is representative of some larger population does not mean that the associations or correlations between variables in the sample will apply to every subgroup of the population (ref.). The overall association or correlation is simply an average value that has been balanced according to the distribution of people in these subgroups. If a sample that is representative of the sex distribution in the target population, the results will not necessarily apply to both males and females, but only to a hypothetical participant that is "weighted" on sex. Subgroups analyses are necessary if one wishes to investigate relationships between variables by subgroups, which we have performed during the criterion validity assessment stage.

In conclusion, we successfully developed a comprehensive race-related attitudes and multiculturalism scale to the Australian context. We used robust, cutting edge psychometric techniques and a large, nation-wide survey. The small number of items (eight) means the instrument will likely be readily used by policymakers and in future research. Future studies should assess the scaling properties of the instrument by using parametric and non-parametric Item Response Theory techniques. The instrument may, nevertheless, be useful to inform on multiculturalism attitudes across the country and hopefully contribute to a public debate aimed to promote multiculturalist inclusive attitudes with the potential to increase social cohesion in Australia.

References

1. Berman G, Paradies Y. Racism, disadvantage and multiculturalism: towards effective anti-racist praxis. Ethnic and Racial Studies. 2010;33(2):214-32.

2. Elias A, Paradies Y. Estimating the mental health costs of racial discrimination. BMC public health. 2016;16(1):1205-.

3. Hage G. Analysing multiculturalism today. In: Bennett T, Frow J, editors. The SAGE handbook of cultural analysis: SAGE Publishing; 2007. p. 488-509.

4. Habtegiorgis AE, Paradies YC, Dunn KM. Are racist attitudes related to experiences of racial discrimination? Within sample testing utilising nationally representative survey data. Soc Sci Res. 2014;47:178-91.

5. Kauff M, Wagner U. Valuable Therefore Not Threatening The Influence of Diversity Beliefs on Discrimination Against Immigrants2012. 714-21 p.

6. Modood T. Multiculturalism. In: Ritzer G, editor. The Blackwell Encyclopedia of Sociology: John Wiley & Sons; 2016.

7. Grigg K, Manderson L. The Australian Racism, Acceptance, and Cultural-Ethnocentrism Scale (RACES): item response theory findings. International Journal for Equity in Health. 2016;15(1):49.

8. Hage GaB. Analysing Multiculturalism Today. The SAGE Handbook of Cultural Analysis2008. p. 488-509.

9. Forrest J, Dunn K. 'Core'Culture Hegemony and Multiculturalism. Ethnicities. 2016;6(2):203-30.

10. Vala J, Costa-Lopes R. Intergroup relations. In: Smelser NJ, Baltes PB, editors. International Encyclopedia of the Behavioral Sciences. 2nd edition ed: Elsevier; 2015.

11. Verkuyten M. Support for Multiculturalism and Minority Rights: The Role of National Identification and Out-group Threat. Social Justice Research. 2009;22(1):31-52.

12. Morrison KR, Ybarra O. The effects of realistic threat and group identification on social dominance orientation. J Exp Soc Psychol. 2008;44(1):156-63.

13. Dunn K, Nelson JK. Challenging the Public Denial of Racism for a Deeper Multiculturalism. Journal of Intercultural Studies. 2011;32(6):587-602.

14. Pratto F, Sidanius J, Levin S. Social dominance theory and the dynamics of intergroup relations: Taking stock and looking forward. European Review of Social Psychology. 2006;17(1):271-320.

15. Adorno TW, Frenkel-Brunswik E, Levinson D, Sandford RN. The Authoritarian personality. New York: Harper; 1950.

16. Bizumic B, Duckitt J. Investigating right wing authoritarianism with a very short authoritarianism scale. Journal of Social and Political Psychology. 2018;6(1):129-50.

17. Duckitt J. Authoritarianism and dogmatism. In: Leary M, Hoyle R, editors. Handbook of individual differences in social behavior. New York: Guilford Press; 2009. p. 298-317.

18. Perry R, Sibley CG. Big-Five personality prospectively predicts Social Dominance Orientation and Right-Wing Authoritarianism. Pers Individ Dif. 2012;52(1):3-8.

19. Sumino T. National identity and public attitudes toward multiculturalism in Canada: Testing the indirect effect via perceived collective threat. Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement. 2017;49(3):183-94.

20. Doane A. Beyond Color-blindness: (Re) Theorizing Racial Ideology. Sociological Perspectives. 2017;60(5):975-91.

21. Bonilla-Silva E. Racism without racists: Color-Blind Racism and the persistence of racial inequality in the United States. New York: Rowman & Littlefield; 2003.

22. Dunn K, Blair K, Alam O, Kamp A. Australians' Views on Cultural Diversity, Nation and Migration, 2015-16. Cosmopolitan Civil Societies: An Interdisciplinary Journal. 2017;9(3):61-84.

23. Marwick A, Ansari Z, Sullivan M, Parsons L, McNeil J. Inequalities in the social determinants of health of Aboriginal and Torres Strait Islander People - A cross-sectional population-based study in the Australian state of Victoria. International Journal for Equity in Health. 2014;13(91):1-12.

24. Curthoys A. An uneasy conversation: the multicultural and the Indigenous. In: Docker J, Fischer G, editors. Race, colour and identity in Australia and New Zealand. Sydney: UNSW Press; 2000. p. 21-36.

25. Denson N, Ovenden G, Wright L, Paradies Y, Priest N. The development and validation of intercultural understanding (ICU) instruments for teachers and students in primary and secondary schools. Intercultural Education. 2017;28(3):231-49.

26. Australian Institute of Health and Welfare. Australia's welfare 2017. Canberra: Australian Institute of Health and Welfare; 2017.

27. R Core Team. R: A language and environment for statistical computing. 2013.

28. Rosseel Y. Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). Journal of statistical software. 2012;48(2):1-36.

29. Jorgensen T, Pornprasertmanit S, Schoemann A, Rosseel Y. semTools: Useful tools for structural equation modeling. R Package Version 0.5-1. 2018.

30. Haynes SN, Richard DCS, Kubany ES. Content validity in psychological assessment: A functional approach to concepts and methods. Psychol Assess. 1995;7(3):238-47.

31. Blair K, Dunn, K., Kamp, A., & Alam, O. Challenging Racism Project 2015-16 National Survey Report. In: University. SWS, editor. 2017.

32. VicHealth. Findings from the 2013 survey of Victorians' attitudes to race and cultural diversity. In: Foundation VHP, editor. Melbourne, Australia2014.

33. Rees S, Silove D, Chey T, Ivancic L, Steel Z, Creamer M, et al. Lifetime prevalence of gender-based violence in women and the relationship with mental disorders and psychosocial function. Journal of the American Medical Association. 2011;306(5):513-21.

34. Fokkema M, Greiff S. How Performing PCA and CFA on the Same Data Equals Trouble. Hogrefe Publishing; 2017.

35. Kline RB. Principles and practice of structural equation modeling. New York, NY: Guilford publications; 2015.

36. Fabrigar LR, Wegener DT. Exploratory factor analysis: Oxford University Press; 2011.

37. Cattell RB. The scree test for the number of factors. Multivariate behavioral research. 1966;1(2):245-76.

38. Horn JL. A rationale and test for the number of factors in factor analysis. Psychometrika. 1965;30(2):179-85.

39. Lawley DN. Vi.—the estimation of factor loadings by the method of maximum likelihood. Proceedings of the Royal Society of Edinburgh. 1940;60(1):64-82.

40. Harman HH. Modern factor analysis: University of Chicago press; 1976.

41. Efron B, Tibshirani RJ. An introduction to the bootstrap: CRC press; 1994.

42. Muthén B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika. 1984;49(1):115-32.

43. Asparouhov T, Muthén B. Simple second order chi-square correction. Mplus technical appendix. 2010.

44. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res. 2007;16(3):219-42.

45. Rubin DB. Multiple imputation for nonresponse in surveys: John Wiley & Sons; 2004.

46. Yu C-Y. Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes: University of California, Los Angeles Los Angeles; 2002.

47. Van Prooijen J-W, Van Der Kloot WA. Confirmatory analysis of exploratively obtained factor structures. Educ Psychol Meas. 2001;61(5):777-92.

48. Saris WE, Satorra A, Van der Veld WM. Testing structural equation models or detection of misspecifications? Structural Equation Modeling. 2009;16(4):561-82.

49. Wu H, Estabrook R. Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. Psychometrika. 2016;81(4):1014-45.

50. Satorra A. Scaled and adjusted restricted tests in multi-sample analysis of moment structures. Innovations in multivariate statistical analysis: Springer; 2000. p. 233-47.

51. Li K-H, Meng X-L, Raghunathan TE, Rubin DB. Significance levels from repeated p-values with multiply-imputed data. Statistica Sinica. 1991:65-92.

52. Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. Structural equation modeling. 2002;9(2):233-55.

53. Meade AW, Johnson EC, Braddy PW. Power and sensitivity of alternative fit indices in tests of measurement invariance. J Appl Psychol. 2008;93(3):568.

54. Byrne BM, Shavelson RJ, Muthén B. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychol Bull. 1989;105(3):456.

55. McDonald RP. Test theory: A unified treatment: Psychology Press; 2013.

56. Zumbo BD, Gadermann AM, Zeisser C. Ordinal versions of coefficients alpha and theta for Likert rating scales. Journal of modern applied statistical methods. 2007;6(1):4.

57. Dunn TJ, Baguley T, Brunsden V. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. Br J Psychol. 2014;105(3):399-412.

58. Turner HJ, Natesan P, Henson RK. Performance Evaluation of Confidence Intervals for Ordinal Coefficient Alpha. Journal of Modern Applied Statistical Methods. 2017;16(2):9.

59. Piedmont RL. Inter-item correlations. Encyclopedia of quality of life and well-being research. 2014:3303-4.

60. Rolstad S, Adler J, Rydén A. Response burden and questionnaire length: is shorter better? A review and meta-analysis. Value Health. 2011;14(8):1101-8.

61. Zijlmans EAO, Tijmstra J, der Ark V, Andries L, Sijtsma K. Item-score reliability as a selection tool in test construction. Front Psychol. 2018;9:2298.

62. Guilford JP. The correlation of an item with a composite of the remaining items in a test. Educ Psychol Meas. 1953;13(1):87-93.

63. Gardner PL. Measuring attitudes to science: Unidimensionality and internal consistency revisited. Research in science education. 1995;25(3):283-9.

64. Kendall SM. Rank correlation. Van Nostrand's Scientific Encyclopedia. 1948.

65. Embretson SE. The new rules of measurement. Psychol Assess. 1996;8(4):341.

66. Ali US, Chang HH, Anderson CJ. Location Indices for Ordinal Polytomous Items Based on Item Response Theory. ETS Research Report Series. 2015;2015(2):1-13.

67. Muthen BO. Instructional Sensitivity in Mathematics Achievement Test Items: Application of a New IRT-Based Detection Technique. 1988.

68. Ferrando PJ. Difficulty, discrimination, and information indices in the linear factor analysis model for continuous item responses. Appl Psychol Meas. 2009;33(1):9-24.

69. Davidson M. Known-groups validity. Encyclopedia of quality of life and well-being research: Springer; 2014. p. 3481-2.

70. Dandy J, Pe-Pua R. Attitudes to multiculturalism, immigration and cultural diversity: Comparison of dominant and non-dominant groups in three Australian states. International Journal of Intercultural Relations. 2010;34(1):34-46.

71. Dunn KM, Forrest J, Burnley I, McDonald A. Constructing Racism in Australia. Australian Journal of Social Issues. 2004;39(4):409-30.

72. Cohen J. Statistical power analysis for the behavioral sciences. 2nd. Hillsdale, NJ: erlbaum; 1988.

73. Hancock GR. Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. Psychometrika. 2001;66(3):373-88.

74. Rabe-Hesketh S, Skrondal A. Assigning values to latent variables. Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models: Chapman and Hall/CRC; 2004.

75. Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society Series B (Methodological). 1991:683-90.

76. Satorra A, Bentler PM. Ensuring positiveness of the scaled difference chi-square test statistic. Psychometrika. 2010;75(2):243-8.

77. Marshall GN, Miles JN, Stewart SH. Anxiety sensitivity and PTSD symptom severity are reciprocally related: Evidence from a longitudinal study of physical trauma survivors. J Abnorm Psychol. 2010;119(1):143.

78. Harnish RJ, Bridges KR, Gump JT. Predicting Economic, Social, and Foreign Policy Conservatism: the Role of Right-Wing Authoritarianism, Social Dominance Orientation, Moral Foundations Orientation, and Religious Fundamentalism. Current Psychology. 2017;37(3):668-79.

79. Faulkner N, Bliuc A-M. 'It's okay to be racist': moral disengagement in online discussions of racist incidents in Australia. Ethnic Racial Stud. 2016;39(14):2545-63.

80. Nunnally JC, Bernstein IH. Psychometric theory: McGraw-Hill New York; 1967.

81. Furr RM, Bacharach VR. Psychometrics: an introduction. Thousand Oaks, CA: Sage; 2013.

82. Wells CS, Wollack JA. An instructor's guide to understanding test reliability. Testing & Evaluation Services University of Wisconsin. 2003.

Supplementary Materials.

Supplementary Table 1. Exploratory Factor Analysis: Factor Loadings (λs) and Bootstrapped 95% CI.

	Factor 1		Factor 2	
Item	Estimate	95% C.I.	Estimate	95% C.I.
1. We need to stop people spreading dangerous ideas and stick to the way things have always been done in Australia.	0.60	[0.43, 0.74]	-0.09	[-0.23, 0.04]
4. We should do what we can to create equal conditions for different racial or ethnic groups.	-0.12	[-0.26, 0.00]	0.57	[0.41, 0.72]
5. Australians from an Anglo background (that is, of British descent) enjoy an advantaged position in our society.	**-0.06**	**[-0.26, 0.09]**	**0.39**	**[0.18, 0.59]**
7. People from racial or ethnic minority groups benefit Australian society.	-0.06	[-0.24, 0.10]	0.45	[0.23, 0.67]
8. People from racial and ethnic minority groups experience discrimination in Australia.	-0.02	[-0.17, 0.09]	0.71	[0.54, 0.86]
9. Something more should be done to reduce discrimination experienced by people from racial or ethnic minority groups in Australia.	0.04	[-0.09, 0.15]	0.91	[0.79, 1.00]
10.Racial or ethnic minority groups take away jobs from other Australians.	0.62	[0.40, 0.79]	-0.09	[-0.30, 0.07]
11.The Australian way of life is weakened by people from minority racial or ethnic backgrounds maintaining their cultural beliefs and values.	0.83	[0.71, 0.93]	0.04	[-0.08, 0.13]
12.People from racial and ethnic minority groups should behave more like mainstream Australians.	0.82	[0.68, 0.94]	0.02	[-0.12, 0.13]

Note. Deleted items were highlighted in bold. This is the result of the Exploratory Factor Analysis after items 2, 3 and 6 were deleted.

Supplementary Table 2. Matrix of residual correlations.

	Item 1	Item 10	Item 11	Item 12	Item 4	Item 7	Item 8	Item 9
Item 1	0.000							
Item 10	-0.032	0.000						
Item 11	-0.019	-0.003	0.000					
Item 12	0.051	-0.040	0.018	0.000				
Item 4	0.033	-0.061	-0.006	0.045	0.000			
Item 7	-0.034	-0.088	-0.049	-0.008	0.020	0.000		
Item 8	-0.022	0.008	0.060	0.020	-0.053	-0.055	0.000	
Item 9	0.061	-0.036	0.037	0.043	0.011	-0.059	0.055	0.000

Note. The residual correlations matrix displays the observed correlation between item responses after the influence of the latent factors was accounted by the model.

Supplementary Table 3. Matrix of inter-item correlations.

	Item 1	Item 10	Item 11	Item 12	Item 4	Item 7	Item 8	Item 9
Item 1	1.000							
Item 10	0.341	1.000						
Item 11	0.394	0.537	1.000					
Item 12	0.429	0.448	0.563	1.000				
Item 4	-0.157	-0.294	-0.275	-0.204	1.000			
Item 7	-0.212	-0.329	-0.314	-0.250	0.345	1.000		
Item 8	-0.222	-0.268	-0.249	-0.252	0.293	0.292	1.000	
Item 9	-0.211	-0.367	-0.332	-0.298	0.438	0.389	0.516	1.000

Note. The inter-item correlations matrix displays the observed marginal correlations between item responses.

Supplementary Table 4. Confirmatory Factor Analysis: Item Thresholds (τ), Item Locations (β) and Item Difficulties (LI_{IRF}).

Item	τ_1 (SE)	τ_2 (SE)	τ_3 (SE)	τ_4 (SE)	$\bar{\tau}$	β_1	β_2	β_3	β_4	LI_{IRF}
Subscale 1: Anglo-centric/Assimilationist attitudes										
1. We need to stop people spreading dangerous ideas and stick to the way things have always been done in Australia.	-0.881 (0.030)	-0.372 (0.026)	0.422 (0.026)	0.834 (0.029)	0.001	-1.400	-0.591	0.671	1.326	0.005
10. Racial or ethnic minority groups take away jobs from other Australians.	-0.455 (0.027)	0.150 (0.026)	1.020 (0.031)	1.504 (0.040)	0.555	-0.580	0.191	1.301	1.918	0.716
11.The Australian way of life is weakened by people from minority racial or ethnic backgrounds maintaining their cultural beliefs and values.	-0.474 (0.027)	0.055 (0.026)	0.709 (0.028)	1.228 (0.034)	0.379	-0.554	0.064	0.828	1.434	0.444
12.People from racial and ethnic minority groups should behave more like mainstream Australians.	-0.989 (0.031)	-0.417 (0.026)	0.442 (0.027)	0.995 (0.031)	0.008	-1.215	-0.512	0.543	1.222	0.011
Subscale 2: Inclusive/Pluralistic attitudes										
4. We should do what we can to create equal conditions for different racial or ethnic groups.	-1.768 (0.047)	-1.470 (0.039)	-0.753 (0.028)	-0.084 (0.026)	-1.019	-2.712	-2.255	-1.155	-0.129	-1.579
7. People from racial or ethnic minority groups benefit Australian society.	-1.698 (0.045)	-1.283 (0.035)	-0.318 (0.026)	0.432 (0.027)	-0.717	-2.708	-2.046	-0.507	0.689	-1.165
8. People from racial and ethnic minority groups experience discrimination in Australia.	-1.624 (0.043)	-1.137 (0.033)	-0.120 (0.033)	0.811 (0.029)	-0.517	-2.388	-1.672	-0.176	1.193	-0.801
9. Something more should be done to reduce discrimination experienced by people from racial or ethnic minority groups in Australia.	-1.794 (0.046)	-1.294 (0.035)	-0.250 (0.026)	0.540 (0.027)	-0.699	-2.095	-1.550	-0.299	0.647	-0.858

Note. τ_i = item thresholds of adjacent categories *i* and *i+1*. β_i = item locations of adjacent categories *i* and *i+1*. $\bar{\tau}$ = average item threshold. LI_{IRF} = item location based on the item response function

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*: American Educational Research Association.

Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review, 8*(1), 77-100. doi:10.1016/0272-7358(88)90050-5

Behrendt, L. (2012). *Indigenous Australia for dummies*. Milton, QLD: John Wiley & Sons.

Béteille, A. (1998). The Idea of Indigenous People. *Current Anthropology, 39*(2), 187-192. doi:10.1086/204717

Bin-Sallik, M. A. (1990). *Aboriginal tertiary education in Australia: How well is it serving the needs of Aborigines*. Adelaide: SACE.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203.

Bowen, J. R. (2000). Should we have a universal concept of 'indigenous peoples' rights'?: Ethnicity and essentialism in the twenty-first century. *Anthropology Today, 16*(4), 12-16. doi:10.1111/1467-8322.00037

Brave Heart, M. Y. H., Chase, J., Elkins, J., & Altschul, D. B. (2011). Historical trauma among indigenous peoples of the Americas: Concepts, research, and clinical considerations. *Journal of Psychoactive Drugs, 43*(4), 282-290.

Brown, A., Scales, U., Beever, W., Rickards, B., Rowley, K., & O'Dea, K. (2012). Exploring the expression of depression and distress in aboriginal men in central Australia: a qualitative study. *BMC Psychiatry, 12*(1), 97.

Brown, A. D., Mentha, R., Rowley, K. G., Skinner, T., Davy, C., & O'Dea, K. (2013). Depression in Aboriginal men in central Australia: adaptation of the Patient Health Questionnaire 9. *BMC Psychiatry, 13*, 271. doi:10.1186/1471-244x-13-271

Budyartati, S. (2015). INDIGENEOUSASI ILMU SOSIAL SEBAGAI DASAR PENDIDIKAN KARAKTER BANGSA DALAM PERSPEKTIF PHILOSHOPIS. *JURNAL SOCIA, 11*(02).

Burnette, C. (2015). Historical Oppression and Intimate Partner Violence Experienced by Indigenous Women in the United States: Understanding Connections. *Social Service Review, 89*(3), 531-563. doi:10.1086/683336

Burnette, C. E., & Figley, C. R. (2016). Historical Oppression, Resilience, and Transcendence: Can a Holistic Framework Help Explain Violence Experienced by Indigenous People? *Social Work, 62*(1), 37-44. doi:10.1093/sw/sww065

Cook, D. A., & Beckman, T. J. (2006). Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *The American Journal of Medicine, 119*(2), 166.e167-166.e116. doi:10.1016/j.amjmed.2005.10.036

Cooper, F. (2005). Colonialism in question. *Theory, knowledge*.

Cuthbert, B. N., & Kozak, M. J. (2013). Constructing constructs for psychopathology: The NIMH research domain criteria. *Journal of Abnormal Psychology, 122*(3), 928-937. doi:10.1037/a0034028

Dove, M. R. (2006). Indigenous People and Environmental Politics. *Annual Review Of Anthropology, 35*(1), 191-208. doi:10.1146/annurev.anthro.35.081705.123235

Downing, L. (2015). Heteronormativity and Repronormativity in Sexological "Perversion Theory" and the DSM-5's "Paraphilic Disorder" Diagnoses. *Archives of Sexual Behavior, 44*(5), 1139-1145. doi:10.1007/s10508-015-0536-y

Drescher, J. (2015). Out of DSM: Depathologizing Homosexuality. *Behavioral Sciences, 5*(4), 565-575. doi:10.3390/bs5040565

Dudgeon, P., Rickwood, D., Garvey, D., & Gridley, H. (2014). A history of Indigenous psychology. In H. M. a. R. W. Pat Dudgeon (Ed.), *Working together: Aboriginal and Torres Strait Islander mental health and wellbeing principles and practice* (2nd ed., pp. 39-54). Canberra, Australia: Commonwealth of Australia.

Follette, W. C., & Houts, A. C. (1996). Models of scientific progress and the role of theory in taxonomy development: A case study of the DSM. *Journal of Consulting and Clinical Psychology, 64*(6), 1120-1132. doi:10.1037//0022-006x.64.6.1120

Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: an introduction*. Thousand Oaks, CA: Sage.

Garuba, H. (2013). On animism, modernity/colonialism, and the African order of knowledge: provisional reflections. *Contested Ecologies*, 42.

Garvey, D. (2008). *Review of the social and emotional wellbeing of Indigenous Australian peoples.* Retrieved from

Ginty, A. T. (2013). Psychometric Properties. In *Encyclopedia of Behavioral Medicine* (pp. 1563-1564): Springer New York.

Goldstein, S., Princiotta, D., & Naglieri, J. A. (2015). Handbook of intelligence. *Evolutionary theory, historical perspective, and current concepts. New York, NY: Springer*.

Gone, J. P. (2013). Redressing First Nations historical trauma: Theorizing mechanisms for indigenous culture as mental health treatment. *Transcultural Psychiatry, 50*(5), 683-706. doi:10.1177/1363461513487669

Guillemin, F., Bombardier, C., & Beaton, D. (1993). Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *Journal of Clinical Epidemiology, 46*(12), 1417-1432.

Haag, A. M., Boyes, A., Cheng, J., MacNeil, A., & Wirove, R. (2016). An introduction to the issues of cross-cultural assessment inspired by Ewert v. Canada. *Journal of Threat Assessment and Management, 3*(2), 65.

Hare, R. D. (2003). Manual for the revised psychopathy checklist. In: Toronto, ON, Canada: Multi-Health Systems.

Haslam, N. (2000). Psychiatric categories as natural kinds: Essentialist thinking about mental disorder. *Social Research*, 1031-1058.

Hawksley, C. M. (2001). Administrative colonialism: district administration and colonial'middle management'in Kelantan 1909-1919 and the Eastern Highlands of Papua New Guinea 1947-1957.

Hodgson, D. L. (2002). Introduction: Comparative Perspectives on the Indigenous Rights Movement in Africa and the Americas. *American Anthropologist, 104*(4), 1037-1049. doi:10.1525/aa.2002.104.4.1037

Ingersoll-Dayton, B. (2011). The development of culturally-sensitive measures for research on ageing. *Ageing Soc, 31*(3), 355-370. doi:10.1017/S0144686X10000917

Johnston, R. J. (1981). The dictionary of human geography. Basil Blackwell.

Jones, K., Parker, E., Mills, H., Brennan, D., & Jamieson, L. (2014). Development and psychometric validation of a Health Literacy in Dentistry scale (HeLD). *Community Dental Health, 31*(1), 37-43.

Kirmayer, L. J. (1991). The Place of Culture in Psychiatric Nosology: Taijin Kyofusho and DSM-III-R. *The Journal of Nervous and Mental Disease, 179*(1), 19-28. doi:10.1097/00005053-199101000-00005

Kirmayer, L. J., Gone, J. P., & Moses, J. (2014). Rethinking Historical Trauma. *Transcultural Psychiatry, 51*(3), 299-319. doi:10.1177/1363461514536358

Kleinman, A. (1987). Anthropology and psychiatry. The role of culture in cross-cultural research on illness. *The British Journal of Psychiatry, 151*(4), 447-454. doi:10.1192/bjp.151.4.447

Kleinman, A., & Kleinman, J. (1991). Suffering and its professional transformation: toward an ethnography of interpersonal experience. *Culture, Medicine and Psychiatry, 15*(3), 275-275. doi:10.1007/bf00046540

Klineberg, O. (1935). Race differences.

Kowal, E., Gunthorpe, W., & Bailie, R. S. (2007). Measuring emotional and social wellbeing in Aboriginal and Torres Strait Islander populations: an analysis of a Negative Life Events Scale. *International Journal for Equity in Health, 6*(1), 18.

Lachman, M. E., & Weaver, S. L. (1998). The sense of control as a moderator of social class differences in health and well-being. *Journal of Personality and Social Psychology, 74*(3), 763.

Lee, E.-H. (2012). Review of the psychometric evidence of the perceived stress scale. *Asian Nursing Research, 6*(4), 121-127.

Lilienfeld, S. O. (2017). Through a glass, darkly: Microaggressions and psychological science. Perspectives on Psychological Science, 12(1), 178-180.

Lindroth, M., & Sinevaara-Niskanen, H. (2013). At the Crossroads of Autonomy and Essentialism: Indigenous Peoples in International Environmental Politics. *International Political Sociology, 7*(3), 275-293. doi:10.1111/ips.12023

LoBiondo-Wood, G., & Haber, J. (2010). Reliability and validity. *Nursing research: Methods and critical appraisal for evidence-based practice*, 285-308.

Loomba, A. (1988). *COLONIALISM/POSTCOLONIALISM*: Taylor & Francis.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 14(2), 117-138.

McCrae, R. R., & Allik, I. (2002). *The five-factor model of personality across cultures*: Springer Science & Business Media.

McCuish, E. C., Mathesius, J. R., Lussier, P., & Corrado, R. R. (2018). The cross-cultural generalizability of the Psychopathy Checklist: Youth version for adjudicated indigenous youth. *Psychological Assessment, 30*(2), 192.

McNally, R. J. (2012). The Ontology of Posttraumatic Stress Disorder: Natural Kind, Social Construction, or Causal System? *Clinical Psychology: Science and Practice, 19*(3), 220-228. doi:10.1111/cpsp.12001

Milfont, T. L., & Fischer, R. (2015). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of psychological research, 3*(1), 111-130.

Miller, K. E., Omidian, P., Quraishy, A. S., Quraishy, N., Nasiry, M. N., Nasiry, S., . . . Yaqubi, A. A. (2006). The Afghan Symptom Checklist: A culturally grounded approach to mental health assessment in a conflict zone. *American Journal of Orthopsychiatry, 76*(4), 423-433. doi:10.1037/0002-9432.76.4.423

Mishra, D. P. (2013). Construct Validity in Cross Cultural Management Research: Classical Test Theory and Latent Trait Theory Approaches.

Mooi, E., & Sarstedt, M. (2010). Data. In *A Concise Guide to Market Research* (pp. 25-44): Springer Berlin Heidelberg.

Ndlovu, M. (2013). Mobilising history for nation-building in South Africa: A decolonial perspective. *Yesterday and Today*(9), 00-00.

Okello, E., & Musisi, S. (2006). Depression as a clan illness (eByekika): an indigenous model of psychotic depression among the Baganda of Uganda. *J World Cult Psychiatry Res Rev, 1*(2), 60-73.

Osborne, R. H., Batterham, R. W., Elsworth, G. R., Hawkins, M., & Buchbinder, R. (2013). The grounded psychometric development and initial validation of the Health Literacy Questionnaire (HLQ). *BMC Public Health, 13*(1), 658.

Paradies, Y. C. (2006). Beyond Black and White. *Journal of Sociology, 42*(4), 355-367. doi:10.1177/1440783306069993

Pearlin, L. I., Menaghan, E. G., Lieberman, M. A., & Mullan, J. T. (1981). The stress process. Journal of Health and Social Behavior, 337-356.

Philip, A., & Ojo, B. O. (2017). Application of Item Characteristic Curve (ICC) in the selection of test items. British Journal of Education, 5(2), 21-41.

Porteus, S. D. (1950). The Porteus Maze Test and intelligence. In. Palo Alto, CA: Pacific Books.

Porteus, S. D. (1961). "Comment". *Current Anthropology, 2*.

Prince, M. (2008). Measurement validity in cross-cultural comparative research. *Epidemiologia e Psichiatria Sociale, 17*(03), 211-220. doi:10.1017/s1121189x00001305

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517.

Ranzijn, R., McConnochie, K. R., & Nolan, W. (2009). *Psychology and Indigenous Australians: Foundations of cultural competence*. South Yarra, Victoria: Palgrave Macmillan Australia.

Robinson, M., Johnson, A. M., Walton, D. M., & MacDermid, J. C. (2019). A comparison of the polytomous Rasch analysis output of RUMM2030 and R (ltm/eRm/TAM/lordif). BMC medical research methodology, 19(1), 36.

Russo, J., & Venâncio, A. T. A. (2006). Classificando as pessoas e suas perturbações: a "revolução terminológica" do DSM III. *Revista Latinoamericana de Psicopatologia Fundamental, 9*(3), 460-483. doi:10.1590/1415-47142006003007

Sanislow, C. A., Pine, D. S., Quinn, K. J., Kozak, M. J., Garvey, M. A., Heinssen, R. K., . . . Cuthbert, B. N. (2010). Developing constructs for psychopathology research: Research domain criteria. *Journal of Abnormal Psychology, 119*(4), 631-639. doi:10.1037/a0020909

Sessle, B. J. (2017). Pain in relation to oral health. *Journal of Oral and Facial Pain and Headache*, 31(1), 5-6.

Slade, G. D., & Spencer, A. J. (1994). Development and evaluation of the oral health impact profile. *Community Dental Health, 11*(1), 3-11.

Stannard, D. E. (1999). Honoring racism: the professional life and reputation of Stanley D. Porteus. *Social Process in Hawai ʻi, 39*, 85-125.

Steenkamp, J.-B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of consumer research, 25*(1), 78-90.

Streiner, D. L. (2003). Being Inconsistent About Consistency: When Coefficient Alpha Does and Doesn't Matter. *Journal of Personality Assessment, 80*(3), 217-222. doi:10.1207/s15327752jpa8003_01

Summerfield, D. (2008). Cross-Cultural Perspectives on the Medicalization of Human Suffering. *In Posttraumatic Stress Disorder: Issues and Controversies* (pp. 233-245). Wiley.

Ewert v. Canada, (2018).

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*: American Psychological Association.

Whitaker, R. (2005). Anatomy of an epidemic: Psychiatric drugs and the astonishing rise of mental illness in America. *Ethical Human Sciences and Services, 7*(1), 23-35.

Wolfe, P. (2006). Settler colonialism and the elimination of the native. *Journal of Genocide Research, 8*(4), 387-409. doi:10.1080/14623520601056240

Young, R. (2009). The stolen generation. *Psychotherapy in Australia, 16*(1), 59.