Implicit Bias Training for New Zealand Medical Students using Cognitive Bias Modification: An Outline of Material Development

Che-Wei Hsu[1] and Zaine Akuhata-Huntington[2]

[1]Psychological Medicine, Dunedin School of Medicine, University of Otago
[2]Kōhatu Centre for Hauora Māori, Dunedin School of Medicine, University of Otago

Health inequity for marginalized groups increases the risk of developing health problems. Healthcare providers' unconscious bias contributes to this inequity. A novel bias modification method—called Cognitive Bias Modification for stereotype (CBM-S)—is a digital tool designed to be used in conjunction to existing bias training to address medical students' biases toward Māori (an indigenous population in Aotearoa New Zealand). CBM-S encourages non-stereotypical interpretations through relevant and specific text-based scenarios and has been tested against control. To improve the scenario's relevance and specificity, here, we adopted a user-centered approach to materials development for CBM-S that involved iterative inputs from medical students and Māori participants to achieve our objectives. We outlined the material development process for CBM-S to provide a guide for researchers and educators in developing CBM-S training.

Keywords: *Cognitive bias modification, implicit bias training, medical student bias, health education, health inequity, indigenous healthcare*

INTRODUCTION

Health inequity refers to the unequal access to health resources and systematic disparity in patients' healthcare experience and overall health status (McCartney et al., 2019). It is a significant global issue, especially for marginalized groups (e.g., indigenous populations, rainbow communities, older persons, people with disabilities; Baah et al., 2019). Health inequity is a risk factor for developing both mental and physical health problems (Paradies, 2006). This is due in part to barriers in accessing healthcare services (Ellision-Loschmann & Pearce, 2006; Harris et al., 2012) and patient's negative experiences during the health-seeking process (Cooper et al., 2012; Harris et al., 2012). It has been well documented that healthcare providers' implicit (unconscious) bias contributes to such barriers (Blair IV et al., 2001; Gonzalez et al., 2014; Smedley et al., 2003; van Ryn & Fu, 2003; White III, 2011). Implicit bias is often difficult to address; conventional bias training is often cognitively and temporally demanding and requires effortful introspection and frequent and long-term education (Forscher et al., 2019). The present paper outlined the material development process for a novel bias training tool called Cognitive Bias Modification for Stereotype (CBM-S). CBM-S is a digital, self-administered training tool that addresses New Zealand medical students' *interpretation bias* of common healthcare scenarios involving Māori patients (Hsu & Akuhata-Huntington, 2024). CBM-S targets bias at the implicit level and is designed to complement existing bias training.

Content Specificity in Interpretation Bias

In information processing theory, scholars have shown that an ambiguous social situation can elicit multiple interpretations, and that subsequent responses to that situation reflect the selected interpretation (Atkinson & Shiffrin, 1968; Lachter et al., 2004). An interpretation bias occurs when an individual consistently follows a similar thought pattern (Nisbett & Ross, 1980), for example, interpreting that Māori patients who ask a series of questions about treatment reflects their lack of education. In many studies, researchers have demonstrated *content specificity* in interpretation bias; that is, interpretation bias is strongest when the situation matches the individual's beliefs (Mathews & MacLeod, 1994; Savulich et al., 2015; Savulich et al., 2017; von Hippel et al., 1997; Yiend & Mackintosh, 2004). For instance, von Hippel et al. (1997) captured participants' gender and ethnic bias by measuring their interpretation of common situations involving females and African Americans, respectively. In that study, male undergraduates completed an implicit bias measure—the Linguistic Intergroup Bias (LIB; Maass et al., 1989), which included text-based scenarios that could elicit stereotypical interpretations about the target group (i.e., women or African Americans). To assess for gender bias, students were randomly assigned to read one of two scenarios involving an ambiguous situation (e.g., *unable to change a blown fuse*). These scenarios were designed to reflect gender stereotype-congruent situations. Each story included either a male name (e.g., *James*) or a female name (e.g., *Molly*). Students provided similarity ratings of four short interpretations of each scenario. The statements varied in their degree of abstraction [e.g., *James/Molly is unable to change the fuse* (concrete) to *James/Molly is dependent* (abstract)]. A higher average rating on abstract interpretations of scenarios with a female name indicated stronger gender bias against women.

To assess ethnic bias, ambiguous scenarios were presented to students (e.g., scenarios involving a slam dunk champion or a spelling-bee winner). These scenarios were designed to reflect ethnic stereotype-congruent situations. Each story was either accompanied by a

photograph of an African American or a Caucasian. Again, students provided similarity ratings of four short interpretations of each scenario that varied in the level of abstraction (e.g., *Johnson performs 360-degree slam-dunk* (concrete) to *Johnson is athletic* (abstract)]. In general, 43% of students showed gender bias towards women; 45% of students showed ethnic bias towards African Americans. Taken together, these results highlighted the use of content-specific experimental materials in capturing biased interpretations of relevant scenarios.

User-Centered Development of Bias Training

Building on from this notion of *content specificity*, employing a user-centered development approach may maximize that specificity and relevance of materials to be used in bias reduction training, such as CBM-S. CBM-S adopts methods of an emerging class of bias modification training called CBM (Hsu, 2023), which was conventionally designed as a therapeutic for various mental health concerns (Hirsch et al., 2018; Koster & Horrelbeke, 2015; Woud et al., 2017). In a clinical application of CBM, researchers have also demonstrated the importance of using content-specific materials in capturing and modifying interpretation bias in people with depression (Lamberton & Oei, 2008). That is, training for depression should include materials that invite *negative* interpretation bias but encourage a *positive* resolution. To improve the content-specificity of materials used in CBM training, Hsu et al. (2023) adopted a user-centered approach to materials development by working collaboratively with experienced clinical psychologists and experts by experience (people with first-hand experience of mental health concerns). Together, these contributors co-created CBM materials that were relevant to their everyday experiences to better capture and modify unhelpful interpretation bias.

In the present paper, we illustrated a similar user-centered development approach to create content for a single-session CBM-S. CBM-S has recently been tested against control in medical education to address students' ethnic bias towards indigenous patient groups in NZ, and results have been promising (see Hsu & Akuhata-Huntington, 2024). Bias modification is achieved by using implicit inferential learning via a word task. To illustrate, CBM-S involves presenting medical students with a set of *user-generated*, *content specific* materials (i.e., healthcare scenarios involving Māori). The final word of each scenario is first omitted to create ambiguity with the purpose of eliciting multiple interpretations. To create the word task, a fragment of the final word of each scenario is revealed ("*…overgen-ralis-d"*). Students enter the first letter of that word to complete the word task, which resolves the ambiguity of the scenario in a *non-ethnic stereotype* manner. This prompted word task is not only effective in guiding the training activity, but more importantly, it is a crucial method to *implicitly* induce a less biased response. According to learning theories, guiding users to respond in a forced direction through cued prompts leads to improved learning outcomes than to provide information in an open-ended manner (Adesope et al., 2017; Rowland, 2014).

CBM-S has several advantages over existing bias training in that it offers an alternative, cost-effective approach that can be self-administrated and delivered through digital platforms. This feature enhances its accessibility and scalability, making it a more feasible option for a wider range of individuals. Beyond this, CBM-S takes a different approach to bias modification. More specifically, existing implicit bias training methods such as metacognition, fact provision, group discussions, and counterstereotype exemplars, involve passive information delivery about social groups, challenging preconceived notions (Joy-Gaba et al., 2010) and promoting awareness of biases concerning those groups (Sabin et al., 2022). CBM-S targets interpretation bias at an *implicit* level, aiming to address biases in a more nuanced and indirect manner through the word task and ambiguous scenarios previously described.

The present paper provides the detailed development and evaluation process of CBM-S content and assessment materials, which had the following objectives and methodology:

- **Objective 1:** Adopt a user-centered approach, with input from medical students and Māori participants to create and evaluate content-specific material to be used in CBM-S and tested against control. The methodology for Objective 1 includes two stages: content creation and evaluation.
- **Objective 2:** Obtain reliability ratings of interpretation bias assessments and user feedback of CBM-S. The methodology for Objective 2 involves one stage of collecting user feedback and reliability data for two interpretation bias assessments. In the present paper, we adopted two interpretation bias measures that are commonly used to assess interpretation bias in CBM studies: the Scrambled Sentence Task (SST; Rude et al., 2003; Rude et al., 2022) and Similarity Rating Task (SRT; Eysenck et al., 1991) The SST and SRT are reliable measures of interpretation bias in psychopathology (e.g., SST: Cronbach's α of .79, Würtz et al., 2022; SRT: α = .82, Berna et al., 2011). Please see below in the method section for more details on the development of these two measures.

METHOD and RESULTS

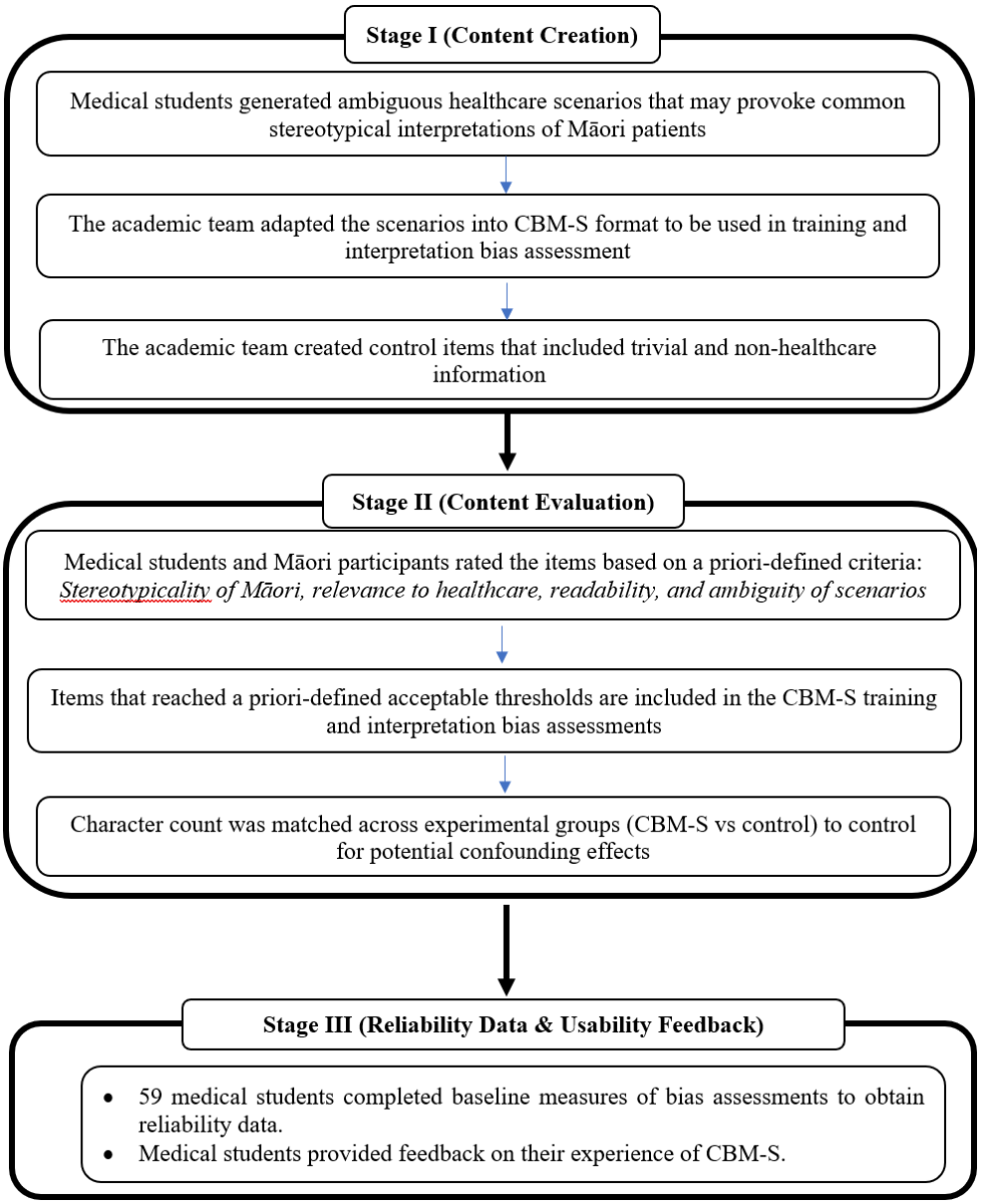
The development and evaluative process of materials for Cognitive Bias Modification for Stereotype (CBM-S) involved varying inputs from medical students and Māori participants (see Figure 1, for a schematic of a three-stage iterative development process). The CBM-S program of work received ethical approval from the University of Otago Ethics Committee (reference: 22/063). We obtained informed consent from all participants in the study.

Stage I: Content Creation

Healthcare Scenario

We invited a group of medical students (N = 5) to each generate 20 common healthcare scenarios involving Māori to be used for developing CBM-S items. From the 100 scenarios, we aimed to develop 80 CBM-S items and 20 interpretation bias assessment items. To do this, we provided a brief introduction about CBM-S, followed by

Figure 1. *Schematic of the material development process*

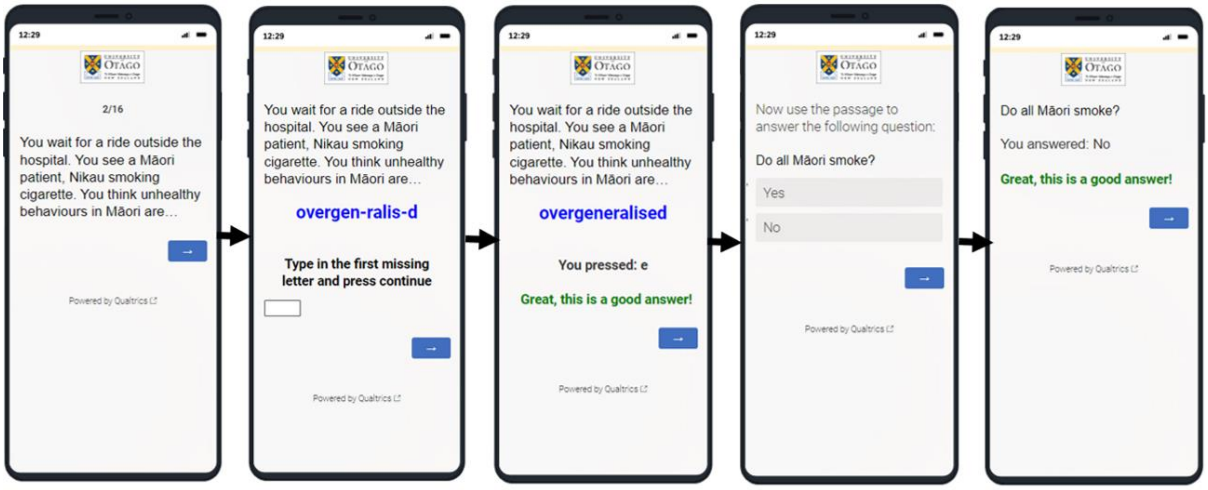


three exemplars of healthcare scenarios involving Māori (e.g., Māori refuses western medicine). Each participant was asked to '*write down 20 scenarios common to healthcare settings that you think may trigger ethnically biased thoughts about Māori patients and to provide both a stereotype/prejudice/discrimination and non-stereotypical/neutral interpretation of that scenario.*' Pre- and post- this task, we asked students to rate their mood using a Visual Analog Scale to monitor for any abrupt mood changes in sadness, anger, distress, anxiety (Gould et al., 2001); we did not find any significant mood changes.

Training Items

Following scenario creation, the academic team—consisting of one Māori researcher and one non-Māori researcher—randomly selected 80 of the 100 scenarios and adapted them into standardized CBM format to be used as training items for CBM-S. CBM format included a three-sentence passage followed by a yes/no question. We removed one or more letters of the final word of each item to create the word task (we removed generally vowels, depending on the length of the word). Participants completed the final word to resolve the ambiguity in a non-ethnic stereotypical manner. We adopted the student-generated *non-stereotypical/neutral* interpretation of each scenario to create the final word for each item. See Figure 2a for an example training item. Gender-neutral pronouns (i.e. they, them) and second-person pronouns (i.e. you, yours) were adopted to generate a first-person perspective of the scenarios. First-person perspective encourages active mental participation in text-based narratives that promotes readers to develop richer perspective-taking (Brunyé et al., 2011), which has been shown to improve the modification effects of the CBM method (Holmes & Mathews, 2005; Holmes et al., 2006).

Figure 2a. *An example CBM-S training item.*



Three additional features were included in the training items. First, to encompass a broader experience of ethnic bias, we arranged the training items into five domains of modern racism toward Māori (Satherley & Sibley, 2018):

1. Negative affect (*negative feelings toward Māori*)
2. Anxiety (*feeling anxious during encounters with Māori*)
3. Denial of historical reparation (*past injustice is non-transferrable to the present day*)
4. Denial of contemporary injustice (*discrimination is not a present-day issue*)
5. Symbolic exclusion (*Māori culture is not representative of Aotearoa New Zealand*)

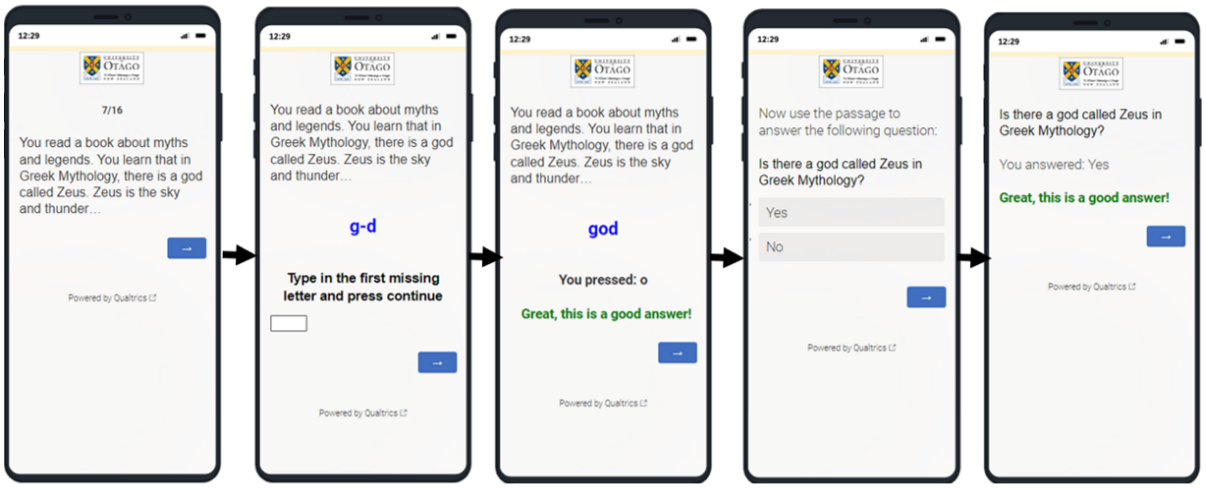
Second, we included dedicated verbs in the final sentence of each item to reflect the pipeline of cognitive information processing, starting from higher-level thoughts to more stable schemata/beliefs (Rumelhart, 1984). Verbs used to reflect processing at each level included: 1) 'think, sense, imagine' (*top level*), 2) 'assume, presume, suppose' (*intermediate level*), and 3) 'believe, are sure, know' (*bottom level*).

Third, researchers have shown that, to achieve effective group bias modification, it is important to both individuate traits and characteristics to the person (Lebrecht et al., 2009; Lee et al., 2017; e.g., *DeAndre is kind*) as well as generalize them to a social category (Gawronski et al., 2018; e.g., *Blacks are kind*). In order to reflect this in CBM-S, initial training blocks used specific Māori names (e.g., Mrs Wiremu) before progressively including group labels (i.e., Māori patients).

Control Items

The academic team created 80 control items to test the modification effects of CBM-S. The aim was to create control items with contents that were unambiguous, non-medical, and non-cultural. We adopted trivial passages from https://mommypoppins.com/kids/fun-facts-for-kids-random-fun-facts), and created non-medical and non-cultural everyday scenarios [e.g., '*It is raining so you carry an umbrella with you. You walk to the bus stop and wait for the bus. You see there are already many people…*(waiting)']. The format of control items was identical to the training items (i.e., a three-sentence

Figure 2b. *An example control item.*



scenario using gender-neutral second-person pronouns with a fragmented final word and a follow-up yes/no question). The fragmented word is used to complete the passage—as opposed to resolving ambiguity. See Figure 2b for an example control item.

Interpretation Bias Measures

The academic team adopted the remaining 20 of the 100 student-generated healthcare scenarios to create two interpretation bias measures that are commonly in CBM studies—Scrambled Sentence Task (SST) and Similarity Rating Task (SRT). Two sets of SSTs and SRTs were developed for pre- and post-assessment of interpretation bias in CBM-S. Interpretation bias measures only included specific Māori names to promote the implicit processing of assessment items (Harris et al., 2016; Paradies et al., 2014).

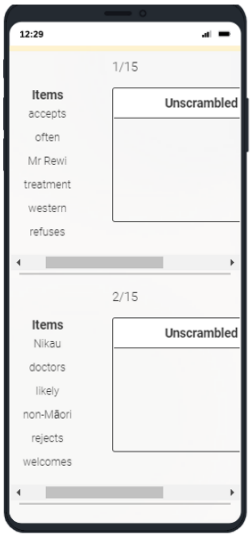
Bias Measure: Scrambled Sentences Task

The SST consisted of scrambled sentences of six words each, including two critical words that were adopted from student-generated interpretations of healthcare scenarios. One word depicted a common stereotype of Māori patients; the other word opposed that depiction in a non-stereotypical/neutral manner. An example of a scrambled sentence is: '*habits / engages / Miss Ropata / in / **unhealthy** / **good**,*' (critical words in bold). Using one of two critical words, participants form either a five-word non-stereotypical interpretation ['*(Miss Ropata) engages in **good** habits*'] or stereotypical interpretation ['*(Miss Ropata) engages in **unhealthy** habits*']. See Figure 3a for an example SST item.

Bias Measure: Similarity Rating Task

The SRT (also known as the Ambiguous Scenario Test) consisted of two parts—encoding and recognition. Encoding items included a three-sentence scenario using gender-neutral second-person pronouns followed by a fragmented final word to induce the ambiguity of the passage; a follow-up yes/no question reinforces that ambiguity. We created titles for each scenario to be used as a cued reminder of the related scenario during the recognition phase. Recognition items included two short statements associated with each encoding item. We adopted student-generated interpretations of healthcare scenarios to develop the two short statements—one statement provided an explanation of the scenario that is consistent with common stereotypes of Māori patients; the other statement explained the scenario in a non-stereotypical/neutral manner. See Figure 3b for an example SRT item.

Figure 3a. *An example SST assessment item.*



Stage II: Content Evaluation

Procedure

Another group of 10 raters (medical students and Māori) evaluated the items created in Stage I. One medical student was unable to complete all of the ratings and did not respond to subsequent communications, resulting in nine final raters (n = 4 medical students; n = 5 Māori participants). Input from each contributor varied according to the task required and is discussed below. Training and control items were interleaved in random order to blind raters from the type of item (i.e., training or control items); for assessment items, stereotype, and non-stereotype statements were presented in random order.

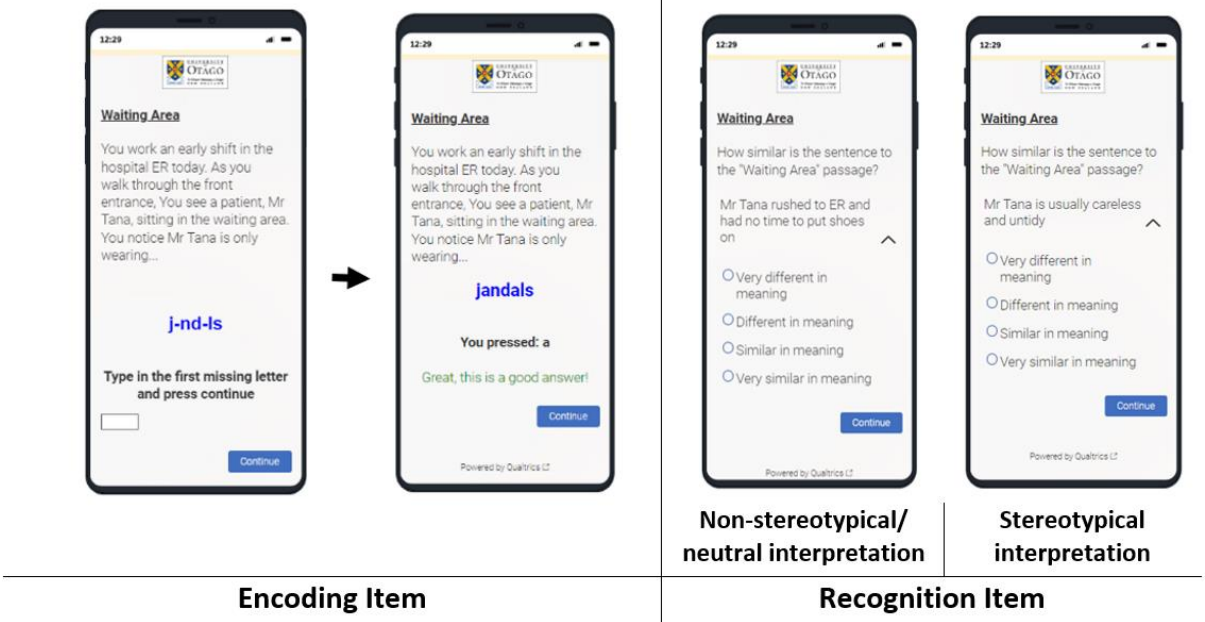
Content Specificity: Training Items

For the purpose of rating, we adopted student-generated interpretations of healthcare scenarios to create two final fragmented words for each scenario—one word resolved the ambiguity of the training item in a non-stereotypical/neutral manner; the other word resolved the ambiguity in a manner that is consistent with common stereotypes of Māori patients. Medical students rated the training items based on a set of a priori-defined criteria (*stereotypicality towards Māori; relevance to a healthcare setting; readability and ambiguity of the scenario*) using a 7-point scale [1 = not at all… to 7 = completely…(*criterion*)]. Māori participants rated the training items on the stereotypical criterion only to reflect a user-centered development approach and promote content specificity. More specifically, to promote an effective and culturally sensitive training approach, it is important to include items that Māori consider to be stereotyping their culture and ethnicity. Items that were excluded based on medical students' ratings of stereotypicality were included if Māori raters had rated those items as stereotypical (21% of items met this criterion). Here is one example of an item that met this criterion: '*While you examine a Māori patient, you tilt their head to the side. As you do this, they appear to move their head away from you. You presume that they are…(uncomfortable/distrustful)*.' Thresholds were set, a priori, to determine acceptable training items: *stereotypical resolution* (≥ 5*); non-stereotypical resolution* (≤ 3*); readability* (≥ 5*); relevance* (≥ 5*); ambiguity* (≥ 2 counts of *'Yes'* responses).

Content Specificity: Control Items

Ratings for the control items followed an identical procedure to that of the training items. For the purpose of rating, two final fragmented words were created. Thresholds were set, a priori, to determine acceptable control items: *stereotypicality of the two fragmented words* (≤ 3*); readability* (≥ 5*); relevance* (≤ 3*);*

Figure 3b. *An example SRT assessment item.*



ambiguity (≤ 1 count of *'Yes'* responses).

Interpretation Bias Measures

Medical students rated 20 SRT encoding items and recognition items based on a set of a priori-defined criteria using a 7-point scale [1 = not at all… to 7 = completely…(*criterion*)]. For encoding items, medical students provided ratings of '*relevance to a healthcare setting'* and '*readability of the scenario*;' for recognition items, both medical students and Māori participants provided ratings of '*stereotypicality.'* Thresholds were set, a priori, to determine acceptable items to be used in interpretation bias assessments: *stereotypical interpretation* (≥ 5*); non-stereotypical interpretation* (≤ 3*); readability* (≥ 5*); relevance* (≥ 5*).* We did not collect ratings of SST items due to the similarity of item contents between SST and SRT as a result of using the same set of student-generated scenarios from Stage I.

Name Ratings

The Māori names to be used in the final set of CBM-S training and assessment items were selected based on user ratings using a 7-point scale (*1 = extremely common non-Māori name to 7 = extremely common Māori name*). The same group of medical students and Māori participants evaluated 16 Māori names and 14 English names that were randomly selected from a list of popular names in New Zealand between 2019-2021 (https://smartstart.services.govt.nz/news/baby-names-maori). Names that reached a priori-defined acceptable value of (≥ 5) were included in the final set of CBM-S items (Note that only the finalized version of CBM-S included the names that were selected here. Materials used for rating during Stage II included group labels (i.e., Māori) and Māori names).

Results

Following the collection of user ratings, we reviewed and refined the items based on a priori-defined acceptable rating values for each criterion and identified any systematic differences between items. This detailed review process is outlined below.

Content specificity: Training and Control Items

First, we examined the ratings for each training and control item independently. Items that fell below a priori-defined acceptable value for each rating criterion were excluded from the final set of CBM-S training and control items. Fifty-nine CBM-S training items and all control items reached acceptable ratings for each criterion

Table 1. *Mean (SD) item ratings (training; control; assessment)*

	Students Ratings (N=4) Mean (*SD*)			Māori Ratings (N=5) Mean (*SD*)		
	Training Items	Control Items	Assessment Items	Training Items	Control Items	Assessment Items
NonStereotypical Interpretation	1.25 (0.64)	1.00 (0.00)	1.27 (0.71)	1.60 (1.38)	1.60 (1.20)	2.04 (1.75)
Stereotypical Interpretation	5.71 (1.26)	1.00 (0.00)	4.97 (2.06)	6.55 (1.06)	1.60 (1.20)	6.21 (1.61)
Relevance to Healthcare	5.98 (1.19)	0.80 (0.47)	6.48 (1.24)	N/A	N/A	N/A
Readability	6.86 (0.58)	6.76 (0.64)	6.58 (1.05)	N/A	N/A	N/A

Table 2. *Participant Descriptive Data*

	Medical Students (N = 59)
Age (Years) Mean (*SD*)	21.59 (1.97)
Ethnicity	
NZ European	31
Māori	8
Pasifika	1
Chinese	9
Other (e.g., Dutch, Japanese)	10
Gender (female:male:non-binary)	39:19:1
Year in Medicine (ELM:ALM)	34:25

(contact author for the complete set of items). Inter-rater reliability on each rating criterion for training items were: $ICC_{non\text{-}stereotype}$ = .42; $ICC_{stereotype}$ = .57; $ICC_{relevance}$ = .72. We do not report ICC scores for ratings of control items and the readability and ambiguity criteria of training items, as might be expected, there was limited variability in these data to make meaningful interpretations of ICC values (Bobak et al., 2018).

Next, we examined any group differences as a function of the mean rating of each criterion (*stereotypical and non-stereotypical resolution, relevance, readability, ambiguity*).

Table 1 shows the mean ratings of the final set of CBM-S training and control items.

For these analyses, we combined ratings of Māori participants and medical students (see Figure 4). Consistent with our aim, Paired Samples t-tests between the training and control groups revealed systematic differences in:

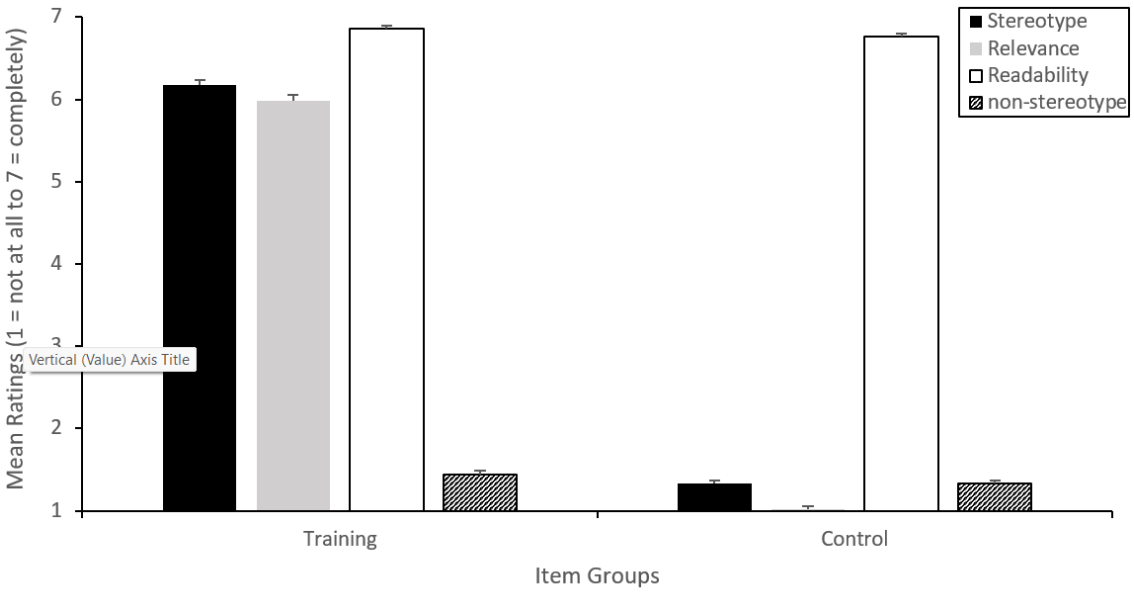
- mean ratings of *stereotypical resolution* between training (M = 6.18, SD = 1.23) and control items (M = 1.33, SD = .94), $t(530)$ = 80.48, $p < .001$.
- mean ratings of *relevance to healthcare* between training (M = 5.98, SD = 1.19) and control items (M = 1.02, SD = .13), $t(234)$ = 63.37, $p < .001$.
- the proportion of students who thought training items (86%) and control items (33%) were *ambiguous*, χ^2 (1) =34.33, $p < .001$.

That is, relative to the control items, training items provided a more consistent depiction of common stereotypes of Māori patients, were more relevant to a healthcare context, and were more ambiguous. On the other hand, we did not find any statistically significant differences in ratings of *readability* between training (M = 6.86, SD = .58) and control items (M = 6.76, SD = .64), $t(234)$ = 1.79, p = .07, and in ratings of *non-stereotypical resolution* between training (M = 1.44, SD = 1.13) and control items (M = 1.33, SD = .94), $t(530)$ = 1.65, p = .10.

In order to match the number of training and control items, we excluded 21 control items based on the item's length, operationally defined as the item's character count. This controls for potential and inadvertent effects of time spent in CBM training (Standage et al., 2009). The Two One-sided Test (TOST) for testing equivalence (Lakens et al., 2018) revealed that the length of training items (M = 134.46, SD = 16.76) was equivalent to that of control items (M = 132.52, SD = 12.55), $t(107)$ = -1.74, p = .04, given an equivalence bound of 6.67 to -6.67 (95% CI).

Interpretation Bias Measures

Figure 4. *Mean ratings of each criterion (excluding ambiguity) across training and control items.*



Next, we examined the ratings for each assessment item independently. Items that fell below a priori-defined acceptable value for each rating criterion were excluded from the final sets of assessment items. Fifteen items reached an acceptable rating for each criterion and were adapted to be used in the final set of SRT and SST items (contact author for complete sets of assessment items). Some items were reworded to create additional content to accommodate for the number of items needed for the bias measures (i.e., the SST consisted of 15 items for each version; the SRT consisted of eight items for each version). Interrater reliability for each of the rating criteria were: $ICC_{non\text{-}stereotype}$ = .68; $ICC_{stereotype}$ = .86; $ICC_{relevance}$ = .52. Again, we do not report ICC scores for ratings of the readability criterion due to floor effect (Bobak et al., 2018).

We then examined any differences in the mean rating of the *stereotypical* and *non-stereotypical resolutions* of the items. Table 1 above shows the mean ratings of the final set of SRT items (recall that we did not collect ratings of SST items due to the similarity of item contents between SST and SRT). For these analyses, we combined ratings of Māori participants and medical students. Consistent with our aim, a Paired Samples t-test revealed a systematic difference in mean ratings of *stereotypical interpretation* (M = 5.66, SD = 1.92) and *non-stereotypical interpretation* (M = 1.70, SD = 1.44), $t(134) = 20.02$, $p < .001$. Specifically, relative to non-stereotypical interpretations, stereotypical interpretations provided a more consistent depiction of common stereotypes of Māori patients. High mean ratings were obtained for items' *relevance to a healthcare setting* and *readability* (see Table 1).

Name Ratings

To obtain a list of names common to Māori to be used in the final set of CBM-S training and assessment materials, we examined the ratings for each name independently. Names that fell below a priori-defined acceptable value for each rating criterion were excluded from the final list of names. Ten names reached the acceptable value, in ascending order of commonality to a Māori name were: *Miss Kahaki, Kiwa, Mr Te Wiata, Mrs Waerea, Mr Tipene, Ms Awatere, Miss Wiremu, Miss Ropata, Nikau, Mr Rewi*. A high degree of reliability was found between raters on name ratings (ICC_{name} = .93). A Paired Samples t-test revealed a difference in mean ratings of these final 10 names (M = 5.30, SD = 1.74) and the names that did not reach acceptable value (M = 2.72, SD = 1.47), $t(298) = 13.49$, $p < .001$.

Stage III: Testing & User Feedback

Participants

As a part of the CBM-S efficacy study (Hsu & Akuhata-Huntington, 2024), we recruited from the University of Otago Medical School 60 medical students in Early Learning in Medicine (ELM: 1st and 2nd year in medicine) and Advanced Learning in Medicine (ALM: 3rd-5th year in medicine). Due to technical issues, one student's data were lost. The final sample included 59 students. Table 2 shows participants' descriptive data.

Procedure

CBM-S testing was hosted on Qualtrics—an online survey platform. After the students received information about the study and provided eConsent, they completed the training and assessments online using either their personal computers or mobile phones/tablets. We obtained user feedback on CBM-S training and examined the reliability of two interpretation bias measures.

User Feedback

A set of questions were included post-training to incorporate user feedback to help refine and improve CBM-S. Medical students rated the training items on the following features using a 7-point scale: **enjoyment**: *1-annoying* to *7-enjoyable*; **clarity:** *1-clear* to *7-confusing*; **interest:** *1-not interesting* to *7-interesting;* **ease of use:** *1-complicated* to *7-easy*. These rating criteria were selected from a 10-scale User Experience Questionnaire based on coverage of the scales and their relevance to CBM-S. The questionnaire was designed to measure users' experience of interactive products (Laugwitz et al., 2008).

Interpretation Bias Measure: SST

Recall that we developed two versions of the SST (SSTv1 and SSTv2) from student-generated healthcare scenarios, with each version consisting of 15 scrambled sentences of six words. Students received one of two versions of SST in a fixed order. Using 5 out of 6 words, students reordered the words to create a grammatically correct sentence. The scrambled sentences were designed to form either a stereotypical or non-stereotypical interpretation using one of two critical words (e.g., '*Ms Waerea likely seeks* ***western/alternative*** *treatment'*); an error occurs when a sentence includes both or none of the critical words or is grammatically incorrect. To reduce deliberate response biases in information processing (Bowler et al., 2012; Rude et al., 2003), the SST included two features. First, students were instructed to '*choose whatever sentence comes to mind and to complete the task as fast as you can as the task is time-limited*.' Students had 5 minutes to complete as many items as possible. Second, students completed the SST user a cognitive load; that is, students were presented with a six-digit number (e.g., 815374) pre-SST and were told '*You will be asked to recall the number later, so please keep the number in your mind while you complete the word task.' S*tudents were asked to recall the same number immediately post-SST.

Using a yes (1) or no (0) dummy coding, two researchers coded all the responses as either *stereotypical*, *non-stereotypical*, or *error*. For example, a sentence that included the 'stereotype' critical word was coded as: (stereotypical = 1; non-stereotypical = 0; error = 0). Kappa scores ranged from moderate to perfect agreement (Landis & Koch, 1977), k = .58–.97, *all $p < .001$*, with only one falling below 0.6. Once all the responses were coded, we summed the scores for each coding category (SUM_stereotypical, SUM_non-stereotypical, SUM_error). An interpretation bias score, ranging from 0% to 100%, was calculated using the following equations:

Bias Score	**= (SUM_stereotypical ÷ Total Items Attempted) x 100**
Non-Bias Score	**= (SUM_non-stereotypical ÷ Total Items Attempted) x 100**

A higher bias score indicated a tendency to interpret scenarios involving Māori in a manner that is congruent to common stereotypes of Māori.

Interpretation Bias Measures: SRT

Recall that SRT consists of two parts—encoding and recognition. During encoding, students received one of the two SRT versions (SRTv1 and SRTv2) in a fixed order. Each version of the SRT consisted of eight items. At the beginning of the encoding phase, students were given the following instructions 'As you *read through each passage, it is important to imagine yourself as the healthcare provider in the situation described*.' For each passage, students were asked to complete the final fragmented word followed by a yes/no question. During the recognition phase, students were presented with two short statements that were related to a passage from the encoding phase. One statement provided an explanation of the passage that is consistent with common stereotypes of Māori patients; the other statement explains the passage in a more helpful/benign manner. Using a 4-point scale (*1 - very different in meaning* to *4 - very similar in meaning)*, students were instructed to rate '*how similar the sentence is to the corresponding passage.*'

We calculated the average score for each type of statement separately to obtain two mean scores: Mean [Stereotype] and Mean [Non-Stereotype]. An interpretation bias score, ranging from +3 to -3, was calculated as follows:

Interpretation Bias = Mean [Stereotype] – Mean [Non-Stereotype]

A more positive score indicated a tendency to interpret scenarios involving Māori in a manner that is congruent to common stereotypes of Māori.

Associations Measure: Implicit Association Test (IAT)

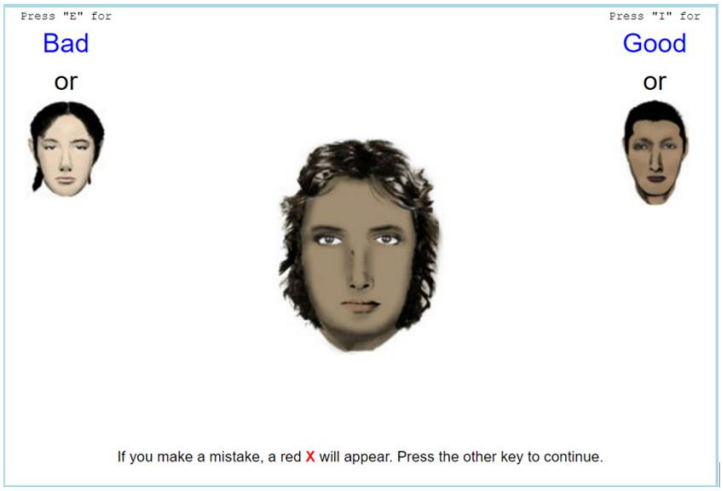
The IAT measures the strength of association between evaluations of targeted categories. It is a widely used measure in social psychology (Greenwald et al., 1998). While we acknowledge the limitations of using skin-tone to measure implicit evaluations of Māori, in the present study, we utilized a readily available online IAT test to assess medical students' automatic evaluations of light and dark skin tones (see Figure 5; Project Implicit, n.d.). During the IAT, students were invited to sort *words* and *pictures of faces* as quickly as possible into one of two categories using two keyboard keys: the 'e' key was pressed to indicate that the *word or picture* belonged to the group on the left; the 'i' key was pressed to indicate that the *word or picture* belonged to the group on the right.

The skin-tone IAT consisted of five parts:

1) **Faces and categories.** Students grouped *faces* of different skin-tone into either 'Dark-Skinned People' (left of screen) or 'Light-Skinned People' (right of screen).
2) **Words and evaluation.** Students grouped different *words* into either 'Bad' (left of screen) or 'Good' (right of screen): ('Good' words: *Delightful, Attractive, Fabulous, Joyful, Adore, Pleasure, Cheer, Appealing*) and ('Bad' words: *Horrific, Sadness, Bothersome, Annoy, Tragic, Selfish, Angry, Despise*).
3) **Category and evaluation combined.** Students grouped both *words and faces* into either the 'Dark-Skinned People/Bad' (left of screen) or 'Light-Skinned People/Good' (right of screen). This step was administered twice.
4) **Faces and categories (with the placement of categories switched).** Students grouped *faces* of different skin-tone into either 'Light-Skinned People' (left of screen) or 'Dark-Skinned People' (right of screen).
5) **Combined categories and evaluations**. Students grouped both *words and faces* into either the 'Light-Skinned People/Bad' (left of screen) or 'Dark-Skinned People/Good' (right of screen). This step was administered twice.

Steps 3 and 5 were counterbalanced in fixed order across the students. For example, if in Step 3, student S1 was shown '***Dark****-Skinned People/****Bad****, **Light****-Skinned People/****Good***,' then in Step 5, S1 was shown '***Light****-Skinned People/****Bad****, **Dark****-Skinned People/****Good***.' The same combination would apply to student S2 in reverse.

Figure 5. *Example skin-tone IAT test*



Note. Adapted from Project Implicit (n.d.).

Table 3. *Medical Students' ratings of the CBM-S training*

Features Mean (SD) ratings Median ratings Rating range (min-max)	**Ratings for each Feature: 7-point Rating Scale (% of Participants)**						
Enjoyment	***1-Annoying***	***2-***	***3-***	***4-***	***5-***	***6-***	***7-Enjoyable***
4.20 (1.52) 4.00 1-7	6.7%	3.3%	20%	30%	20%	13.3%	6.7%
Clarity	***1-Clear***	***2-***	***3-***	***4-***	***5-***	***6-***	***7-Confusing***
3.53 (4.00) 4.00 1-6	13.3%	20%	13.3%	16.7%	26.7%	10%	0%
Interest	***1-Not interesting***	***2-***	***3-***	***4-***	***5-***	***6-***	***7-Interesting***
4.23 (5.00) 5.00 1-7	3.3%	10%	20%	13.3%	36.7%	13.3%	3.3%
Ease of Use	***1-Complicated***	***2-***	***3-***	***4-***	***5-***	***6-***	***7-Easy***
5.13 (5.00) 5.00 2-7	0%	3.3%	6.7%	16.7%	36.7%	20%	16.7%

The average time students spend sorting words and faces accurately into their corresponding categories (i.e., good/bad; dark/light skin tone) were recorded and inferred students' automatic evaluations of light and dark skin tone. Results from Project Implicit (n.d.) were presented in categories, ranging from (*1 - strong preference for light skin tone* to *7 - strong preference for dark skin tone*, with labels *moderate, slightly, and no preference* between the two endpoints of the scale).

Results

Reliability of SST

Using reliability analysis, Cronbach's α for the bias score for SSTv1 was .45 and for SSTv2 was .73, with an average Cronbach's α of .63, indicating a moderate level of internal consistency (Hair et al., 2010). The corrected item-total correlation for SSTv1 and SSTv2 had a mean of .22 and .35, respectively. Pearson's Product Moment Correlation showed that the correlation between bias and non-bias scores was statistically significant for SSTv1, $r = -.76$, $p < .001$, and for SSTv2, $r = -.81$, $p < .001$. Next, to assess the split-half reliability of the SST, we correlated bias scores based on odd and even numbered items. Results revealed Spearman-Brown-corrected reliability between the two halves of .38 for SSTv1 and .72 for SSTv2.

Reliability of SRT

Using reliability analysis, Cronbach's α for the bias score for SRTv1 was .35 and for SRTv2 was .66, with an average Cronbach's α of .49, indicating a poor level of internal consistency (Hair et al., 2010). The corrected item-total correlation for SSTv1 and SSTv2 had a mean of .23 and .35, respectively. Pearson's Product Moment Correlation showed that the correlation between interpretation bias and non-bias scores was statistically non-significant for SRTv1, $r = .24$, $p = .25$, and for SRTv2, $r = .20$, $p = .27$, suggesting that SRT items likely had low sensitivity to assessing interpretation bias in the present study.

Next, to assess the split-half reliability of the SRT, we correlated bias scores based on odd and even numbered items. Results revealed Spearman-Brown-corrected reliability between the two halves of .03 for SRTv1 and .69 for SRTv2.

Correlation: Bias and Association Tests

Next, we examined whether there was an association between the SRT and SST, and the IAT. As point-biserial correlations determined the relation between the SST and SRT scores and skin-tone IAT. Results showed a statistically non-significant correlation between IAT scores and SST scores (SSTv1: $r = -.10$, $p = .64$; SSTv2: $r = -.21$, $p = .24$) and SRT scores (SRTv1: $r = .34$, $p = .09$; SRTv2: $r = -.24$, $p = .18$). Additionally, a Pearson's Product Moment Correlation showed that the correlation between the SRT and SST interpretation bias scores was statistically non-significant, $r = .02$, $p = .87$, suggesting that the two interpretation bias measures were likely non-convergent.

Usability Feedback

The results of students' ratings of CBM-S are shown in Table 3. As shown in Table 3, the mean ratings on all features revealed positive experiences with using CBM-S. That is, on average, students found CBM-S enjoyable, clear, interesting, and easy to use.

DISCUSSION

In the present paper, we reported the detailed development process of creating training and assessment materials for a novel digital self-administered ethnic bias training called Cognitive Bias Modification for Stereotype (CBM-S). CBM-S was developed in New Zealand and aims to address medical students' bias toward Māori

patients by presenting students with a series of ambiguous healthcare situations that invite multiple interpretations but leading students to respond with a non-stereotypical/neutral interpretation. An integral part of a high-quality and efficacious educational tool is the *relevance and specificity* of its materials, which can be achieved through adopting a user-centered approach to materials development. The present paper involved iterative inputs from medical students and Māori, and we aimed to achieve our two objectives: 1) to create and evaluate contents to be used in CBM-S that address common Māori stereotypes in healthcare settings and 2) to obtain user experience data of CBM-S and reliability data of interpretation bias measures.

Medical students created 100 common healthcare scenarios involving Māori, which researchers adapted into standard CBM format. CBM-S items adopted specific verbs to reflect a progression from higher-level thought processes (*what we think*) to more stable schemata/beliefs (*what we believe*). All items were rated by medical students and Māori participants using a set of pre-defined criteria (*stereotypicality, relevance to healthcare, readability, and ambiguity of the scenario*). Additionally, a progression from using specific Māori names to using more generic labels of ethnicity (i.e., Māori) was adopted to promote individual- and group-level attribution of traits/behaviours from the scenarios. The names that were rated by medical students and Māori to create a pool of names commonly given to Māori.

Out of the 100 scenarios created, 59 CBM-S training items and all control items reached a priori-defined acceptable value for each criterion. Further analyses of CBM-S content revealed that relative to control, training items were more ambiguous, more relevant to healthcare settings, and provided a more consistent depiction of common Māori patient stereotypes, with both training and control items rated as highly comprehensible. In the final set of CBM-S items, we included the 59 training items and selected 59 control items based on the average character count of scenarios so that the length of the scenarios was equivalent across training and control items. By matching the character count of the scenarios across experiment groups, we aimed to reduce the potential confounding effects of the duration spent in CBM-S training. Overall, medical students found CBM-S easy to use and to comprehend, interesting, and enjoyable.

The remaining 20 student-generated healthcare scenarios were adapted into interpretation bias assessment items. Fifteen items reached a priori-defined acceptable value for each criterion: *stereotypicality, relevance to healthcare, and readability.* These 15 items were used to create two versions of the Scrambled Sentence Task (SST) and the Similarity Rating Task (SRT)—two interpretation bias measures that are commonly used in clinical studies of the CBM method. Each version of the SST consisted of 15 scrambled sentences of six words each; each version of the SRT consisted of eight items. Reliability testing of interpretation bias assessments showed an overall moderate internal consistency for the SST but poor internal consistency for the SRT. There was a negative correlation between bias and non-bias scores on the SST. Our analyses did not reveal a correlation between bias and non-bias scores on the SRT or between the SRT and SST scores. These results suggest that despite adopting from the same pool of student-generated scenarios to create both the SRT and SST, the reliability of the SRT in measuring medical students' interpretation bias of scenarios involving Māori patients may be limited. One possible explanation for the difference in reliability between the two interpretation measures is that, unlike the SST, students' responses to the SRT may have been subjected to demand characteristics—a reported issue of the SRT assessment (Schoth & Liossi, 2018). The SST is a time-limited task under cognitive load. This feature of the SST was designed to limit conscious response bias (Bowler et al., 2012; Rude et al., 2003).

In addition to the overall reliability data, our analyses revealed that relative to version one of both interpretation bias measures, version two showed better internal consistency and split-half reliability. There may be several explanations for this difference. First, 15 user-generated exemplars that reached a priori-defined acceptable values were adopted to create two versions of 15 SST items and eight SRT items (46 items in total). This is to accommodate for the conventional number of items used for each measure. As a result, some assessment items were similar but not the same as student-generated scenarios. Furthermore, the finalized assessment items were not re-rated by service users, meaning that these items did not go through the same iterative process of review and refinement as during the first phase of item creation. Items that were included in version two of the bias measures may have been more similar to the student-generated scenarios than did version one, which may have accounted for the differences in reliability across versions. Through retrospection of our data, we observed that the mean stereotypicality rating of the items for version one was $M = 5.40$, and the rating for version two was $M = 5.81$, although not statistically significant.

Another possible explanation is that despite adopting student-generated exemplars to create assessment items, the personal relevance to any one individual is limited and may not necessarily reflect the interpretation of the scenarios of all our participants. A re-examination of our data revealed a dispersion of ratings of assessment items. More specifically, stereotypicality ratings of the items ranged from 3.33 – 7, with a variation of 1.26. These data suggest individual variations in the interpretation of assessment scenarios, which may have impacted the reliability across versions.

Consistent with findings from previous studies, we did not find a correlation between interpretation bias measures and the Implicit Association Test (IAT). IAT measures the strength of associations between categories and automatic evaluations of those categories—as opposed to biased interpretations of healthcare situations. In other words, at the basic level of social information processing, IAT may be assessing a different cognitive component (i.e., pattern matching) to that of students' interpretation of healthcare scenarios (Sekaquaptew et al., 2003). Specifically, the skin-tone IAT assesses whether participants recognize and associate 'good' (e.g., attractive, joyful) or 'bad' (e.g., horrific, despise) words with light- or dark-skinned faces. Interpretation bias measures, such as SRT and SST, assess participants' interpretation of scenarios involving the target group,

which may have manifest in different ways depending on the given situation.

The content development process discussed in this paper should consider the following limitations. First, as previously discussed, five medical students from the same university created 100 healthcare exemplars involving Māori, which may limit the generalization of these scenarios. Future work should include a larger sample within a broader context to create content that directly reflects personal experiences of ethnic stereotyping in different healthcare contexts. By the same token, interpretation bias assessment items should adopt one-to-one mapping of student-generated scenarios to individual test items to improve the reliability and relevance of the measures.

A second limitation is that further development of assessments for measuring interpretation bias toward marginalized ethnic groups, particular for the SRT, should address issues of demand characteristics. In some studies, researchers have included indirect ratings of interpretation bias by measuring participants' ratings of pleasantness (Berna et al., 2011), or their level of concern (Davey et al., 1992), pertaining to the SRT scenarios. These ratings invite participants to report their feelings rather than their explicit interpretations, which may mask the intent of the assessment.

Finally, although Māori are often (inaccurately) associated with dark skin, using skin tone to measure associations of concepts of Māori is unidimensional and may have limited the validity of our findings. Māori is rich in cultural values and practices derived from *mātauranga Māori* (Māori knowledge) that encompasses multidimensional concepts and traditions (Mead, 2016). As such, using skin tone is a limited and superficial representation of Māori that may not have captured students' perceptions of Māori when engaged in assessment and training.

We anticipate that CBM training will complement existing training methods, such as metacognitive strategies, fact provision, and open discussions, as a part of a weekly independent learning approach in medical education and professional development. There is empirical evidence in the clinical literature suggesting a reduction in interpretation bias following six-session weekly trainings (40-items per session), with evidence of effects after the third session that remained at a 1-month and 3-month follow-up (Yiend et al., 2022). We envisage a similar weekly training session approach for CBM-S in medical education, with future studies needed on the 'dose-response' of the training in modifying ethnic bias.

Conclusion

In conclusion, CBM is a class of training methods that has now been extended into a digital training tool to address implicit ethnic bias, which we called CBM-S. In the NZ context, CBM-S aims to address medical students' bias toward Māori in healthcare settings. When developing any educational tool, it is important to follow a user-centered development approach to maximize *content specificity and relevance* of training content. This approach will most likely also optimize user acceptability and engagement, and the effectiveness and reliability of the training and assessment, to create an ethnic bias training tool to achieve true health equity.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: a meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659-701. https://doi.org/10.3102/0034654316689306

Atkinson, R.C., & Shiffrin, R.M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence, & J. T. Spence (eds). *The psychology of learning and motivation: II*. New York: Academic Press.https://doi.org/10.1016/S0079-7421(08)60422-3

Baah, F.O., Teitelman, A.M., & Riegel, B. (2019). Marginalization: conceptualizing patient vulnerabilities in the framework of social determinants of health—an integrative review. *Nursing Inquiry, 26*(1), e12268. https://doi.org/10.1111/nin.12268

Berna, C., Lang, T. J., Goodwin, G. M., Holmes, E. A. (2011). Developing a measure of interpretation bias for depressed mood: an ambiguous scenarios test. *Personality and Individual Differences, 51*(3), 349–354. https://doi.org/10.1016/j.paid.2011.04.005

Blair, I.V., Steiner, J.F., & Havranek, E.P. (2001). Unconscious (implicit) bias and health disparities: Where do we go from here? *The Permanente Journal, 15*(2), 71–78. https://doi.org/10.7812/tpp/11.979

Bobak, C. A., Barr, P. J., & O'Malley, A. J. (2018). Estimation of an inter-rater intra-class Correlation coefficient that overcomes common assumption violations in the assessment of Health measurement scales. *BMC Medical Research Methodology, 18*(1), 93. https://doi.org/10.1186/s12874-0180550-6

Bowler, J. O., MacKintosh, B., Dunn, B. D., Mathews, A., Dalgleish, T., & Hoppitt, L. (2012). A comparison of cognitive bias modification for interpretation and computerized cognitive behavior therapy: effects on anxiety, depression, attentional control, and interpretative bias. *Journal of Consulting and Clinical Psychology, 80*(6), 1021–1033. https://doi.org/10.1037/a0029932

Brunyé, T.T., Ditman, T., Mahoney, C.R., & Taylor, H. (2011). Better you than I: perspectives and emotion simulation during narrative comprehension. *Journal of Cognitive Psychology, 23*(5), 659–666. https://doi.org/10.1080/20445911.2011.559160

Cooper, L.A., Roter, D.L., Carson, K.A., Beach, M.C., Sabin, J.A., Greenwald, A.G., et al. (2012). The associations of clinicians' implicit attitudes about race with medical visit communication and patient ratings of interpersonal care. *American Journal of Public Health, 102*(5), 979–87. ttps://doi.org/10.2105/AJPH.2011.300558

Davey, G. C., Hampton, J., Farrell, J., & Davidson, S. (1992). Some characteristics of worrying: evidence for worrying and anxiety as separate constructs. *Personality and Individual Differences, 13*(2), 133–147. https://doi.org/10.1016/0191-88692(92)90036-OEllision-Loschmann, L., & Pearce, N. (2006). Improving access to health care among New Zealand's Māori population. *American Journal of Public Health, 96*(4), 612–617. https://doi.org/10.2015/ajph.2005.070680

Eysenck, M.W., Mogg, K., May, J., Richards, A., & Mathews, A. (1991). Bias in interpretation of ambiguous

sentences related to threat in anxiety. *Journal of Abnormal Psychology, 100*(2):144–150. https://doi.org/10.1037//0021-843x.100.2.144

Forscher, P., Lai, C.K., Axt, J.R., Ebersole, C.R., Herman, M., Devine, P.G., & Nosek, B. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology, 117*(3), 522–559. https://doi.org/10.1037/pspa0000160

Gawronski, B., Rydell, R. J., De Houwer, J., Brannon, S. M., Ye, Y., Vervliet, B., & Hu, X. (2018). Contextualized attitude change. In J. M. Olson (Eds). *Advances in experimental Social psychology*. Elsevier Academic Press. https://doi.org/10.1016/bs.aesp.2017.06.001

Gonzalez, C.M., Kim, M.Y., & Marantz, P.R. (2014). Implicit bias and its relation to health disparities: A teaching program and survey of medical students. *Teaching and Learning in Medicine, 26*(1), 64–71. https://doi.org/10.1080/10401334.2013.857341

Gould, D., Kelly, D., Goldstone, L., & Gammon, J. (2001). Examining the validity of pressure ulcer risk assessment scales: developing and using illustrated patient simulations to collect the data. *Journal of Clinical Nursing, 10*(5), 697–706. https://doi.org/10.1046/j.1365-2702.2001.00525.x

Greenwald, A. G., McGhee, D. E., Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology, 74*(6):1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Hair, Jr. J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis: a global perspective (7th edition)*. Upper Saddle River, New Jersey: Prentice Hall; 2010.

Harris, R., Cormack, D., Curtis, E., Jones, R., Stanley, J., & Lacey, C. (2016). Development and testing of study tools and methods to examine ethnic bias and clinical decision-making among medical students in New Zealand: the bias and decision-making in medicine (BDMM) study. *BMC Medical Education, 16*, 173. https://doi/org/10.1186/s12909-016-0701-6

Harris, R., Cormack, D., Tobias, M., Yeh, L.C., Talamaivao, N., & Timutimu, R. (2012). Self-report experience of racial discrimination and health care use in New Zealand: Results from 2007/07 New Zealand health survey. *American Journal of Public Health, 102*(5),1012–1019. https://doi.org/10.2105/ajph.2011.300626

Hastie, R. (1984). Causes and effects of causal attribution. *Journal of Personality and Social Psychology, 46*(1):44–56. https://doi.org/10.1037/0088-3514.46.1.44

Hirsch, C.R., Krahé, C., Whyte, J., Loizou, S., Bridge, L., Norton, S., et al. (2018). Interpretation training to target repetitive negative thinking in generalized anxiety disorder and depression. *Journal of Consulting and Clinical Psychology, 86*(12), 1017–1030. https://doi.org/10.1037/ccp0000310

Holmes, E.A., & Mathews, A. (2005). Mental imagery and emotion: a special relationship? *Emotion, 5*(4), 489–497. https://doi.org/10.1037/1528-3542.5.4.489

Holmes, E.A., Mathews, A., Dalgleish, T., & Mackintosh, B. (2006). Positive interpretation training: effects of mental imagery versus verbal training on positive mood. *Behavior Therapy, 37*(3), 237–247. https://doi.org/10.1016/j.beth.2006.02.002

Hsu, C. W. (2023). Mind over prejudice: An implicit bias training in medical education using cognitive bias modification. *Journal of Graduate Medical Education.*

Hsu, C. W. & Akuhata-Huntington, Z. (2024). I have a dream: Altering medical students' ethnic bias towards Indigenous population (NZ Māori) using a digital training called cognitive bias modification. *Stigma and Health.* Advance online publication. https://doi.org/10.1037/sah0000505

Hsu, C. W., Stahl, D., Mouchlianitis, E., Peters, E., Vamvakas, G., Keppens, J., Watson, M., Schmidt, N., Jacobsen, P., McGuire, P., et al. (2023). User-centered development of STOP (Successful Treatment for Paranoia: Material development and usability testing for a digital therapeutic for paranoia. *JMIR Human Factors, 8*(10): e45453. https:/doi.org/10.2196/45453

Joy-Gaba, J. A., & Nosek, B. A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology, 41*(3), 137-146. https://doi.org/10.1027/1864-9335/a000020

Koster, E. H. W., & Horrelbeke, K. (2015). Cognitive bias modification for depression. *Current Opinion in Psychology, 4*, 119–123. https://doi.org/10.1016/j.copsyc.2014.11.012

Lachter, J., Forster, K.I., & Ruthruff, E. (2004). Forty-five years after Broadbent (1958): still no identification without attention. *Psychological Review, 111*(4), 880–913. https://doi.org/10.1037/0033-295X.111.4.880

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: a tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lamberton, A., & Oei, T. P. S. (2008). A test of the cognitive content specificity hypothesis in depression and anxiety. *Journal of Behavior Therapy and Experimental Psychiatry, 39*(1), 23–31. https://doi.org/10.1016/j.jbtep.2006.11.001.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a User Experience Questionnaire. In: A. Holzinger (Eds). *HCL and Usability for Education and Work*, 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB. 2008 Nov 20-21; Graz, Austria. Lecture Notes in Computer Science, *5298*. https://doi.org/10.1007/978-3-540-89350-9_6

Lebrecht, S., Pierce, L.J., Tarr, M.J., & Tanaka, J.W. (2009). Perceptual other-race training reduces implicit racial bias. *PLoS One, 4*(1), e4215. https://doi.org/10.1371/journal.pone.0004215

Lee, K., Quinn, P. C., & Heyman, G.D. (2017). Rethinking the emergence and development of implicit racial bias: A perceptual-social linkage hypothesis. In N. Budwig, E. Turiel, & P. D. Zelazo (Eds). *New perspectives on human development*. Cambridge University Press. https://doi/org/10.1017/CBO9781316282755.004

Maass, A., Salvi, D., Arcuri, L., & Semin, G. (1989). Language use in intergroup contexts: the linguistic intergroup bias. *Journal of Personality and Social Psychology, 57*(6), 981–993. https://doi.org/10.1037//002-3514.57.6.981

Mathews, A., & MacLeod, C. (1994). Cognitive approaches to emotion and emotional disorders. *Annual Review of Psychology, 45*(1), 25-50. https://doi.org/10.1146/annurev.ps.45.020194.000325

McCartney, G., Popham, F., McMaster, R., & Cumbers A. (2019). Defining health and health inequalities. *Public Health. 172*, 22–30. https://doi.org/10.1016/j.puhe.2019.03.023
Mead, H. (2016). *Tikanga Māori living by Māori values. Revised edition*. Wellington, New Zealand: Huia Publishers.
Nisbett, R.E., & Ross, L. (1980). *Human inference: strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice-Hall
Paradies, Y. (2006). A systematic review of empirical research on self-reported racism and health. *International Journal of Epidemiology, 35*(4), 888–901. https://doi.org/10.1093/ije/dy1056
Paradies, Y., Truong, M., & Priest, N. (2014). A systematic review of the extent and measurement of healthcare provider racism. *Journal of General Internal Medicine, 29*(2), 364–387. https://doi.org/10.1007/s11606-013-2583-1
Project Implicit (n.d.). *Skintone-Science IAT*. https://implicit.harvard.edu
Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychology Bulletin, 140*(6), 1432-1463. https://doi.org/10.1037/a0037559
Rude, S.S., Valdez, C.R., Odom, S., & Ebrahimi, A. (2003). Negative cognitive biases predict subsequent depression. *Cognitive Therapy and Research, 27*(4), 415–429. https://doi.org/10.1023/A:1025472413805
Rude, S.S., Wenzlaff, R.M., Gibbs, B., Vane, J., & Whitney, T. (2022). Negative processing biases predict subsequent depressive symptoms. *Cognition & Emotion, 16*(3), 423–440. https://doi.org/10.1080/02699930143000554
Rumelhart, D.E. (1984). Schemata and the cognitive system. In R.S. Wyer, Jr. & T. L. Srull (Eds). *Handbook of social cognition, 1*. Eribaum, Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.
Sabin, J., Guenther, G., Ornelas, I. J., et al. (2021). Brief online implicit bias education increases bias awareness among clinical teaching faculty. *Medical Education Online, 27*(1), 2025307. https://doi.org/10.1080/10872981.2021.2025307
Satherley, N., & Sibley, C.G. (2018). The modern racism toward Māori scale. *The New Zealand Journal of Psychology, 47*(2), 4–13
Savulich, G., Freeman, D., Shergill, S.S., & Yiend, J. (2015). Interpretation biases in paranoia. *Behavior Therapy, 46*(1), 110-124. https://doi.org/10.1016/j.beth.2014.08.002
Savulich, G., Shergill, S.S., & Yiend, J. (2017). Interpretation biases in clinical paranoia. *Clinical Psychological Science, 5*(6). 985-1000. https://doi.org/10.1177/2167702617718180
Schoth, D. E., & Liossi, C. (20180). A systematic review of experimental paradigms for exploring biased interpretation of ambiguous information with emotional and neutral associations. *Frontier in Psychology*, 8, 171. https://doi.org/10.3389/fpsyg.2017.00171
Sekaquaptewa, D., Espinoza, P., Thompson, M., Vargas, P., & von Hippel, W. (2003). Stereotypic explanatory bias: Implicit stereotyping as predictor of discrimination. *Journal of Experimental Social Psychology, 39*(1), 75–82. https://doi.org/10.1016/S0022-1031(02)00512-7
Smedley, B.D., Stith, A.Y., & Nelson, A.R. (2003). *Unequal Treatment: Confronting Racial and Ethnic Disparities in Healthcare* (Eds). Washington, DC: National Academy Press. https://www.ncbi.nlm.nih.gov/books/NBK220358/
Standage, H., Ashwin, C., & Fox, E. (2009). Comparing visual and auditory presentation for the modification of interpretation bias. *Journal of Behavior Therapy and Experimental Psychiatry*, 40(4):558–570. https://doi.org/10.1016/j.jbtep.2009.07.006
van Ryn, M., & Fu, S.S. (2003). Paved with good intentions: do public health and human service providers contribute to racial/ethnic disparities in health? *American Journal of Public Health, 93*(s), 248–255. ttps://doi.org/10.2105/ajph.93.2.248
von Hippel, W., Sekaquaptewa, D., & Vargas, P. (1997). The Linguistic Intergroup Bias as an implicit indicator of prejudice. *Journal of Experimental Social Psychology*, 33(5):490–509. https://doi.org/10.1006/jesp.1997.1332
White III AA. (2011). *Seeing patients: Unconscious bias in health care*. Cambridge, MA: Harvard University Press.
Woud, M.L., Verwoerd, J., & Krans, J. (2017). Modification of cognitive biases related to posttraumatic stress: A systematic review and research agenda. *Clinical Psychology Review, 54*, 81–95. https://doi.org/10.1016/j.cpr.2017.04.003
Würtz, F., Zahler, L., Blackwell, S.E., Margraf, J., Bagheri, M., Woud, M.L. (2022). Scrambled but valid? The scrambled sentences task as a measure of interpretation biases in psychopathology: A systematic review and meta-analysis. *Clinical Psychology Review, 93*, 102133. https://doi.org/10.1016/j.cpr.2022.102133
Yiend, J., & Mackintosh, B. (2004). The experimental modification of processing biases. In J. Yiend (ed). *Cognition, emotion and psychopathology*. Cambridge, UK: Cambridge University Press.
Yiend, J., Lam, C., Schmidt, N., Crane, B., Heslin, M., Kabir, T., McGuire, P., Meek, C., Mouchlianitis, E., Peters, E., Stahl, D., Trotta, A., & Shergill, S. (2022). Cognitive bias modification for paranoia (CBM-pa): a randomised controlled feasibility study in patients with distressing paranoid beliefs. *Psychological Medicine, 14*, 1–13. https://doi.org/10.1017/S0033291722001520

Corresponding author:
Dr Che-Wei Hsu, Department of Psychological Medicine, University of Otago, PO Box 54, Dunedin, New Zealand 9054.
Tel +64 3 470 0999 ext 57362
E-mail: jerry.hsu@otago.ac.nz
ORCID: 0000-0002-3297-3961

Statements and Declaration
This is to acknowledge that there is no conflict of interest nor is there any financial interest or benefit that has arisen from the direct applications of this research. Survey elements, Copyright Qualtrics, LLC. Used With Permission

Acknowledgments
This research was funded by the Accelerator Grant 2022, Division of Health Sciences, University of Otago.