Manuscript Details

Manuscript number	JAD_2017_1504_R1
Title	Distinguishing transient versus stable aspects of depression in New Zealand Pacific Island children using Generalizability Theory
Article type	Research Paper

Abstract

Abstract Background: The distinction between temporary versus enduring or state/trait aspects of depression is important. More precise distinction would improve understanding of the aetiology of depression and those aspects most amenable to intervention thus identifying more homogeneous, dynamic targets for clinical trials. Generalizability Theory has been proposed as useful for disentangling state and trait components of psychopathology. Methods: We applied Generalizability Theory to determine the relative contributions of temporary and enduring aspects of depression in a widely used screening measure of depression the - 10-item Children's Depression Inventory (CDI-10; Kovacs, 1985). Participants were children of Pacific Island descent living in New Zealand (n= 668). Data were collected at ages - 9, 11, and 14 years. Results: The CDI-10 demonstrated acceptable generalizability across occasions (G=.79) with about one third of variance in total scores attributed to temporary and two thirds to more enduring aspects of depression. There were no other significant sources of error variance. Two items were identified as more sensitive than the remaining eight to more dynamic symptoms. Limitations: Studies with briefer test-retest intervals are warranted. Use of this Pacific Island cohort limits generalisability of findings to other cultures and ethnicities. No data were collected on whether participants had received intervention for depression. Conclusions: While the CDI-10 reliably measures both stable and transient aspects of depression in children, the scale does not permit clear distinction between them. We advocate application of Generalizability Theory for developing state/trait depression measures and determining which existing measures are most suitable for capturing modifiable features of depression.

Keywords	Depression; Generalizability Theory; Children's Depression Inventory; State; Trait

Corresponding Author	Richard Siegert
Order of Authors	Janis Paterson, Oleg Medvedev, Alexander Sumich, El-Shadan Tautolo, Christian Krägeloh, Rose Sisk, Robert McNamara, Michael Berk, Ajit Narayanan, Richard Siegert

Suggested reviewers	Bill Hoyt, Geoffrey Norman, Alan Tennant, Caroline Terwee

Submission Files Included in this PDF

File Name [File Type]

G_CDI_JAD_CoverLetter_Siegert_10.11.17.docx [Cover Letter]

G_CDI_JAD_Response to Reviewers_Siegert_10.11.17.docx [Response to Reviewers]

State_Trait_CDI_G_Siegert_Highlights.docx [Highlights]

State_Trait_CDI_G_Siegert_Abstract_10.11.17.docx [Abstract]

State_Trait_CDI_G_Siegert_10.11.17.docx [Manuscript File]

State_Trait_CDI_G_Siegert_Table1.docx [Table]

State_Trait_CDI_G_Siegert__Table2.docx [Table]

State_Trait_CDI_G_Siegert_Table3.docx [Table]

State_Trait_CDI_G_Siegert_Table4.docx [Table]

State_Trait_CDI_G_Siegert_Table5.docx [Table]

State_Trait_CDI_G_Siegert__Conflict of Interest Statement.docx [Conflict of Interest]

State_Trait_CDI_G_Siegert__Authors_Statement_10.11.17.docx [Author Statement]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

Research Data Related to this Submission

There are no linked research data sets for this submission. The following reason is given:
The authors do not have permission to share data

Professor Paolo Brambilla 10.11.17
Editor-in-Chief
Journal of Affective Disorders

Ref: JAD_2017_1504
Title: Distinguishing transient versus stable aspects of depression in New Zealand Pacific Island children using Generalizability Theory
Journal: Journal of Affective Disorders

Dear Professor Brambilla,

Thank-you for the opportunity to revise and resubmit our manuscript to JAD. Thank-you also for the extension to accommodate the fact that I was travelling in September. We found the reviewers' comments to be both insightful and constructive and we have attempted to address all of the points that they raised. We believe that the manuscript is considerably improved as a consequence. I shall detail the reviewers' comments and our responses to each point in the attached file Response to Reviewers.

Please note that in our revised manuscript all new or modified text is highlighted in red ink

In summary we have addressed each point that was raised by both reviewers and the manuscript is substantially improved. To this end I have added a sentence thanking and acknowledging the input of two anonymous reviewers. I trust that this is appropriate and satisfactory to you. We look forward to hearing from you in due course concerning the final decision on our manuscript.

Kind regards

Richard J. Siegert

10.11.17
Journal of Affective Disorders

Ref: JAD_2017_1504
Title: Distinguishing transient versus stable aspects of depression in New Zealand Pacific Island children using Generalizability Theory
Journal: Journal of Affective Disorders

Please note that in our manuscript all new or modified text is highlighted in red ink:

Comments from the editors and reviewers:

Reviewer 1 This longitudinal study, conducted among a large sample of children in New Zealand, examined the relative contributions of temporary and stable aspects of depression. Generalizability theory was used; the examined measure was the brief 10-item Children's Depression Inventory (CDI). Results show that generalizability across occasions is acceptable (G= .79). Moreover, about one third of the variance in total scores can be attributed to dynamic/changing aspects of depression; while about two thirds can be attributed to more stable/enduring aspects of depression. Items differ in their sensitivity to changing depression symptoms.

The paper addresses an important issue, and is overall quite well written. However, several issues need to be addressed before the paper can be considered for publication. My detailed comments are provided below.

General

COMMENT: The paper is difficult to follow for those not familiar with Generalizability Theory.

RESPONSE: We have added several new sentences/paragraphs in response to a number of the comments from both reviewers and the explanation of Generalizability Theory is considerable expanded. Please see several examples in our responses below.

COMMENT: Throughout, several typos have to be corrected (for instance, insert a comma after e.g.)

RESPONSE: We have added a comma after 'e.g.' throughout the paper and also corrected some other typographical errors.

Introduction

COMMENT: The authors should provide a definition/description of the "generalizability coefficient" (page 5)

RESPONSE: We have added the following to the Introduction in the first paragraph on p.4:

This includes calculation of a *G coefficient* which represents how generalizable the test scores are across different situations. Bloch and Norman (2012, p. 968) describe the G coefficient as

the ratio of signal to noise or 'true' variance to 'true + error variance' and provide further details on its derivation and calculation.

COMMENT: An explicit rationale for using the short form of the CDI as opposed to the 27 item full scale is needed.

RESPONSE: We have added this new text to the final paragraph of the Introduction on p. 7:

The brief CDI was originally chosen for this longitudinal cohort study because its brevity made it practical to include in a much longer battery of assessment tools. The availability of data from three points in time made it particularly suitable for the present Generalizability analysis.

Method

COMMENT: Page 6: how does the final sample (N= 668) relate to the original sample (N= 1398) exactly? Were there any differences between those included and those not included at baseline or at age 9?

RESPONSE: We have added the following text to the Methods section under Participants on p.8.

The present sample included 668 participants who have completed the CDI three times with a two/three year interval between assessments at ages 9, 11, and 14 years and who had no more than 5 missing items at any one assessment. This sample of 668 had the same distribution of Pacific Island ethnic groups as the remainder of the original birth cohort (N=730) but a higher female:male ratio (53:47% versus 45:55%, p=0.007).

COMMENT: Page 6: how were the 10 items selected exactly?

RESPONSE: We have added new text to the Methods section under Measure on p.8.

These items were selected from the full 27 by removing the one item that would most attenuate Cronbach's alpha in a stepwise procedure until a reliable set of 10 items remained (Kovacs, 1985, 2003).

COMMENT: Page 7: mean imputation has drawbacks (e.g., it attenuates any correlations involving the variables that are imputed). Were more sophisticated imputation methods considered?

RESPONSE: Please see our response to the detailed comments regarding imputation made by Rev. 2, # 10 below.

COMMENT: Pages 7-8: the discussion of the theory is very hard to follow for one who has no or little knowledge of Generalizability theory (e.g., what is Whimbey's correction coefficient?;

RESPONSE: We have added the following text explaining more about Whimbey's correction in the first paragraph on p.4 in the Introduction.

An important step in a Generalizability analysis is the calculation of individual variance components for each facet (e.g., person, item, occasion) in the measurement situation. These variance components are first calculated by a conventional ANOVA and corrected by a method known as 'Whimbey's correction which accounts for the type of sampling involved (i.e. random, fixed or random finite). This also includes calculation of a *G coefficient* which represents how generalizable the test scores are across different situations. Bloch and Norman (2012, p. 968) describe the G coefficient as the ratio of signal to noise or 'true' variance to 'true + error variance' and provide details on its derivation and calculation.

COMMENT: what is the difference between a differentiation facet and a source of error?;

RESPONSE: We have added this sentence to the section headed Generalizability Theory at the bottom of p.3:

In G terms the variance associated with participants or persons is considered the central concern and is known as the *differentiation facet* with other facets (e.g., items, occasion) viewed as sources of measurement error.

COMMENT: how is the ANOVA analysis performed exactly?)

RESPONSE: We have added these three sentences to answer this query to the first paragraph on p.4 in the Introduction under Generalizability Theory header.

In most Generalizability studies there are two stages known as the G-study and the D-study. The G-study involves a standard factorial ANOVA with the calculation of variance components for each facet and for their interactions (rather than the significance tests typically associated with an ANOVA). The D-study allows the researcher to then estimate the impact on reliability of variations in different facets such as increasing the number of participants or the number of items in a scale.

Results

COMMENT: Page 9: was a 2 factor solution examined (e.g., using CFA in MPlus)?

RESPONSE: No, we did not examine a 2 factor solution. As explained in the manuscript the Scree plot indicated a single factor solution and this was supported by consistent high loadings on the first principal component. While this is an interesting suggestion from the reviewer the focus of this paper is on a Generalizability theory approach to psychometrics. With all due respect we fear that introducing a 2-factor confirmatory factor analysis would complicate the paper to an extent that could deter most readers not expert in psychometrics.

COMMENT: Page 11: please explain the cutoff score of 4 for the subgroup formation (e.g., why not 3 or 5)

RESPONSE: We have added the following sentence to the Results section on pp.11.

We chose a score of 4 since the full CDI has 27 items with 12 as the cut-off point for depression and 4 on a 10 item version represents a roughly equivalent cut-off point.

Discussion

COMMENT: Page 13: how can individual items be used as trait vulnerability markers exactly?

RESPONSE: We have added a sentence to explain this more fully on p. 16.

……..could be used to index risk. So when individuals score high on several of the most stable or trait-like items it could be used as an indicator that these children are most prone to depression and might benefit from early intervention.

Reviewer 2

The manuscript utilized G Theory analyses to assess for state versus trait aspects of depression using the CDI-10 with a group of children at ages 9, 11 and 14. The study used a novel method with the G Theory that provided an innovative and interesting approach to better understanding the depression symptoms assessed by the CDI-10 and how they may be more state-like versus trait-like over a 5-year span. Below I have outlined some concerns about the manuscript.

Major Concerns:

COMMENT: 1. The introduction is a bit disjointed. Perhaps adding in some transitions would help with this. This was particularly noticeable when jumping from the G-theory section to the Stable versus Transient section.

RESPONSE: We have added the following sentence to smooth the transition from the G-theory section to the Stable Versus Transient section (p.5):

In the present article we argue that G Theory may have particular value for distinguishing between state and trait or stable versus enduring aspects of behaviour and depression in particular.

COMMENT: 2. In line with the previous comment, perhaps consider adding in a mental health example into the G-theory section and shortening the current example of the nursing program.

RESPONSE: (a) We have deleted the following text to shorten the current nursing example:

If this hypothetical examination involved a very large cohort of students and ran over several days we could also include the day of the week or *occasion* as another facet – thus permitting us to determine if examiners were consistent from Monday to Friday. Importantly, G Theory also allows us to calculate the variance component and error variance due to the interactions between student and station or examiner and occasion. Similarly, if examiners rotated around the ten stations, we could then determine if the same examiners rated different students more or less stringently on different clinical problems (i.e. rater x student x station) and the amount of error variance this interaction accounts for. In a similar way, such methods may be used to improve reliability in multisite research studies.

(b) We have added the following new text to provide an example from a mental health context to the bottom of p.4:

Consider another example from a mental health setting. In examining the inter-rater reliability of a new rating scale for assessing the severity of depression in outpatients we might have the following facets: Person (the differentiation facet), rater, and occasion. The calculation of variance components for each facet and their interactions allows us to determine precisely how much variance is due to the patient (Person facet) and changes in their condition (Person x Occasion) as well as the raters. High inter-rater reliability will be reflected in a very small variance component for the facet Rater.

COMMENT: 3. Please make it clearer to the reader that the first set of analyses discussed are the results from the full sample and not the sub-group sample.

RESPONSE: We now start the Results section on p.11 with the following sentence:

We report the results of the analyses for the full sample of 668 children first and then report the sub-group analysis for the 244 children with a score ≥ 4 on the brief CDI.

COMMENT: 4. The manuscript states that the "major contribution of the present study is in highlighting the potential value of G Theory for research on state-trait distinction." Yet the intro stated that the aims of the study were "to examine the psychometric properties of the 10-item version of the CDI using G theory" and "to explore the degree to which state and trait or stable versus transient aspects of depression can be quantified for this scale." Given the aims of the study, I expected there to be more of a discussion about the CDI-10 in the discussion section of the manuscript. For instance, given that the results indicated that it is mostly assessing trait-like symptoms, what does that say about the CDI-10 and its clinical utility? Why was the CDI-10 initial developed? Why use a measure of mostly only stable traits? These were all questions that I expected to be answered as I read through the results and into the discussion.

RESPONSE: We have moved the following sentence up on p.14:
The CDI-10 was initially developed as a rapid screening tool to quantify the severity of depressive symptoms in children for clinical and research purposes (Kovacs, 1985).

And we have added the following text to the second half of the first para of Discussion section on p.14):

This raises the question of whether a measure of predominantly stable traits is the best measure for those situations where change is a primary interest such as treatment studies. Interestingly Wu's (2016) structural equation modelling study of the Beck Depression Inventory – II using a 'state-trait-occasion model' reported that it measured both 'trait-like and occasion-specific variance' (p.48) and the trait variance was greater than 50%. La Grange et al. who examined the full 27-item CDI using a structural equation model concluded that the CDI was mostly measuring state variance. In contrast, the present results suggest that the brief CDI is a good measure of relatively stable individual differences in depressive symptoms while also being reasonably sensitive to more dynamic or occasion specific depressive symptoms.

COMMENT: 5. The test-retest reliability of the CDI-1- was low. Why would this be low when the G Theory analyses were indicating that the measure is mostly assessing traits of depression? Wouldn't a trait measure be expected to have high test-retest reliability?

RESPONSE: We have added new text to the opening paragraph of the Discussion on p.15.

Interestingly, the G coefficient for this sample seemed somewhat in contrast to the relatively low test-retest correlations of 0.41 at two years and 0.22 at four years test-retest interval. However, these results are comparable with previous findings. For example, Saylor, Finch, Spirito and Bennett (1984) reported a one week test-retest correlation for the full CDI of .38 for 69 'normal' 5th and 6th graders and .59 and .87 for two samples of 'emotionally disturbed' children. Similarly, Nelson and Politano (1990) reported a test-retest coefficient of .47 with a 30 day test-retest interval for 96 psychiatrically hospitalized boys and girls. Certainly, further work examining the relationship between generalizability, as reflected in the G coefficient, and temporal stability, as measured by test-retest correlations, seems indicated.

COMMENT: 6. The discussion mentions the potential limitation that the current sample was only Pacific Island children. Are there previous studies illustrating potential differences between this population and populations with other demographics? Is there any research to indicate depression may present differently? Any previous research could be cited here to illustrate potential population differences and why it is important to investigate constructs across different populations.

RESPONSE: We have added the following new text to the Discussion on p.18.

…… use of the PIFS cohort limits generalizability of findings to other cultures and ethnicities, which should be investigated in future research. New Zealand has a high lifetime prevalence of depression by international standards (Kessler and Bromet, 2013), and Pacific Island youth within New Zealand have about double the rate of depression and suicide attempts of the general population (Statistics New Zealand and Ministry of Pacific Island Affairs, 2011).

COMMENT: 7. In the introduction, the authors discuss a study on G Theory and the QoL measure. Explain further what is meant or concluded from the remark that the "reliability of change for three QoL subscales was consistently less than Cronbach's alpha for each time point."

RESPONSE: We have added the following text to the Introduction to explain this study and its findings more clearly and more fully at bottom of p.15:

A recent paper by Chavez et al., (2016) demonstrated this approach with the Adolescent Quality of Life-Mental Health Scale (AQOL-MHS) in adolescents attending mental health clinics assessed on three occasions over eight months. The focus of that study was on the reliability of change scores, and the authors used a method developed by Cranford et al. (2006) based on G Theory to compare conventional estimates of reliability (Cronbach's alpha, test-retest) with G Theory estimates. The authors observed that the reliability of change assessed by G Theory for the three subscales of the AQOL-MHS was adequate but consistently lower than Cronbach's alpha for each time point. They concluded Cronbach's alpha is sufficient when comparing scores across people at the same point in time but the

reliability of change scores must be considered when comparing the change in persons across multiple time points.

COMMENT: 8. Provide more detail on the G theory study with the BDI. Suggest taking out or shortening the other examples of G theory including the above section about nursing school (see comments 1 and 2), and providing more details on the BDI. Are there any other studies in the literature assessing G Theory in relation to any other measures of Depression?

RESPONSE: We have shortened the example given of G Theory applied to assessment of nursing skills and added a mental health example (see response to Reviewer 2, Major concern #2 above).

We have also extended the detail given about the BDI study using G Theory (Wu, 2016) and added new material on structural equation methods applied to measures of depression in the Introduction p.2., para.2, See below:

An alternative approach to assessing state versus trait has been to use structural equation modeling to compare the covariance structures of a state/trait measure on at least two separate occasions (Cole, Martin & Steiger, 2005). Using this "latent trait-state-occasion model" with the Beck Depression Inventory II (BDI-II) in two large samples of young people, Wu reported a trait component accounting for more than 50% of variance and an 'occasion-specific factor' that explained between 7 – 12% of variance (Wu, 2016, p.39). LaGrange et al. (2011) applied the latent trait-state-occasion model to four measures of depressive cognitions and symptoms in a longitudinal study of 515 children and adolescents and found three of these measures characterised by "two types of longitudinal factors: a time invariant (or trait-like) factor and a series of time-varying (more state-like) factors" (p.13). Of particular relevance to the present study was their fourth measure - the Children's Depression Inventory - was almost entirely accounted for by time varying or state like variance.

COMMENT: 9. How did the current results align with the previous study by Crowley et al. (1994)? It is shown in the table and mentioned vaguely in the discussion but it is unclear to the reader how they are similar/different. Please expand and explain fully.

RESPONSE: We have added new text to the Discussion on p.17 (bottom of page) as follows:

The most notable differences were that Items accounted for a substantial amount (5.8%) of error variance only for the Crowley et al. data whereas the reverse was true for Items x Occasion. The former could reflect the fact that Crowley et al. used the full 27-item CDI whereas we used a 10-item version in which items had been selected to maximise internal consistency - hence minimizing error due to the items. The finding that I x O variance was notable in the present study (4.5%) but close to zero in Crowley et al.'s study probably reflects the greater number of occasions and larger test retest intervals. We assessed three times at 24-month intervals compared with Crowley et al. who assessed on two occasions 7 months apart. The rapid developmental changes in children means that item difficulty varies at different ages.

COMMENT: Provide rationale for imputing data as mentioned in the Measure section. Particularly given that there are only 10 items. Did you input data for those who were only missing 1 item, or more items? Up to how many items per individual were imputed? How does that potentially influence the results? Do the results change if you compare the imputed

data with a sample without the imputed data? I also suggest providing citations to justify the imputing of data.

RESPONSE: (a) We have added the following new text to the Participants section on p.8.

The present sample included 668 participants who have completed the CDI three times with a two/three year interval between assessments at ages 9, 11, and 14 years and who had no more than 5 missing items at any one assessment.

(b) We also added the following text to Measure section on p.8-9 plus new reference.

Missing values for individual items were imputed for up to five items per individual/time point using person mean substitution where missing values were estimated based on the mean for every person at that particular time point (Huisman, 2000). Of a total group of 875 participants screened for depression at age 9 the mean CDI score for those included compared with those excluded due to missing data at one or more time points was not significantly different (3.15 v 2.78, p=0.07).

COMMENT: 11. The authors mention that two items were indicated to be more state-like, however, these two items also increase the reliability of the measure. What does this mean? Expand on this.

RESPONSE: We have added a new sentence to the Discussion at bottom of p.15. as follows.

However, these two items also increase the reliability of the measure which might reflect the fact that these two items were relatively stable as reflected by their G-coefficient when the 244 children with higher levels of symptoms were analyzed separately (see Table 5).

COMMENT: 12. What does it mean that the sub-group results indicated that a different item had more sensitivity to change and that the other two items that initially did no longer appeared to be as sensitive? Extend this part of the discussion to discuss why this might be and what that means?

RESPONSE: We have extended the Discussion to discuss this by adding the following text to para. 2 on p15.

Interestingly, …. In this regard it is notable that the Person x Item variance component was larger for the entire sample (10.6%) than for the subgroup with high levels of symptoms (3.4%). Burns (1998, p.85) explains the P x I variance as due to 'differences in the ordering of subjects on different items'. This raises the possibility that some items might work differently for healthy and depressed participants or Differential Item Functioning - an issue that we intend to pursue in the future using item response theory (Siegert, Walkey & Turner-Stokes, 2009)…….

COMMENT: 13. Extend discussion on similarity/differences in results of current study and previous study by Crowley et al. (1994).

RESPONSE: We have extended the discussion comparing our findings with Crowley et al. See response to Reviewer 2 # 9 above.

COMMENT: 14. Is there evidence to suggest that the state aspects of depression are more amenable through treatment than the more trait aspects? This is stated multiple times throughout the introduction and discussion, but no research is cited.

RESPONSE: This is an excellent point made by the reviewer and really got us thinking. We have added new text in two places in the Discussion to emphasize this is an assumption and also acknowledge an alternative perspective.

(a) Added to Discussion p.16:
We are assuming here that state symptoms are more amenable to treatment than trait symptoms of depression but that also needs to be established. Indeed highly dynamic symptoms might constitute a more difficult 'moving target' for therapy.

(b) Final two sentences on p.19 modified to acknowledge assumption identified by reviewer.

This would permit a sharper distinction between risk factors for depression, which are the trait features and might be less susceptible to modification, and dynamic features which we assume to be more amenable to change and might be the likely targets for intervention. It raises the prospect of a state-trait scale for depression in which the state scale would have maximal responsiveness to change.

Minor Concerns:

COMMENT: 1. Please add in the citation for the CDI-10 in the method section of the Abstract.

RESPONSE: Citation added to Abstract.

COMMENT: 2. There is an incorrect use of APA style in regards to a citation in the first sentence of the last paragraph in the introduction.

RESPONSE: We have corrected this citation.

COMMENT: 3. In the Abstract Method second, please add a comma or dash in the following sentence to help distinguish the sentence from the name of the measure, for example "… in a widely used screening measure of depression – the 10-item Children's Depression Inventory."

RESPONSE: We have added the dash as suggested to the Method section of the Abstract.

COMMENT: 4. The running head is in a different font than the manuscript. Please correct. Also, it is not in correct APA style. The words "Running head" should only appear on the title page. Then the rest of the headings should just include "CDI GENERALIZABILITY THEORY." Please include a correct title page (APA style).

RESPONSE: We have corrected the font and font size of the running head to Times New Roman 12 and altered the running head to be consistent with the APA Publication Manual 6th ed.

COMMENT: 5. There is a comma missing in the second citation in the second paragraph in the participants section of the manuscript.

RESPONSE: We have added a comma to this citation.

COMMENT: 6. Please specific what ages the CDI designed for? Is it all children, ages 1 day old through 18? I doubt it, but it is not clear. What is the reading level of the measure?

RESPONSE: Under the heading Measure on p.8 we have modified the opening sentence to read as follows:

The CDI (Kovacs, 1981) is a 27-item self-report measure of symptoms of depression in children from age 7 to 17.

In summary we have addressed each and every point that was raised by both reviewers and the manuscript is substantially improved. To this end I have added a sentence thanking and acknowledging the input of two anonymous reviewers. I trust that this is appropriate and satisfactory to you. We look forward to hearing from you in due course concerning the final decision on our manuscript.

Kind regards

Richard J. Siegert

Highlights:

- Generalizability Theory provides a robust method for estimating the state and trait components of psychometric measures.
- We used G Theory analysis on the 10 item Children's Depression Inventory with a large sample of Pacific Island children who were assessed at three ages.
- The CDI-10 demonstrated acceptable generalizability for the target population across occasions with about one third of variance in total scores attributed to dynamic and two thirds to enduring aspects of depression.
- Using G it may be possible to refine existing instruments or develop new ones where there is a clearer distinction between state and trait symptoms and the items representing them.

Abstract

Background: The distinction between temporary versus enduring or state/trait aspects of depression is important. More precise distinction would improve understanding of the aetiology of depression and those aspects most amenable to intervention thus identifying more homogeneous, dynamic targets for clinical trials. Generalizability Theory has been proposed as useful for disentangling state and trait components of psychopathology.

Methods: We applied Generalizability Theory to determine the relative contributions of temporary and enduring aspects of depression in a widely used screening measure of depression the - 10-item Children's Depression Inventory (CDI-10; Kovacs, 1985). Participants were children of Pacific Island descent living in New Zealand (*n*= 668). Data were collected at ages - 9, 11, and 14 years.

Results: The CDI-10 demonstrated acceptable generalizability across occasions (G=.79) with about one third of variance in total scores attributed to temporary and two thirds to more enduring aspects of depression. There were no other significant sources of error variance. Two items were identified as more sensitive than the remaining eight to more dynamic symptoms.

Limitations: Studies with briefer test-retest intervals are warranted. Use of this Pacific Island cohort limits generalisability of findings to other cultures and ethnicities. No data were collected on whether participants had received intervention for depression.

Conclusions: While the CDI-10 reliably measures both stable and transient aspects of depression in children, the scale does not permit clear distinction between them. We advocate application of Generalizability Theory for developing state/trait depression measures and determining which existing measures are most suitable for capturing modifiable features of depression.

Key words: Depression; Generalizability Theory; Children's Depression Inventory; State; Trait

Distinguishing transient versus stable aspects of depression in New Zealand Pacific Island children using Generalizability Theory

Janis Paterson, School of Public Health and Psychosocial Studies, Auckland University of Technology, Auckland, New Zealand; Oleg N. Medvedev, Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand; Alexander Sumich, Division of Psychology, Nottingham Trent University, Nottingham, United Kingdom and School of Public Health and Psychosocial Studies, Auckland University of Technology; El-Shadan Tautolo, School of Public Health and Psychosocial Studies, Auckland University of Technology; Christian U. Krägeloh, School of Public Health and Psychosocial Studies, Auckland University of Technology; Rose Sisk, Biostatistics and Epidemiology, Auckland University of Technology; Robert K. McNamara, Department of Psychiatry and Behavioral Neuroscience, University of Cincinnati College of Medicine; Michael Berk, Deakin University, IMPACT Strategic Research Centre, School of Medicine, Geelong, Victoria, Australia; Ajit Narayanan, School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology; Richard J. Siegert[1], School of Public Health and Psychosocial Studies, Auckland University of Technology.

[1] Corresponding author: Richard J. Siegert, Auckland University of Technology, Psychology, School of Public Health and Psychosocial Studies, Room AR321, 90 Akoranga Drive, Northcote, Auckland 0627, Private Bag 92006, Auckland 1142, New Zealand, +64 9 921 9999 x. 7885, richard.siegert@aut.ac.nz.

Introduction

The distinction between dynamic and more stable aspects of behavior is well established in psychology (Chaplin et al., 1988). Generally, relatively stable or enduring aspects of an individual's behavior are referred to as a *trait*, while more dynamic aspects are referred to as a *state*, meaning changeable behavior that is much more context bound or situation specific. The distinction between such stable and dynamic aspects of behavior has sometimes been demonstrated by comparing test - retest correlations for the state and trait subscales of a measure, e.g., the State-Trait Anxiety Inventory (STAI; Gaudry et al., 1975), with a higher correlation expected for trait than state components. For example, Barnes et al., (2002) reviewed the reliability of the STAI and found a mean test-retest correlation of .88 (range .82 - .94) for trait anxiety and .70 (range .34 - .96) for state anxiety, based on seven studies reporting test-retest reliability. An obvious limitation of this approach to distinguishing state from trait is that it uses total score correlations and fails to examine the relative strengths of individual items.

An alternative approach to assessing state versus trait has been to use structural equation modelling to compare the covariance structures of a state/trait measure on at least two separate occasions (Cole, Martin & Steiger, 2005). Using this "latent trait-state-occasion model" with the Beck Depression Inventory II (BDI-II) in two large samples of young people Wu reported a trait component accounting for more than 50% of variance and an 'occasion-specific factor' that explained between 7 – 12% of variance (Wu, 2016, p.39). LaGrange et al. (2011) applied the latent trait-state-occasion model to four measures of depressive cognitions and symptoms in a longitudinal study of 515 children and adolescents and found three of these measures characterised by "two types of longitudinal factors: a time invariant (or trait-

like) factor and a series of time-varying (more state-like) factors" (p.13). Of particular relevance to the present study was their fourth measure - the Children's Depression Inventory - was almost entirely accounted for by time varying or state like variance. While this approach has the advantage that it can compare the strength of the loadings of individual items on state or trait factors at different points in time it fails to partial out the precise proportion of variance across occasions due to the person, the occasion and the item (as well as their interactions). A method which does exactly this, namely to account for all major sources of variance in a measurement situation, and arguably (Medvedev et al., 2017) should be the preferred technique for demonstrating a state-trait distinction is Generalizability Theory.

Generalizability Theory

Generalizability or G Theory was developed by Cronbach and represents a major advance upon Classical Test Theory (CTT) particularly in regard to evaluating reliability (Cronbach et al., 1963). While CTT conceptualizes test scores as the sum of a true score and error variance, G Theory uses analysis of variance (ANOVA) to calculate precise estimates for the error variance due to each important measurement *facet*, where facet refers to a distinct element that might influence variance and error in test scores in any testing situation. For example, the persons tested, the test items and the testing occasion are three examples of facets. Hence, whereas CTT restricts analysis of reliability and error variance to a single element such as the test items (i.e., Cronbach's alpha), the occasion (test-retest) or the rater (inter-rater), G Theory permits the researcher to break reliability down into all the important facets contributing to measurement error in a single analysis. In G terms the variance associated with participants or persons is considered the central concern and is known as the *differentiation facet* with other facets (e.g., items, occasion, rater) viewed as sources of measurement error.

An important step in a Generalizability analysis is the calculation of individual variance components for each facet (e.g., person, item, occasion) in the measurement situation. These variance components are first calculated by a conventional ANOVA and corrected by a method known as 'Whimbey's correction which accounts for the type of sampling involved (i.e. random, fixed or random finite). This also includes calculation of a *G coefficient* which represents how generalizable the test scores are across different situations. Bloch and Norman (2012, p. 968) describe the G coefficient as the ratio of signal to noise or 'true' variance to 'true + error variance' and provide details on its derivation and calculation. In most Generalizability studies there are two stages known as the G-study and the D-study. The G-study involves a standard factorial ANOVA with the calculation of variance components for each facet and for their interactions (rather than the significance tests typically associated with an ANOVA). The D-study allows the researcher to then estimate the impact on reliability of variations in different facets such as increasing the number of participants or the number of items in a scale.

As an illustration, medical education is one setting where G theory has been widely used (e.g., Prion et al., 2016). For example, when nursing students are examined on clinical skills, they might encounter ten structured clinical problems (or *stations*) that they are required to respond to in the presence of an examiner who rates their performance on a 1-10 scale for each problem. Assuming a different examiner (i.e. rater) at each station, a G Theory approach permits us to calculate the variance component for each facet – namely student, station, and rater - and their interactions. Consider another example from a mental health setting. In examining the inter-rater reliability of a new rating scale for assessing the severity of depression in outpatients we might have the following facets: Person (the differentiation facet), rater, and occasion. The calculation of variance components for each facet and their interactions allows us to determine precisely how much variance is due to the patient (Person

facet) and changes in their condition (Person x Occasion) as well as the reliability of raters. High inter-rater reliability will be reflected in a very small variance component for the facet Rater. In the present article we argue that G Theory may have particular value for distinguishing between state and trait or stable versus enduring aspects of behaviour and depression in particular.

Stable Versus Transient Features of Depression

A key issue for measuring change in symptoms of psychopathology is the extent to which the underlying construct is stable or changes over time – the state-trait distinction. This distinction between stable and dynamic symptoms has been applied less often for conceptualizing depression than for anxiety. However, there has been increasing interest recently in separating dynamic versus more stable aspects of depression from both neurobiological and personality perspectives (Bhagwar et al., 2002; Graham et al., 2013; Natoli et al., 2016; Wu 2016). In a recent psychometric study, Wu (2016) tested a latent state-trait-occasion model using the Beck Depression Inventory II (BDI-II) in two cohorts, of adolescents and young adults, concluding that the BDI-II measures "both trait-like and occasion-specific variances for individuals during late adolescence to early adulthood" (p.48). Lakes and Hoyt (2009) have argued that G Theory is a particularly useful conceptual framework for measuring the stable and transient components of a latent construct among adolescents where development is paramount. A recent paper by Chavez et al., (2016) demonstrated this approach with the Adolescent Quality of Life-Mental Health Scale (AQOL-MHS) in adolescents attending mental health clinics assessed on three occasions over eight months. The focus of that study was on the reliability of change scores and the authors used a method developed by Cranford et al. (2006) based on G Theory to compare conventional estimates of reliability (Cronbach's alpha, test-retest) with GT estimates. The authors observed that the reliability of change assessed by G Theory for the three subscales of

the AQOL-MHS was adequate but consistently lower than Cronbach's alpha for each time point. They concluded Cronbach's alpha is sufficient when comparing scores across people at the same point in time but the reliability of change scores must be considered when comparing the change in persons across multiple time points.

The Children's Depression Inventory

The Children's Depression Inventory is a widely used measure of depression in children and adolescents. It is a 27-item self-report questionnaire for quantifying the severity of symptoms of depression in young people aged 7 – 17 years. We are aware of only one previous article using G Theory to investigate the psychometric characteristics of the CDI (Crowley et al., 1994). In that study they administered the full 27-item CDI twice, with a 28-week retest interval, to 164 children aged 11 – 16 years (mean age 12.6) in Texas schools. Crowley et al., (1994) found a relatively low generalizability coefficient for a single CDI administration (.63) compared with repeat testing (.81) and advocated multiple assessments before diagnosing depression in a child. They noted that the largest source of error in their analysis was the Persons x Items interaction suggesting that the CDI items were interpreted differently by different participants. The variance component for the P x I interaction (17.3%) in that analysis was notably larger than that for Persons (13.4%), Items (5.8%) and Persons x Time (5.2%).

In a recent paper Medvedev et al. (2017) argued that G Theory offers unique advantages for separating the dynamic from stable dimensions of a latent trait. They argued that the variance due to the person facet in a repeated measures test situation represents trait behavior and that state behavior is indexed by the variance due to the person x occasion interaction. Moreover, a state measure should be characterized by low person and low occasion variance, but a relatively large person x occasion interaction variance component.

Conversely, a trait measure should display large amounts of variance due to the person facet and low person x occasion variance. In the same article Medvedev et al. proposed a new index for characterizing the state (state component index, SCI) and trait (trait component index, TCI) components of a measure. G Theory produced variance components consistent with this interpretation and thereby demonstrated the construct validity of trait and state behavior. In the present paper, we use G Theory to examine the short version of the CDI for state versus trait characteristics in a large cohort of Pacific children born in Auckland, New Zealand. The brief CDI was originally chosen for this longitudinal cohort study because its brevity made it practical to include in a much longer battery of assessment tools. The availability of data from three points in time made it particularly suitable for the present Generalizability analysis. The aims of the present research were: (1) to examine the psychometric properties of the 10-item version of the CDI using G Theory and (2) to explore the degree to which state and trait or stable versus transient aspects of depression can be quantified for this scale.

Methods

Participants

All participants were children of Pacific Island descent living in New Zealand participating in the longitudinal Pacific Islands Families Study https://niphmhr.aut.ac.nz/research-centres/centre-for-pacific-health-and-development-research/pacific-islands-families-study. The Pacific Islands Families (PIF) Study is a longitudinal investigation of Pacific children born in Auckland, New Zealand, and their parents. All potential participants from Middlemore hospital in South Auckland were selected from births where at least one parent identified as being of a Pacific ethnicity and was a NZ permanent resident. The original cohort included 1,376 mothers of 1,398 Pacific infants

(including 44 twins). Children and their mothers were followed up at six-weeks (baseline data collection), 1, 2, 4, 6, 9, 11 and 14 years postpartum. Compared with data available from Statistics New Zealand's 1996 and 2001 censuses, the inception cohort was broadly representative of the Pacific census figures (Statistics New Zealand, 2002; Paterson et al., 2008).

The present sample included 668 participants who have completed the CDI three times with a two/three year interval between assessments at ages 9, 11, and 14 years and who had no more than 5 missing items at any one assessment. This sample of 668 had the same distribution of Pacific Island ethnic groups as the remainder of the original birth cohort (N=730) but a higher female:male ratio (53:47% versus 45:55%, p=0.007). This sample was previously reported to have a prevalence of 7.3% for depressive symptoms at age nine years (Paterson et al., 2014) based on a T-score > 65 as the cut off score for depression. The sample size of 668 is large enough to satisfy criteria for both generalizability analysis (Bloch and Norman, 2012) and a reliability study in health research (Shoukri et al., 2004). Ethics approval has been granted for all phases of the longitudinal PIFS study by the Auckland Regional Ethics Committee.

Measure

The CDI (Kovacs, 1985, 2003) is a 27-item self-report measure of symptoms of depression in children from age 7 to 17. This study used the 10-item brief version a screening tool for depression. These items were selected from the full 27 by removing the one item that would most attenuate Cronbach's alpha in a stepwise procedure until a reliable set of 10 items remained (Kovacs, 1985, 2003). The brief CDI has ten items scored on a 0 – 2 Likert scale with five items reverse scored. Scores can range from 0 to 20 with a higher score representing greater depression. Missing data were below 1%. Missing values for individual

items were imputed for up to five items per individual/time point using person mean substitution where missing values were estimated based on the mean for every person at that particular time point (Huisman, 2000). Of a total group of 875 participants screened for depression at age 9 the mean CDI score for those included compared with those excluded due to missing data at one or more time points was not significantly different (3.15 v 2.78, p=0.07).

Generalizability Analyses

Generalizability analyses were conducted following the guidelines described elsewhere (Medvedev et. al., 2017) and employed EduG 6.1-e software (Swiss Society for Research in Education Working Group, 2006). Both the G (generalizability) and D (decision) studies used a person (P), by item (I), by occasion (O) random effects design expressed as P × O × I, where the I facet is fixed and the P and O facets are infinite. Persons were the object of measurement defined as a differentiation facet and not a source of error and items and occasions were instrumentation facets (Cardinet et al., 2010). Error variance due to person x occasion interaction in a scale score can be interpreted as reflecting a state component or a scale's sensitivity to state changes (Medvedev et al., 2017). Table 1 includes definitions of components used for two-facet generalizability analysis.

Table 1 about here

Variance components for each facet and their interactions were computed based on traditional ANOVA estimates using equations introduced by Brennan (1977, 1992).

Whimbey's correction coefficient was applied to traditional ANOVA estimates that consider finite facets such as items that are not derived from infinite populations (Cardinet et al., 2010). Whimbey's correction has no effect on facets derived from infinite populations (e.g., persons) and is expressed as (N(f)−1)/N(f), where N(f) is the population size of the f facet in the G-study design. The unique contribution of each facet to the total variance of universe scores was estimated using generalizability analysis and included relative and absolute error variance and G-coefficients for the differentiation facet (persons). The relative G-coefficient is commonly expressed as ρ^2, $\square^2$ or intermediate value if Whimbey's correction is applied (Cardinet et al., 2010) and only considers variance directly related to the object of measurement. The absolute G-coefficient is equivalent to Phi (Φ) coefficient after Whimbey's correction is applied and accounts for other variance sources (e.g., item x occasion interaction) that may influence an absolute measure indirectly (Cardinet et al. 2010). A state component index (SCI) and trait component index (TCI) were computed using formulas developed by Medvedev et al. (2017) that reflect the proportion of variance attributed to a dynamic and a stable component in a measure:

$$\text{SCI} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_t^2} \qquad \text{TCI} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_s^2}$$

In these formulas, variance due to the object of measurement or person is $\sigma_t^2 = \sigma_p^2$ and variance due to temporal changes across 3 time points of assessment in this study refers to person-occasion interaction $\sigma_s^2 = \sigma_{po}^2$. It can be seen that these coefficients have inverse relationship and they are not affected by error due to other sources (e.g., item-person interaction), which is the major benefit of using them to distinguish between dynamic and stable components in a measure (Medvedev et al., 2017). The D-study investigated properties of individual facets by varying facet designs aimed at optimizing measurement.

After the initial analysis with all 668 participants we repeated the G analysis with a sub-group of 224 children who all displayed at least some depressive symptoms (score ≥ 4) on the CDI) at Time 1 aged 9 years. We chose a score of 4 since the full CDI has 27 items with 12 as the cut-off point for depression and 4 on a 10 item version represents a roughly equivalent cut-off point. This group of 244 children all scored above the 63rd percentile of the whole sample. Our rationale for this post-hoc, sub-group analysis was that the sample is a community sample comprising mostly healthy or non-depressed young people – and this could exaggerate the trait component of variance. Consequently we attempted to confirm the initial results in a sub-group of children who all reported some symptoms of depression.

Results

We report the results of the analyses for the full sample of 668 children first and then report the sub-group analysis for the 244 children with a score ≥ 4 on the brief CDI. Descriptive statistics for the CDI data collected at three time points with two and three year intervals are presented in Table 2. The average internal consistency or Cronbach's alpha over three occasions was .71 and test-retest reliability with a two and five year interval was .41 and .22 respectively (both $p < 0.001$). The mean total CDI score was significantly higher at Time 1 (age 9 years, 3.14, ± 2.73, range 0-14), than at Time 2 (11 years, 2.05, ± 2.30, range 0-12) and Time 3 (14yrs, 1.73, ±2.65, range 0-20). The data were normally distributed with a mean total CDI score across three occasions of 2.31 (±2.63, range 0-20). Principal component analysis was conducted to examine the CDI structure with the full data set. Unidimensionality of the measure was supported by the scree plot showing a clear cut-off point after the first principal component, and item-component loadings ranged from .47 (item 10) to .63 (item 4).

Table 2 about here

G-Study

Table 3 includes classical ANOVA estimates for person (P), item (I), occasion (O) and their interactions. Whimbey's correction was applied to estimate corrected variance components presented in columns 7 and 8 (in %). It can be seen that the largest amount of error variance is due to person x item x occasion interaction, which is influenced by person x occasion, person x item, and item x occasion interaction in the traditional ANOVA. Therefore, these estimates were used in our G-study to extract error variances contributed uniquely by each potential source of error and compute absolute and relative G-coefficients as presented in Table 3. Table 3 also includes the variance components reported by Crowley et al. (1994) for the 27-item CDI (n=164) for comparison and for a sub-group analysis of 244 symptomatic children.

Table 3 about here

Table 4 about here

In a G-study, the object of measurement (persons) refers to a differentiation facet that is treated independently. SCI and TCI were computed using the relevant person (P) and person x occasion (P x O) estimates from table 4. Person x occasion interaction is reflecting individual state (Medvedev et al., 2017) and suggests that the CDI has a degree of sensitivity to individual changes over time (SCI= .33). However, the trait component index (TCI = .67)

indicated that the CDI satisfies criteria proposed for a valid trait measure. Overall, the scale shows acceptable generalizability of .79 for both the absolute and relative Coefficient G (Arterberry et al., 2014; Cardinet et al., 201) for the target population over time. Table 2 shows that the only relevant source of error variance is person x occasion interaction that comprises 100% of error variance, absolute and relative. Person x occasion interaction is reflecting individual state (Medvedev et al., 2017) and suggests that the CDI has a degree of sensitivity to individual changes over time (SCI= .33). However, the trait component index (TCI = .67) indicated that the CDI satisfies criteria proposed for a valid trait measure.

D-Study

Facets analysis was conducted to estimate G-coefficients for every individual item, which are presented in Table 5. Lower value G-coefficients (e.g., G< .50) reflect an item sensitive to state changes. Item 4 '*I hate myself/I do not like myself/I like myself' '* and item 7, '*I look O.K./There are some bad things about my looks/I look ugly'* both have G-coefficients about .20 (absolute and relative) meaning that these items have higher sensitivity to changes over time. Removing these items from the scale resulted in a decrease of the variance explained by person-occasion interaction to 64.8%. However, the overall G-coefficient has also decreased to .72 (absolute and relative), which suggests that these items make an important contribution to the overall reliability of the measure. All other items exhibit G-coefficients in the range from .40 to .51 indicating a modest sensitivity to temporal state changes. Removing any of the occasions from observation design produced a slight drop of the overall generalizability coefficients (i.e. G=.72) and provided support for the original measurement design with 3 time points.

Table 5 about here

Sub-Group Analysis

For the 244 young people with scores ≥ 4 on the CDI the G coefficient both relative and absolute was 0.88. The variance components for the ANOVA for this analysis are included in Table 3 (n=244) alongside those for the full sample and for Crowley et al.'s (1994) study. The absolute and relative G coefficients for individual items are included in Table 5 in parentheses alongside the results from the full sample. The Trait Component Index was .97 (SCI = .03).

Discussion

This study examined the psychometric properties of the 10-item version of the CDI-10 using G Theory and to determine the extent to which the scale has the capacity to parse depression as a temporary or enduring construct. The CDI-10 was initially developed as a rapid screening tool to quantify the severity of depressive symptoms in children for clinical and research purposes (Kovacs, 1985). The measure was found to have a G coefficient of .79, which is very close to the accepted criterion of .80 for good generalizability indicating that it measures predominantly enduring depression symptoms. Given the very long intervals between testing and the rapid developmental changes occurring in children from ages 9 to 14 this seems to indicate a high degree of generalizability for scores on the 10 item CDI. While the CDI also captures significant proportions of dynamic variance in scores with about one third of variance in the scores attributed to dynamic and two thirds to stable aspects of depression, the scale structure does not permit clear distinction between these aspects. This raises the question of whether a measure of predominantly stable traits is the best measure for those situations where change is a primary interest such as treatment studies. Interestingly Wu's (2016) structural equation modelling study of the Beck Depression Inventory – II using

a 'state-trait-occasion model' reported that it measured both 'trait-like and occasion-specific variance' (p.48) and the trait variance was greater than 50%. La Grange et al. who examined the full 27-item CDI using a structural equation model concluded that the CDI was mostly measuring state variance. In contrast, the present results suggest that the brief CDI is a good measure of relatively stable individual differences in depressive symptoms while also being reasonably sensitive to more dynamic or occasion specific depressive symptoms.

Interestingly the G coefficient for this sample seemed somewhat in contrast to the relatively low test-retest correlations of 0.41 at two years and 0.22 at four years test-retest interval. However, these results are comparable with previous findings. For example, Saylor, Finch, Spirito and Bennett (1984) reported a one week test-retest correlation for the full CDI of .38 for 69 'normal' 5th and 6th graders and .59 and .87 for two samples of 'emotionally disturbed' children. Similarly, Nelson and Politano (1990) reported a test-retest coefficient of .47 with a 30 day test-retest interval for 96 psychiatrically hospitalised boys and girls. Certainly, further work examining the relationship between generalizability, as reflected in the G coefficient, and temporal stability, as measured by test-retest correlations, seems indicated.

The present study was also able to examine each of the 10 CDI items in terms of their state-trait properties. Two items (items 4 and 7) were notably more unstable or more sensitive to dynamic changes than the remaining eight. These were '*I hate myself/I do not like myself/I like myself*' and '*I look O.K./There are some bad things about my looks/I look ugly''*. However, these two items also increase the reliability of the measure which might reflect the fact that these two items were relatively stable as reflected by their G-coefficient when the 244 children with higher levels of symptoms were analyzed separately (see Table 5). In the case of the brief CDI, there is little scope for removing any of the 10 items comprising it as they are all essential to maintain internal reliability. However, G Theory could be fruitfully

used with other longer depression measures to establish which items are most important for capturing state or trait elements of mood. Importantly, it could be used as a component of item analysis in developing new measures specifically intending to measure separate state and trait dimensions of a latent variable. Similarly, it could be used to compare among standard measures of depression to establish which is most able to capture state versus trait aspects of depression. Some individual items may serve as trait vulnerability markers, like trait neuroticism, and these could be used to index risk. So when individuals score high on several of the most stable or trait-like items it could be used as an indicator that these children are most prone to depression and might benefit from early intervention. On the other hand, more dynamic items might better be used in clinical trials, as more sensitive markers of response. It might even be possible to determine which symptoms of depression (i.e., cognitive, emotional, motivational, somatic) are more or less amenable to change or treatment. We are assuming here that state symptoms are more amenable to treatment than trait symptoms of depression but that also needs to be established. Indeed highly dynamic symptoms might constitute a more difficult 'moving target' for therapy.

We also repeated the G analysis with a sub-group of 244 children who demonstrated symptoms of depression at the age of 9. The results were similar although the G coefficient increased to .88 suggesting that with a clinical sample the 10 item CDI is even more of a trait measure. Interestingly, when the G coefficients for individual items were considered, items 4 and 7 in the depressed sub-group did not appear to be as sensitive to state changes as they were in the full sample. However, item 8 (loneliness) showed greater sensitivity. In this regard it is notable that the Person x Item variance component was larger for the entire sample (10.6%) than for the subgroup with high levels of symptoms (3.4%). Burns (1998, p.85) explains the P x I variance as due to 'differences in the ordering of subjects on different items'. This raises the possibility that some items might work differently for healthy and depressed participants

or Differential Item Functioning - an issue that we intend to pursue in the future using item response theory (Siegert, Walkey & Turner-Stokes, 2009). This highlights the potential for using G analysis in the future for developing new, or refining existing instruments, so as to maximize their capacity to accurately reflect mood changes in those with and without clinically relevant depression.

Although the PIFS longitudinal dataset provided a unique opportunity to estimate the extent to which depressive symptoms can be considered relatively stable or dynamic through development (9 to 14 years), this study is not without its limitations. For example, data from such long inter-assessment intervals (2-3 years) might not generalise to fluctuations in depression across shorter intervals (weeks, months). Many depression questionnaires ask individuals to consider the preceding 2-4 week period, and typical intervention studies often take place over a period of 2-6 months. Thus, further studies with briefer test-retest intervals are warranted, and these might also compare the stability of depression to other emotional constructs (e.g., anxiety, anger). At the same time it is worth noting that a fairly similar pattern of variance components to those found in the present study were reported by Crowley et al. (1994) for a 28 week test-retest interval using the full 27 item CDI (see Table 3). The most notable differences were that Items accounted for a substantial amount (5.8%) of error variance only for the Crowley et al. data whereas the reverse was true for Items x Occasion. The former could reflect the fact that Crowley et al. used the full 27-item CDI whereas we used a 10-item version in which items had been selected to maximise internal consistency - hence minimising error due to the items. The finding that I x O variance was notable in the present study (4.5%) but close to zero in Crowley et al.'s study probably reflects the greater number of occasions and larger test retest intervals. We assessed three times at 24 month intervals compared with Crowley et al. who assessed on two occasions 7 months apart. The rapid developmental changes in children means that item difficulty varies at different ages.

In addition, while the need to better understand the aetiology, prevalence and progression of depression in Pacific Island adolescents has been previously highlighted (Wyatt et al., 2015), use of the PIFS cohort limits generalisability of findings to other cultures and ethnicities, which should be investigated in future research. New Zealand has a high lifetime prevalence of depression by international standards (Kessler and Bromet, 2013) and Pacific Island youth within New Zealand have about double the rate of depression and suicide attempts of the general population (Statistics New Zealand and Ministry of Pacific Island Affairs, 2011). A further limitation of the current study is that no data were collected on whether participants had received intervention for depression, as this was not an initial focus of the PIFS study.

Finally, the use of a brief, 10-item screening measure for depression provided a limited range of depression items for analysis. Use of the complete 27-item CDI would provide considerably more information on the extent to which items can range in their degree of 'stateness' versus 'traitness', and should be further investigated. However, even with only 10 items, we were able to demonstrate proof of principle for the potential of a G Theory approach to item analysis in this regard. First, in quantifying the state and trait variance inherent in the overall measure and second in identifying those individual items most sensitive to state or dynamic aspects of mood.

The major contribution of the present study is in highlighting the potential value of G Theory for research on the state-trait distinction In particular its potential value for improving measures used for screening for mood disorders and tracking progression. This has important implications for the measurement of depression and for the design of clinical trials. In measuring depression, it may be possible to refine existing instruments or develop new ones where there is a clearer distinction between dynamic and stable symptoms and the items reflecting them. This would permit a sharper distinction between risk factors for depression,

which are the trait features and might be less susceptible to modification, and dynamic features which we assume to be more amenable to change and might be the likely targets for intervention. It raises the prospect of a state-trait scale for depression in which the state scale would have maximal responsiveness to change.

REFERENCES

Arterberry, B. J., Martens, M. P., Cadigan, J. M., & Rohrer, D. (2014). Application of generalizability theory to the big five inventory. *Personality and individual differences*, *69*, 98-103.

Barnes, L. L., Harp, D., & Jung, W. S. (2002). Reliability generalization of scores on the Spielberger State-Trait Anxiety Inventory. *Educational and Psychological Measurement*, *62*(4), 603-618.

Bhagwagar, Z., Whale, R., & Cowen, P. J. (2002). State and trait abnormalities in serotonin function in major depression. *The British Journal of Psychiatry*, *180*(1), 24-28.

Bloch, R., & Norman, G. (2012). Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. *Medical teacher*, *34*(11), 960-992.

Brennan, R. L. (1977). *Generalizability analyses: Principles and procedures*. Technical Bulletin No. 26, Research and Development Division, American College Testing Program.

Brennan, R. L. (1992). Elements of generalizability theory. (2nd ed.). Iowa City: ACT Publications.

Burns, K. J. (1998). Focus on Quantitative Methods-Beyond Classical Reliability: Using Generalizability Theory to Assess Dependability. *Research in Nursing and Health*, *21*, 83-90.

Chaplin, W. F., John, O. P., & Goldberg, L. R. (1988). Conceptions of states and traits: Dimensional attributes with ideals as prototypes. *Journal of personality and social psychology*, *54*(4), 541-557.

Chavez, L. M., Garcia, P., Ortiz, N., & Shrout, P. E. (2016). Applying generalizability theory methods to assess continuity and change on the Adolescent Quality of Life - Mental Health Scale (AQOL-MHS). *Quality of Life Research*, *25*(12), 3191-3196.

Cole, D. A., Martin, N. C., & Steiger, J. H. (2005). Empirical and conceptual problems with longitudinal trait-state models: Introducing a trait-state-occasion model. *Psychological methods*, *10*(3), 3-20.

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably?. *Personality and Social Psychology Bulletin*, *32*(7), 917-929.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Mathematical and Statistical Psychology*, *17*, 137-163.

Crowley, S. L., Thompson, B., & Worchel, F. (1994). The Children's Depression Inventory: A comparison of Generalizability and Classical Test Theory analyses. *Educational and Psychological Measurement*, *54*(3), 705-713.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG.* New York: Routledge.

Gaudry, E., Vagg, P., & Spielberger, C. D. (1975). Validation of the state-trait distinction in anxiety research. *Multivariate Behavioral Research*, *10*(3), 331-341.

Graham, J., Salimi-Khorshidi, G., Hagan, C., Walsh, N., Goodyer, I., Lennox, B., & Suckling, J. (2013). Meta-analytic evidence for neuroimaging models of depression: State or trait?. *Journal of affective disorders*, *151*(2), 423-431.

Huisman, M. (2000). Imputation of missing item responses: Some simple techniques. *Quality and Quantity*, *34*(4), 331-351.

Kessler, R. C., & Bromet, E. J. (2013). The epidemiology of depression across cultures. *Annual review of public health*, *34*, 119-138.

Kovacs, M. (1985). The Children's Depression Inventory (CDI). *Psychopharmacology Bulletin*, *21*, 995-998.

Kovacs, M. & Staff, M. (2003). Children's Depression Inventory (CDI): Technical Manual Update. Multi-Health Systems. *Inc.: Toronto, Ontario, USA.*

LaGrange, B., Cole, D. A., Jacquez, F., Ciesla, J., Dallaire, D., Pineda, A., Truss, A., Weitlauf, A., Tilghman-Osborne, C. & Felton, J. (2011). Disentangling the prospective relations between maladaptive cognitions and depressive symptoms. *Journal of abnormal psychology*, *120*(3), 511.

Lakes, K. D. and Hoyt, W. T. (2009) Applications of Generalizability Theory to Clinical Child and Adolescent Psychology Research. *Journal of Clinical Child & Adolescent Psychology*, *38*(1), 144-165.

Medvedev, O, N., Krägeloh, C.U., Narayanan, A., & Siegert, R.J. (2017). Measuring Mindfulness: Applying Generalizability Theory to distinguish between state and trait. *Mindfulness*, 8(4), 1036-1046.

Natoli, A., Nelson, S. M., Lengu, K. J., & Huprich, S.K. (2016). Sensitivity to criticism differentially mediates the relationship between interpersonal problems and state and trait depression. *Personality and Mental Health*, *10*(4), 293-304.

Nelson III, W. M. & Politano, P. M. (1990). Children's Depression Inventory: Stability over repeated administration in psychiatric inpatient children. *Journal of Clinical Child Psychology*, *19*(3), 254-256.

Paterson, J., Iusitini, L. & Taylor, S. (2014). Pacific Islands Families Study: depressive symptoms in 9-year-old Pacific children living in New Zealand. *The New Zealand Medical Journal*, *127*(1390), 13-21.

Paterson, J., Percival, T., Schluter, P., Sundborn, G., Abbott, M., Carter, S., Cowley-Malcolm, E., Borrows, J., & Gao, W. (2008). Cohort profile: The Pacific Islands Families (PIF) Study. *International Journal of Epidemiology*, *37*(2), 273-79.

Prion, S. K., Gilbert, G. E., & Haerling, K. A. (2016). Generalizability Theory: An introduction with application to simulation evaluation. *Clinical Simulation in Nursing*, *12*(12), 546-554.

Saylor, C. F., Finch, A. J., Spirito, A., & Bennett, B. (1984). The Children's Depression Inventory: A systematic evaluation of psychometric properties. *Journal of consulting and clinical psychology*, *52*(6), 955-967.

Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: review and new results. *Statistical Methods in Medical Research*, *13*(4), 251-271.

Siegert, R.J., Walkey, F.H., & Turner-Stokes, L. (2009). An examination of the factor structure of the Beck Depression Inventory-II in a neurorehabilitation inpatient sample. *Journal of the International Neuropsychological Society*, *15*(1), 142-147.

Statistics New Zealand. (2002). Pacific progress: A report on the economic status of Pacific peoples in New Zealand. *Wellington: Statistics New Zealand*.

Statistics New Zealand and Ministry of Pacific Island Affairs (2011). *Health and Pacific peoples in New Zealand.* Wellington: Statistics New Zealand and Ministry of Pacific Island Affairs.

Swiss Society for Research in Education Working Group. (2006). EDUG user guide. *Euchatel, Switzerland: IRDP*.

Wu, P. C. (2016). Longitudinal stability of the Beck Depression Inventory II: A latent trait-state-occasion model. *Journal of Psychoeducational Assessment*, *34*(1), 39-53.

Wyatt, L. C., Ung, T., Park, R., Kwon, S. C. & Trinh-Shevrin, C. (2015). Risk factors of suicide and depression among Asian American, Native Hawaiian, and Pacific Islander youth: a systematic literature review. *Journal of Health Care for the Poor and Underserved*, *26*, 191 – 237.

Table 1. Definition of components for two-facet Generalizability analysis (P x I x O).

Persons (P)	Universe score for person *p* (averaged deviation from grand mean over items and occasions)
Items (I)	Effect for item *i* (averaged deviation from grand mean over persons and occasions)
Occasions (O)	Effect for occasion *o* (averaged deviation from grand mean over persons and items)
P x I	Interaction effect between person *p* and item *i* averaged over occasions
P x O	Interaction effect between person *p* and occasion *o* averaged over items
P x I x O, *e*	Interaction effect between person *p*, item *i* and occasion *o,* containing random error

Table 2. Descriptive statistics including mean, variance, standard deviation (SD), Cronbach's alpha (α) and test-retest coefficients for the CDI (n=668 x 3 occasions).

Item	Mean	Variance	SD	*α*	Test-retest *r*
1 How often sad	0.24	0.22	0.47	-	-
2 Will things work out	0.19	0.24	0.49	-	-
3 Does things wrong	0.25	0.25	0.50	-	-
4 How feels about self	0.26	0.25	0.50	-	-
5 How often feel like	0.14	0.16	0.40	-	-
6 How often bothered by	0.30	0.33	0.58	-	-
7 How feels about looks	0.20	0.21	0.46	-	-
8 How often feel alone	0.22	0.22	0.47	-	-
9 Amount of friends	0.29	0.33	0.57	-	-
10 Loved by anybody	0.22	0.24	0.49	-	-
Occasion					
1 (Baseline)	0.23	0.25	0.50	.64	-
2 (2 years later)	0.23	0.25	0.50	.68	.41
3 (4 years later)	0.24	0.25	0.50	.83	.22

Note: Grand mean= 0.23; Variance= 0.2

Table 3. ANOVA for the CDI including sum of squares (SS), degrees of freedom (df), mean squares (MS), variance components (in %) and standard errors (SE) using Person (P) x Item (I) x Occasion (O) design with interactions (*n*=668).

Source	SS	df	MS	Random	Mixed	Corrected[a]	%	SE[b]	% [c]	Crowley et al.[d]
P	523.33	667	0.78	0.02	0.02	0.02	7.6	0.00	14.1	13.4
I	42.31	9	4.70	0.00	0.00	0.00	0.0	0.00	0.0	5.8
O	0.25	2	0.13	0.00	0.00	0.00	0.0	0.00	0.0	0.2
P x I	1683.42	6003	0.28	0.03	0.03	0.03	10.6	0.00	3.4	17.3
P x O	220.68	1334	0.17	0.00	0.02	0.02	6.1	0.00	5.6	5.2
I x O	150.17	18	8.34	0.01	0.01	0.01	4.5	0.00	4.9	0.1
P x I x O	2328.90	12006	0.19	0.19	0.19	0.19	71.3	0.00	72.1	58.0
Total	4949.07	20039					100%		100%	100%

Note: [a] Corrected components are calculated by applying Whimbey's correction to the ANOVA estimates. [b] SE in the right column is related to the mixed effects presented in column 6.[c] Replication with subsample (*n*=244) that exhibit CDI score ≥4 [d] Data from the study of Crowley et al. (1994) conducted for the full 27-it

Table 4. Estimated variance components with standard errors (SE) and G-coefficients for the SDI G-study P x I x O design (n=668).

Source of variance	Differentiation variance	Relative error variance	% Relative	Absolute error variance	% Absolute
P	0.02	.....		.....	
I	.....	.....		(0.00)	0.0
O	.....	.....		(0.00)	0.0
P x I	.....	(0.00)	0.0	(0.00)	0.0
P x O	.....	0.01	100.0	0.01	100.0
I x O	.....	.....		(0.00)	0.0
P x I x O	.....	(0.00)	0.0	(0.00)	0.0
Sum of variances	0.02	0.01	100%	0.01	100%
Standard deviation	0.14	Relative SE: 0.07		Absolute SE: 0.07	
Coef_G relative	0.79				
Coef_G absolute	0.79				

Note: Grand mean = 0.23; Variance error of the mean for levels used = 0.00; SE of the grand mean = 0.01

Table 5. Absolute and relative G-coefficients for each individual item of the CDIS for the full sample of children and the symptomatic sub-group of children [1]

Items	Response option to score 2	G-Relative N=688 (N=244)	G-Absolute N=688 (N=244)
1	I am sad all the time	0.46 (0.39)	0.45 (0.38)
2R	Things will work out for me OK	0.51 (0.51)	0.48 (0.47)
3	I do everything wrong	0.48 (0.47)	0.47 (0.47)
4R	**I like myself**	**0.21 (0.38)**	**0.20 (0.37)**
5R	I feel like crying once in a while	0.45 (0.41)	0.44 (0.40)
6R	Things bother me once in a while	0.47 (0.48)	0.46 (0.48)
7	**I look ugly**	**0.22 (0.36)**	**0.21 (0.36)**
8	**I feel alone all the time**	**0.45 (0.22)**	**0.42 (0.20)**
9	I do not have any friends	0.42 (0.41)	0.40 (0.38)
10R	I am sure that somebody loves me	0.47 (0.45)	0.46 (0.45)

[1] Note: R indicates reverse scoring

Conflict of Interest Statement

The authors declare no conflicts of interests.

Author's contributions:

JP secured funding, collected data and contributed to writing up. OM led statistical analysis and contributed to writing up. AS contributed to study design and writing up. ET collected data, advised on culture, and contributed to writing up. CK contributed to design, statistical analysis, and writing up. RS managed the data set. RM contributed to study design and writing up. MB contributed to study design and writing up. AN advised on statistics and contributed to writing up. RJS secured funding, led the project and wrote the initial draft. All authors read and approved the final manuscript before submission.

Role of the Funding Source: The funders had no role in the collection, analysis or interpretation of data.

Acknowledgements: This research was conducted within the School of Public Health and Psychosocial Studies of the Faculty of Health and Environmental Sciences at the Auckland University of Technology. This article resulted from a meeting of the authors in April 2017 funded by a Royal Society of New Zealand Catalyst (Seeding General) Grant awarded to Richard Siegert in 2016 16-AUT-024-CSG. The funding covered travel and accommodation costs for AS, RM and MB. MB is supported by a National Health and Medical Research Council (NHMRC) Senior Principal Research Fellowship (grant number 1059660). The funding for the PIFS project was supported by the Ministry of Business, Innovation, and Employment (CONT-33797-HASTR-AIT). The funding covered data collection. The authors gratefully acknowledge the children and families who participated in the study, the Pacific Community Advisory Board and other members of the PIFS team. The authors would also like to thank two anonymous reviewers whose detailed and constructive comments substantially improved the manuscript.