The development and preliminary validation of a brief scale of emotional distress in young people using combined classical test theory and item response theory approaches: The Brief Emotional Distress Scale for Youth (BEDSY)

Author

Spence, Susan H, Rapee, Ronald M

Published

2022

Journal Title

Journal of Anxiety Disorders

Version

Accepted Manuscript (AM)

DOI

10.1016/j.janxdis.2021.102495

Title: The development and preliminary validation of a brief scale of emotional distress in young people using combined classical test theory and item response theory approaches: The Brief Emotional Distress Scale for Youth (BEDSY)

Susan H Spence[1]

Ronald M Rapee[2]

[1] Australian Institute of Suicide Research and Prevention and School of Applied Psychology, Griffith University, Brisbane, Queensland, Australia

[2]Department of Psychology, Centre for Emotional Health, Macquarie University, Sydney, New South Wales, Australia

Corresponding Author: Susan H Spence, Australian Institute of Suicide Research and Prevention and School of Applied Psychology,
Level 1, Building M24, Mount Gravatt Campus,
Griffith University, 176, Messines Ridge Rd., Mt Gravatt, QLD 4121, Australia
Email: s.spence@griffith.edu.au

Declarations of interest*:* none

ABSTRACT

Reliable, valid, and brief measures are important for identifying young people in community contexts who experience mental health problems. This paper reports the development and preliminary validation of the Brief Emotional Distress Scale for Youth (BEDSY), a measure based on anxiety and depression symptoms that load strongly upon the general construct of emotional distress. Participants, aged 11-17 years, included 2663 from a community population and 281 referred anxious youth. From a pool of 20 items, eight were selected for the final scale using methods from classical test theory, followed by item response theory (IRT). The final eight items met the pre-specified criteria for skewness and kurtosis, item-total correlations, IRT characteristics, and discrimination between referred vs. community samples. Exploratory structural equation modeling for a bi-factor model indicated that 81% of total variance was explained by the general emotional distress factor. The 8-item BEDSY showed strong internal consistency, good construct validity, and acceptable sensitivity and specificity in discriminating between a community sample vs anxious youth, and between youth with and without high levels of depressive symptoms. As such the scale has strong potential as a brief screen for identifying emotionally distressed young people in community contexts.
(193 words)

Keywords: emotional distress, anxiety, depression, children, adolescents, assessment, screening

1. Introduction

Brief measures that enable reliable and valid identification of young people with mental health problems are critical to facilitating the transition to appropriate early intervention. This is particularly the case for internalizing problems, such as anxiety and depression, which are difficult to observe and are characterized by very low levels of help-seeking. Although detailed assessment measures are required to accurately identify young people with mental health issues, to guide the content of intervention, and to evaluate outcomes, these scales are time-consuming for both participant and professional and require considerable resources. The use of such detailed measures is generally precluded in large-scale community screening of young people and in population-level research. Thus, there is a need for brief, standardized, and psychometrically sound instruments of high *sensitivity* that provide a valid indication of the possible presence of mental health conditions such as anxiety and depression. At the same time, such measures need have good *specificity* so that they do not falsely identify large numbers of young people without clinical levels of the problem concerned.

The brief assessment of anxiety and depression presents a challenge as there are several different types of anxiety disorder each with specific features, in addition to the emotional, cognitive, and physiological components of depression. Thus, measures that provide in-depth symptom evaluation of all aspects of anxiety and depression are extensive. There are several scales that assess anxiety and/or depression in young people, but these typically include well in excess of 20 items. Comprehensive reviews of such measures have been reported by Spence (2018) for anxiety and by Stockings et al. (2015) for depression. In some instances, short forms have been developed, such as for the Spence Children's Anxiety Scale (Ahlen et al., 2018; Reardon et al., 2018) and a 10-item version of the Centre for Epidemiological Studies – Depression (CES-D; Radloff, 1977). There is also a shorter, 25-item version of the 47-item Revised Children's Depression and Anxiety Scale (RCADS: Chorpita et al., 2000) (RCADS) developed by (Ebesutani et al., 2012). Recently, an 11-item

RCADS has been reported, although sensitivity/specificity were weak among adolescent girls (.72 and .68 respectively) and younger adolescents (.78/.66 respectively) (Radez et al., 2021). Other brief instruments include the 9-item Patient Health Questionnaire for Adolescents (PHQ-9-A: Johnson et al., 2002) for depression and the Generalized Anxiety Disorder-7 item scale (Spitzer et al., 2006) for anxiety which have the disadvantage of being downward extensions of adult measures rather than including items specifically developed for use with young people, and also focus on only anxiety or depression.

An alternative approach to identifying young people with elevated anxiety and/or depression symptoms is to identify those who experience symptoms that reflect an underlying construct that is common to both conditions rather than aiming to assess symptoms that are specific to each disorder. There is a good deal of evidence demonstrating strong covariation between anxiety and depression in both adults and children (Brodbeck et al., 2011; Cole et al., 1997; Hankin et al., 2016) and this has been proposed to reflect a shared general construct, in addition to unique variance associated with specific symptom dimensions of anxiety and depression. This shared, general component has been variously referred to as "negative affectivity", "emotional distress", "general distress", "psychological distress", or "internalizing problems" (Achenbach & Edelbrock, 1978; Iani et al., 2014). This paper will use the term "emotional distress". Several studies with both children and adults have found good support for a bifactor model in which most of the variance in anxiety and depression test items is explained by a strong general factor of emotional distress, with the specific factors of anxiety and depression explaining a relatively small proportion of the variance in item scores (Shaw et al., 2017; Zanon et al., 2021). Thus, we proposed that it would be feasible to develop a brief measure of emotional distress that is sensitive to the detection of young people with high levels of anxiety and/or depression symptoms by use of items that load strongly on the general construct of emotional distress that is common to both anxiety and depression.

Some other measures specifically designed for children have also taken the approach of assessing the general construct of emotional distress including, for example, the 13-item anxious-depressed subscale of the Child Behavior Checklist (CBCL: Achenbach, 1995; Achenbach & Edelbrock, 1978). However, this measure has been shown to not discriminate well between children with and without anxiety or depressive disorders (Burleson Daviss et al., 2006; Knepley et al., 2019; Read et al., 2015). The Strengths and Difficulties Questionnaire also includes a brief emotional symptoms scale, with five emotional symptoms items, including three relating to worry and fears, one concerning depression, and one for headaches/stomach-aches or sickness. Child-report on this subscale has been found to have weak internal consistency and poor factor loadings, (Ortuño-Sierra et al., 2015). Furthermore, youth report on the SDQ emotional symptoms subscale has shown relatively poor sensitivity for detecting clinical levels of anxiety (.50) and depression (.75) in a community sample (Goodman et al., 2009), a finding that was consistent with the results of a recent meta-analysis (Bergström & Baviskar, 2020).

Given these limitations, we aimed to develop a brief measure of the general factor of emotional distress that is psychometrically sound and provides a reliable and valid screen to identify young people with elevated symptoms of anxiety and/or depression. To achieve this aim we used a rigorous methodology involving both classical test theory and IRT.

2. Method

2.1. Participants and Procedure

Ethics approval for data collection for each sample was obtained on a site-specific basis. Ethical approval for the data collection with the clinically referred anxious sample was obtained from the human research ethics committees of the University of Queensland, University of Southern Queensland, and Griffith University. Ethical approval for data

collection from the community sample was obtained from the Macquarie University Human Research Ethics Committee and the NSW State Education Research Applications Process.

2.2.1. *Community Sample*

The 2663 community participants were aged 11-17 years (Mean = 13.41, SD = 1.74), and included 1300 (48.8%) boys, 1354 (50.8%) girls, 9 (0.3%) who identified as intersex or other sexual identity. Participants were recruited from grades 6 - 12 of 15 state, co-educational schools in Sydney, Australia. Schools were from predominantly middle-class areas. Students completed the 20 items for this study as part of a broader project on mental health screening. In addition, community participants completed the Short Mood and Feelings Questionnaire (SMFQ: Angold & Costello, 1987). Parents of all students provided written consent for their child's participation (approx. 60% return rate) and students provided assent. Students completed measures online in class groups, at school under supervision of a welfare teacher. All students who had informed consent and were present on the day of testing participated.

2.1.2. *High-Depression-Symptom Community (HiDepCom) Sample*

To examine the discriminative validity of the new scale, a sub-group of the community sample was identified as having a high level of depressive symptoms (HiDepCom) based on a score equal to or greater than the recognized clinical cut-point of 11 on the SMFQ (Angold et al., 2002). The 536 young people in the HiDepCom sample (20.1% of the total community sample) consisted of 195 (36.4%) boys and 337 (62.9%) girls and 4 (0.7%) who identified as other gender category. Mean age was 13.79 (*SD* = 1.75) years. The 2121 young people in the group that did not show elevated depressive symptoms (referred to as "not-depressed") (80.0% of the community sample) consisted of 1105 (52.0%) boys and 1017 (47.8%) girls and 5 (0.2%) who identified as other gender. Mean age was 13.32 (*SD* = 1.72) years.

2.2.2 Anxious-referred sample

The 281 anxious-referred participants were referred by health professionals (e.g. GP, pediatrician, nurse, psychologist, psychiatrist, social worker) to an open-access, free, self-help, online, 10-session, CBT treatment program for anxious children and adolescents (March et al., 2018). All enrolled in the BRAVE Self-Help program between July 1, 2014, and June 30, 2016. These participants were aged 11-17 years (Mean = 13.55, SD = 1.74), and included 112 (39.9%) boys, 166 (59.1%) girls, and 3 (1.1%) who identified as intersex or other sexual identity. In terms of residence, 152 (54.1%) resided in major Australian cities, 75 (26.5%) from Inner Regional Australia, 25 (8.9%) from Outer Regional Australia, 13 (4.6%) from remote or very remote Australia, and 16 [5.7%] provided data that could not be accurately coded).

At the time of enrolment, participants completed an online questionnaire that included the 20 items examined in the present study. Informed consent (and parental consent in the case of children aged <16 years) was obtained during online enrolment. Participants were not provided with any reimbursements for participation.

2.2. Missing data

Complete data were available for 95.8% of individuals in the community sample with total missing values of 0.32% of data points. For the anxious-referred participants, complete data were available for 98.8% of youth, with missing values found for 0.14% of data points. Missing values were handled using multiple imputation to create 10 data sets with pooled results used in subsequent analyses examining predictors of outcome.

2.3. Measures

2.3.1. The Brief Emotional Distress Scale for Youth (BEDSY)

The final BEDSY items were selected from 20 items tested in the present study, of which 6 reflected symptoms of major depressive disorder and 14 were symptoms of anxiety

(either social, separation, generalized anxiety disorder or panic disorder) consistent with DSM-5 criteria (American Psychiatric Association, 2013). These 20 items were selected after pilot tests with a larger pool of items involving two different community samples of children and adolescents to progressively identify items that, in exploratory factor analyses, loaded strongly on factors relating to anxiety or depression (≥.50) and also showed high commonality loadings on the principal component (≥.50). The pilot items included (i) 13 items from the Spence Children's Anxiety Scale (Spence, 1998) reflecting symptoms of separation, social, generalized anxiety and panic, but excluding obsessive compulsive disorder or physical injury fear items, ii) 2 items from the Children's Anxiety Scale (Spence et al., 2014) iii) 5 pilot items generated by the authors to reflect anxious reassurance seeking and anxiety regarding uncertainty, and iv) 10 pilot depression items developed by the authors to reflect symptoms described in the DSM-5. All items were rated on a 4-point scale regarding frequency of occurrence over the last 4 weeks, ranging from 0 (never) to 3 (always).

2.3.2. Short Mood and Feelings Questionnaire (SMFQ: Angold & Costello, 1987)

The SMFQ is a 13-item scale developed to assess depressive symptoms in children and adolescents aged 6 –18 years. Items reflect clinical and taxonomic knowledge of depression in youth, and include affective and cognitive aspects, plus items relating to tiredness, restlessness, and poor concentration. Respondents rate each statement regarding occurrence over the past 2 weeks as 0 = not true, 1 = sometimes true, or 2 = true, with scores ranging from 0 to 26. The SMFQ has been shown to have a good sensitivity (0.71) and specificity (0.83) at cut-point of 11 (Turner et al., 2014) which is the generally accepted cut-off for extreme scores on the SMFQ (Angold et al., 2002). Internal consistency in the present study was 0.92.

2.4. Data analytic procedure

2.4.1. Step 1: Rationalizing from 20 items using classical test theory analyses

In order to reduce the number of items, the pool of 20 items was examined first using classical test theory approaches. Given that the aim was to develop a scale of emotional distress that would identify both anxious and depressed young people, this first step selected items for further examination that:-

a. showed acceptable skewness, kurtosis and level of difficulty (see below for more explanation) and strong item-total correlation. To be retained for further examination, items were required to have skewness between -2 to +2 and kurtosis between -7 to +7 (Byrne, 2010) and an item-total correlation of .50 or greater.

b. discriminated significantly between i) anxious-referred youth vs community controls and ii) youth in the community sample with scores above vs below the clinical cut-off for depression on the SMFQ. Logistic regression analyses were conducted to examine each item's capacity to discriminate between a) the community and anxious-referred samples, and b) those in the clinical vs non-clinical ranges on SMF depression scores in the community sample. Items were only retained if they showed an odds ratio Exp(B) value > 2.0 for both comparisons, and an item-total correlation > 0.5 with the total score of all items.

All classical test theory analyses were conducted using SPSS26.

2.4.2. Step 2: Determining whether the data for items remaining after Step 1 met criteria of unidimensionality sufficient to justify use of IRT

To justify the use of IRT to examine the properties of the selected items, it was first necessary to demonstrate that there was a sufficiently strong general factor to justify use of IRT. Reise et al. (2011) noted that some degree of multidimensionality is evident in most psychological constructs and the presence of multidimensionality does not necessarily invalidate either the scoring of an instrument as a measure of a single construct, nor the fitting of a subsequent unidimensional IRT model if there is a dominant general factor running through the items. Thus, if the item data fit a bifactor model but are explained primarily by a

strong general factor in addition to significant but weak specific factors, then it is acceptable to fit a unidimensional IRT model to examine the properties of the items. The analytical approach followed that of Reise et al. (2011) using a bi-factor exploratory structural equation model (BESEM) analysis to evaluate the relative percentage of variance explained by the general factor compared to the specific factors ie. to determine the extent to which item responses on each instrument are saturated by a general factor or by meaningful multidimensionality. In a BESEM model, each item is allowed, but not forced, to load on a general factor and one or more specific factors. The specific factors account for the residual common variance that is unexplained by the general factor. BESEM was selected rather than confirmatory factor analysis (CFA) which has been criticized for not permitting cross-loadings between items, resulting in inflation of the general factor (Dierendonck et al., 2021; Gomez et al., 2020). An exploratory structural equation modeling (ESEM) approach has been proposed as a better alternative to overcome this issue (Asparouhov & Muthén, 2009).

BESEM was conducted with MPlus, with weighted least square mean and variance adjusted (WLSMV) estimation and BI-GEOMIN rotation. WLSMV adjusts for non-normality in the data and is appropriate for polytomous data with four or less response categories (Gomez et al., 2020). The general and specific factors were constrained to be orthogonal. The fit of the bifactor model was compared with 1-factor and 2-correlated factor models using ESEM, examining the comparative fit index (CFI), Tucker-Lewis Index (TLI) and Root Mean Square Error of Approximation index (RMSEA). A good model fit is reflected by CFI and TLI values >.95 and RMSEA <.06 (Hu & Bentler, 1999).

Co-efficient Omega (Zinbarg et al., 2006) was then computed to determine the relative strength of the general and specific factors using the Dueber (2017) calculator. Omega Hierarchical (*OmegaH*: ω_H) reflects the percent of variance in the total score that can be explained by the general factor. In contrast *Omega Hierarchical Subscale (OmegaHS* ω_{HS}) reflects the percent of variance in the total score that can be explained by the subscale-specific

factors that is independent from the general factor. Omega scores range from 0 to 1. When *OmegaH* is high (≥.80) and the *OmegaHS* values are low, it suggests that the scale can be considered as predominantly unidimensional and the use of IRT and the total score is justified rather than the use of subscale scores. The degree to which the scale is essentially unidimensional is also reflected by the degree of explained common variance (ECV) and percentage of uncontaminated correlations (PUC), with ECV values >.85 supporting unidimensionality (Rodriguez et al., 2016).

If a strong general factor for items was evident following Step 2 above, sufficient to justify use of IRT analysis, it was planned that IRT would be used at Step 3 to determine whether item responses were consistent with a latent factor of emotional distress as reflected by good infit and outfit, and item characteristic curves (ICCs) that demonstrate the item's ability to discriminate between respondents at different levels of the trait, with measurement precision across levels of the latent construct of interest, and invariance in response tendencies across gender and age groups.

2.4.3. Step 3: Confirming Item and Scale Characteristics using IRT

Step 3 aimed to examine the properties of the selected items using IRT and Rasch analyses, conducted using jMetrik 4.1.1. This approach enables examination of the data against a Rasch model that examines the construct validity of each item, and the unidimensionality, reliability and separation characteristics of the scale. As such, it determines whether the summation of raw scores can be justified as a reflection of an interval measurement scale of the latent construct (emotional distress). To meet criteria for inclusion in the final scale, items were required to show Rasch Joint Maximum Likelihood Estimation Difficulty scores between 0 and +/- 1.5 rather than being excessively "easy" or "difficult" indicating that respondents tend to provide very high or very low ratings on an item rather than an appropriate spread of scores. To determine how well the items fit the Rasch model, Mean-Square fit statistics for infit and outfit were examined as an average of squared

differences between the observed and expected responses from the Rasch model. Items retained for the final scale were required to show a Weighted Mean Square (WMS) Infit and Unweighted Mean Square (UMS) Outfit statistic between 0.50 and 1.50 (Linacre, 2012).

Item characteristic curves (ICC) were examined with jMetrik. As items had polytomous response options, the ICC for an item shows a separate curve for each response option. Items were discarded if they did not show at least two response option curve discriminations between Theta values of -2 to +2 (reflecting trait values of 2SDs above/below the mean). Also, to be retained, the Test Information Function (TIF) of the item had to show that the item contributed maximum measurement precision at a point between the mean (0) and +2 SDs on the latent trait, given that the test was intended to identify elevated levels of emotional distress.

Differential Item Functioning (DIF) of items was then examined between age groups (11-13 yrs vs 14-17yrs) and genders. DIF examines the differences in responses to an item between a reference and a focal group (eg. between boys and girls) for individuals with the same level of the latent trait. Such differences reflect a response bias to the item rather than differences in the latent trait. To be retained items had to be classified as AA according to the Cochran-Mantel-Haenszel chi-square statistic according to DIF criteria for polytomous items outlined by Meyer (2014). AA items show a non-significant, negligible difference between expected and observed item responses ($p<.05$), CC items are regarded as items of significant concern ($p \geq .10$), with BB items showing moderate DIF warranting further examination ($p=.05$ to $p < .10$).

The person separation index and number of strata index of the final brief scale were examined through IRT Rasch analysis to determine the extent to which the scale could i) detect statistically significant reliable differences in levels of severity of the construct among respondents (the person separation index) and ii) reliably separate the sample into measurably distinct groups/levels (the number of person strata index). A separation index of ≥ 2.0

(reflecting a person reliability co-efficient of 0.8) would indicate that at least 3 strata could be reliably identified (Eckes, 2011).

2.4.4. Step 4: Determining Scale Properties of Internal Consistency and Criterion Validity

After examining internal consistency of the 8-item scale, criterion validity was examined by comparing the total scores for the anxious-referred vs community samples, and the HiDepCom vs non-depressed community samples using SPSS26. The criterion validity of the test was examined using Receiver Operating Characteristic (ROC) curve analyses to determine the degree to which the brief scale could accurately discriminate between individuals with and without clinically elevated symptoms of anxiety or depression. Comparisons were made between the anxious-referred vs community samples, and between the HiDepCom vs non-depressed community samples using SPSS26. Given that the community sample included a proportion of participants who experienced elevated depression and anxiety, this would lower the test's capacity to discriminate between young people with and without high emotional distress in comparisons using the community sample. Thus, a third analysis was conducted comparing community youth who did not exceed the cut-off point for depression vs a combined sample of youth who were considered to have high levels of emotional distress consisting of the community HiDepCom plus the anxious-referred group.

The area under the curve (AUC) value from the ROC analysis provides a measure of the test's capacity to correctly classify those with and without a target condition, with values of .80 –.90 indicating "excellent" and .90 –1.00 indicating "outstanding" discrimination (Hosmer et al., 2013). Thus, we examined the capacity of the scale to accurately identify the proportion of individuals correctly identified as meeting criteria for anxiety or depression (ie. sensitivity) and to correctly identify those who did not meet criteria for such problems (ie. specificity). Effective screening measures need to ensure high sensitivity so that positive cases are not missed but also need to ensure high specificity to reduce the costs of identifying large

numbers of individuals who do not actually meet the criteria for the target condition. Although there is a case to suggest that priority should be given to sensitivity, given the purpose of the screening test to identify youth with high emotional distress, it is also important to ensure high specificity to avoid high rates of false positives that can result in unnecessary costs if identified youth are referred into early intervention programs. In determining the optimal cut-off points for the total score, we set a goal of $\geq$ .80 for both sensitivity and specificity, based on reviews and evaluations of measures for screening for depression and/or anxiety in children in community samples (e.g., Reardon et al., 2018; Stockings et al., 2015; Vilagut et al., 2016).

3. Results

3.1. Step 1. Characteristics of initial 20 items and item rationalization using classical test theory analyses

Supplementary Table 1 shows the mean scores, SDs, skewness and kurtosis, item-total correlations, and the logistic regression discrimination parameters (anxious-referred vs. community/HiDepCom vs. non-depressed community) for the initial 20 items. Ten items were excluded from consideration after failing to meet the criteria specified above (See Supplementary Table 1). The remaining 10 items (6 depression and 4 anxiety) met all criteria for further consideration. In order to balance the number of anxiety and depression items in the scale and to produce the shortest scale with strong psychometric properties and discriminative validity, the six depression items were examined to delete the two with weaker psychometric criteria. First, the item "I feel like I am a real loser" was deleted as it showed higher skew and kurtosis than the other items, then "I feel like crying" was dropped as it showed weaker capacity to discriminate between HiDepCom vs non-depressed community samples (See Supplementary Table 1). The remaining 8 items (4 anxiety and 4 depression) were then submitted for further examination. The final 4 depression items (I feel sad; I feel

really alone; I feel like there is nothing to look forward to; I just don't enjoy things anymore) were all investigator generated items. Of the 4 anxiety items 3 were from the SCAS (I worry that something bad will happen to me; All of a sudden I feel really scared for no reason at all; I feel afraid) and one from the CAS (I feel nervous).

3.2. Step 2: Results determining whether the 8-items met criteria of unidimensionality sufficient to justify use of IRT

3.2.1. Basic exploratory factor analysis of the 8 short-listed items showed 2 factors with eigenvalues of 5.11 and 1.01. This produced a 5:1 ratio of the first to second eigenvalue suggesting a strong general factor. The correlation between anxiety and depression factors was 0.68 (p<.001). Exploratory structural equation modeling (ESEM) was then used to compare the relative fit of the 1-factor, 2-correlated-factor, and bi-factor models using the MPlus DIFFTEST. The results indicated that the two-correlated factor model provided a significantly better fit than the 1-factor bifactor model ($\Delta \chi^2 = 693.11$, $\Delta df = 7$, $p <. .001$), and the bi-factor model provided a significantly better fit than the 2-correlated factor model ($\Delta \chi^2 = 80.56$, $\Delta df = 6$, $p <. .001$) (See Table 1). Fit indices for the bi-factor model were CFI = 0.999 and TLI =0.994 and RMSEA index = 0.042 indicating a strong fit of the data. All items loaded strongly on the general factor (.60 - .91), with weaker loadings on the specific factors.

Insert Table 1 about here

We then examined the degree to which the general factor of the bi-factor model was of sufficient magnitude to justify the use of IRT. Omega Hierarchical co-efficients were computed to determine the degree to which the specific factors of depression and anxiety in the bifactor model provide unique information over and above the general factor. For the general factor *OmegaH* was 0.920, with *OmegaHS* being 0.004 and 0.022 for the specific factors of depression and anxiety respectively. The higher the *OmegaH*, the more the general factor is the dominant source of systematic variance, with values $\geq$0.80 suggesting that total scores are essentially measuring a unidimensional construct (Reise et al., 2013). The

explained common variance index (ECVI) indicated that 81.1% of the total variance in the data was accounted for by the general factor, with 7% and 12% of unique variance in the specific factors of depression and anxiety respectively (Stucky & Edelen, 2015). Thus, despite the bi-factor structure of the scale, there is good evidence of sufficient unidimensionality to justify examination of its individual items using IRT. We therefore proceeded to examine the item characteristics of the 8 items using IRT to further confirm that they provided a valid reflection of the latent factor of general emotional distress.

3.3. Step 3: Results Confirming Item and Scale Characteristics using IRT

3.3.1. Results of Rasch IRT analysis

The results of IRT analyses for the 8 items are shown in Table 2, with all items showing acceptable difficulty levels, ranging between -0.97 to 0.47 and WMS infit and UMS outfit values within the range 0.50 to 1.50 suggesting that each item provided a good fit to the Rasch model.

The Item Characteristic Curves (ICCs) for the 8 items are shown in Supplementary Figure 1. These curves show the probability of reporting each of the 4 possible response levels (never to always) at varying levels of the latent factor of emotional distress (theta). For all items, the probability of reporting low (never) to high (always) responses increased in line with the latent factor scores, with no reversals. All items showed at least two cross-over points between response option curves, within the theta value range of +/-2 indicating a satisfactory spread of scores across the latent factor of the scale. The information curves showed that the maximum precision point for each item lay between theta values of 0 to +2 which, for a clinical scale, is the range of scores in which it is most important to detect differences.

In terms of Differential Item Functioning, Table 2 shows an absence of DIF (rated AA, $p<.05$) across age groups and genders for all items, indicating that individuals of an equivalent level of the latent trait responded in an equivalent way to each item irrespective of their age or gender, indicating no age or gender bias in item responding.

The Rasch analysis revealed a person reliability index of 0.78 and separation index of 1.90 indicating an acceptable level of reliability (Meyer, 2014). The number of strata index was 2.86 suggesting that the test could accurately identify ~3 strata of individuals on the latent construct of emotional distress.

Insert Table 2 about here

3.4. Step 4: Results Determining Scale Internal Consistency, Normative Data, and Criterion Validity

3.4.1. Internal consistency

All eight items correlated significantly with the scale total score (>.55) with Cronbach's alpha co-efficient of .88.

3.4.2. Means and SDs by Age and Gender

Table 3 shows the means and SDs for the total score for the BEDSY by age-group and gender for the anxious-referred and community (total, HiDepCom and non-depressed) samples. The anxious-referred sample showed a significantly higher BEDSY total score than the full community sample, $F(1,2930) = 526.21$, $p<.001$, with a large effect size ($d = 1.45$). In the community sample, those who exceeded the clinical cut-off for depression on the SMFQ showed significantly higher total scores on the BEDSY compared to those who did not $F(1,2652) = 2756.50$, $p<.001$, again with a large effect size ($d = 2.55$). Data for the full community sample indicated a significant interaction between age and gender, $F(2,2654) = 22.23$, $p<.001$, with older girls showing significantly higher total scores than boys or younger girls.

Insert Table 3 about here

3.4.3. Sensitivity and Specificity

The results for the Receiver Operating Characteristic (ROC) curve analysis to compare the ability of the total score of the 8-item scale to correctly identify individuals with/without the target "condition" of anxiety or depression, using the Area Under the Curve (AUC) statistics. ROC curve analyses for boys and girls combined (Table 4) indicated that:-

i) the anxious-referred vs community samples showed an AUC = 0.84.

ii) the HiDepCom vs non-depressed community samples showed AUC = 0.94.

iii) the anxious-referred plus HiDepCom community samples vs the non-depressed community sample showed AUC = .94.

Table 4 also shows cut-off scores that provided optimal levels of sensitivity and specificity for boys and girls combined and for separate genders. Detailed information showing specificity and sensitivity values at differing cut-offs are provided in Supplementary Table 2 (by gender) and Supplementary Table 3 (by age and gender). In determining the optimal cut-off scores, consideration needs to be given to sensitivity and specificity, to ensure high levels of correct identification of youth who are experiencing high levels of emotional distress (sensitivity) while also ensuring that the test does not erroneously pick up large numbers of young people who do not have problematic levels of emotional distress (specificity). Table 4 reveals that there is a trade-off in determining optimal cut-offs depending upon whether priority is given to sensitivity or specificity. On balance, we suggest the following cut-offs:- a) for girls, a cut-off of 8 provides sensitivity/specificity values of 0.83/0.73; 0.81/0.91; and 0.81/.91 in discriminating between anxious-referred vs community youth, HiDepCom vs non-depressed community participants, and between a combined sample of community HiDepCom youth plus the anxious-referred participants vs the non-depressed community sample and b) for boys, a cut-off of 6 provides sensitivity/specificity values of 0.76/0.73; 0.91/0.85 and 0.85/0.85 across the three sample comparisons respectively.

Insert Table 4 about here

4. Discussion

This paper describes the development and preliminary validation of a brief screening measure to identify children and adolescents in a community context who are experiencing emotional distress associated with anxiety and depression and are likely to benefit from early intervention. The article identified a need for, and lack of, psychometrically sound brief measures of emotional distress that are suitable for screening large numbers of young people in community contexts.

A series of steps based on classical test theory identified 8 items that met pre-determined psychometric criteria, with these items consisting of 4 items generally associated with depression and 4 items generally associated with anxiety. An exploratory factor analysis using structural equation modeling of a two-correlated factor model indicated that the items loaded strongly on their hypothesized anxiety or depression factors. However, when a bifactor model was examined, it provided a stronger fit, with the general factor explaining the vast majority (81%) of test variance, and weak specific factors of anxiety and depression symptoms once the general factor was included in the model. The findings support a general factor that is sufficiently strong to justify the use of IRT to examine the properties of the 8 items. This supports the view that the scale is essentially assessing the general construct of emotional distress in line with the goals of the study. Furthermore, the results suggest that the scale should be scored as a total score, rather than separate subscale scores.

All items of the new scale, named the Brief Emotional Distress Scale for Youth (BEDSY), met pre-set criteria for skewness and kurtosis, item-total correlations, and showed significant differences between anxious vs community youth and HiDepCom vs non-depressed participants. The item characteristics in IRT analyses also met criteria for level of difficulty, in-fit and out-fit, and showed the capacity to detect differences in levels of emotional distress. All items showed invariance in response patterns across age groups and gender.

The BEDSY was shown to have good internal consistency (0.88), and acceptable specificity and sensitivity. Optimal levels of sensitivity and specificity were obtained using gender-specific cut-offs. For girls the sensitivity/specificity values were 0.83/0.73; 0.81/0.91; and 0.81/.91 at cut-off of 8 respectively in discriminating between anxious-referred vs community youth, HiDepCom vs non-depressed community participants, and between a combined sample of community HiDepCom youth plus the anxious-referred participants vs

the non-depressed community sample. For boys, sensitivity/specificity values at a cut-off of 6 were 0.76/0.73; 0.91/0.85 and 0.85/0.85 across the three sample comparisons respectively.

Although it is difficult to directly compare sensitivity and specificity estimates across studies given differences in sample characteristics and methodologies, we note that the levels of sensitivity and specificity for the BEDSY are at least as high as those reported elsewhere in the literature for identifying young people with anxiety and/or depression using full-length measures of these constructs in community settings. For example, DeSousa et al. (2012) reported specificity/sensitivity of 0.82/0.52 for the Screen for Child Anxiety Related Emotional Disorders (SCARED) (Birmaher et al., 1997) in discriminating between children (aged 9-18 years) with and without a diagnosed anxiety disorder in a school setting. Similarly, Canals et al. (2012) found sensitivity/specificity values of 0.76/0.68 for youth report on the SCARED comparing children with and without an anxiety disorder in a community sample. Reardon et al., (2018) found sensitivity/specificity values of 0.67/0.64 comparing youth aged 7-11yrs from a community sample and clinically referred anxious youth on an 8-item version of the Spence Children's Anxiety Scale (Spence, 1997).

Similar results have been obtained for the identification of youth depression in community samples. For example, sensitivity/specificity levels of the 20-item Centre for Epidemiological Scale – Depression (Radloff, 1977) have been reported as 0.75/0.80 (Tran et al., 2019); 0.84/0.72 (Yang et al., 2018), and 0.90/0.74 (Cuijpers et al., 2008) for samples including both genders; and 0.83/.77 and .85/.49 for girls and boys respectively (Garrison et al., 1991).

In terms of strengths and limitations, the study had the advantage of including large samples for both the community and clinical populations in the development of the scale and examination of its psychometric properties. The selection of items was based on a rigorous process drawing on criteria from both classical test theory, followed by confirmation of item properties using item response theory. However, there are limitations in the study design that

need to be addressed in future research. For example, the anxious-referred sample was based on referral by a health practitioner to an online CBT program and no formal diagnosis of an anxiety disorder was conducted. Similarly, the high depression community sample was selected based on scores that exceeded the recommended clinical cut-off on the SMFQ (Thapar & McGuffin, 1998) rather than a diagnosis using a structured clinical interview. Furthermore, although the selection of the initial 20 items was based on a separate pilot study sample, in the current study the selection of the final 8 items and testing of the scale's ability to discriminate both involved the same community sample. It is possible that this could have resulted in sample-specific effects. Therefore, it will be important for future studies to test the predictive validity of the measure in independent samples.

A further issue is that we did not exclude youth with anxiety or depression problems from the community sample in the comparison with the anxious-referred group. Given that Lawrence et al. (2015) found that 8.1% of the Australian general population of 4-17 year-olds met criteria for an anxiety or depressive disorder, this is likely to have been the case in the community sample. This creates a challenge for comparing clinical vs community samples given that the community sample are not all necessarily non-clinical, resulting in an underestimate of the test's ability to discriminate between the samples. For this reason, the present study also compared the high-depression vs non-depressed community samples and compared the combined anxious-referred and community high-depression samples with the non-depressed community sample. Indeed, in these analyses the AUC was higher (0.94 rather than 0.84). Ideally, the study should have included both the screening measure and a standardized, structured diagnostic interview with all community participants to determine who did and did not meet diagnostic criteria for an anxiety and/or depressive disorder. In practice, however, such studies tend to be prohibitively expensive as they require interviewing very large samples of which the majority are not experiencing a clinical disorder. Future research should also examine concurrent validity of the scale in relation to existing longer

measures of associated constructs such as the internalizing subscale of the CBCL, and similarly to demonstrate divergent validity in terms of differentiation from different constructs such as the externalizing subscale on the CBCL.

It will also be important in future research to examine parent and teacher versions of this new scale to determine its properties with other informants and to examine whether inclusion of additional informants provides meaningful improvements in detection of youth with emotional distress. It will also be valuable to examine its test-retest reliability and sensitivity to treatment effects in order to determine its capacity to evaluate community-based interventions or the impact of early intervention initiatives. If the scale is to be of value in other contexts, such in primary health settings, it will be important to demonstrate its ability to discriminate between those with and without emotional disorders of anxiety and depression among youth referred to mental health clinics. While the results provided good support for the scale as a screener to identify young people in community contexts who are experiencing emotional distress, it would not be appropriate for use by clinicians for the diagnosis of anxiety and depressive disorders. Finally, the current study was limited to youth aged 11 to 17 years, and it would be valuable to evaluate the properties of the scale with a younger age group (e.g. 8-10 years), and with youth from a broader range of socio-economic and other cultural backgrounds. The results are consistent with previous findings of a strong common construct underpinning anxiety and depression in adolescents (eg. Brodbeck et al., 2011) and it is likely that this will also be evident among the younger age group given the findings of Cole et al. (1997) that what they called the "unified" or broadband construct was stronger for children in grade 3 than in grade 6. However, this proposition remains to be tested.

In terms of practical implications, the BEDSY would appear to offer promise as a screening tool to identify young people in community contexts who are experiencing emotional distress associated with high levels of anxiety or depression symptoms. For example, in school settings, those identified could be offered a group-based early intervention

or targeted prevention program, to enhance coping skills and reduce anxiety and depressive symptoms. Alternatively, young people with mild to moderate elevations in emotional distress could be monitored over time and re-assessed with the BEDSY after several weeks to determine whether such elevation in scores is a transient or more stable pattern, as this may influence decisions about the need for action. Those with very high scores on the BEDSY (eg. 2 standard deviations above the mean) may warrant a more in-depth assessment to identify youth who require individualized intervention with a mental health professional. The BEDSY may also be of value in screening youth for participation in online, self-help programs for anxiety and/or depression. For example, those below the cut-off could be provided with brief health literacy information, while those exceeding the cut-off may be accepted into the online intervention.

In conclusion, the Brief Emotional Distress Scale for Youth (BEDSY) was shown to have strong psychometric properties and demonstrates good potential as a short screening tool for identifying children who are experiencing symptoms of emotional distress associated with anxiety and/or depression. It is quick to administer and score and provides good identification of emotional distress in young people aged 11-17 years in a school setting, with good sensitivity and specificity at identified cut-off points. Further research is now needed to examine its ability to discriminate between youth in a school setting in which the presence and absence of anxiety and depressive disorders have been determined through clinical diagnostic interview. Similarly, it will be important to examine its properties with younger age groups, and as a quick test for monitoring treatment progress and program outcomes in community and clinical contexts.

Author Note

Declarations of interest: none

Funding: Data collection for the anxious-referred sample was supported as part of a study funded by *beyondblue.* Rapee's time was partially supported by an Australian Research Council Laureate Fellowship (FL150100096)

Authors' Contributions: Authors Spence and Rapee were both involved in the data collection. Data analysis was conducted by Spence, and both authors were involved in the preparation of the manuscript.

Appendix A. Supplementary materials:

Supplementary material related to this article can be found in the online version at doi: LINK

References

Achenbach, T. M. (1995). Empirically based assessment and taxonomy: Applications to clinical research. *Psychological Assessment, 7*(3), 261-274. https://doi.org/10.1037/1040-3590.7.3.261

Achenbach, T. M., & Edelbrock, C. S. (1978). The classification of child psychopathology: A review and analysis of empirical efforts. *Psychological Bulletin, 85*(6), 1275-1301. https://doi.org/10.1037/0033-2909.85.6.1275

Ahlen, J., Vigerland, S., & Ghaderi, A. (2018). Development of the Spence Children's Anxiety Scale - Short Version (SCAS-S). *Journal of Psychopathology and Behavioral Assessment, 40*(2), 288. https://doi.org/10.1007/s10862-017-9637-3

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. American Psychiatric Association.

Angold, A., & Costello, E. J. (1987). *Short Mood and Feelings Questionnaire* Durham, NC, Duke University.

Angold, A., Erkanli, A., Silberg, J., Eaves, L., & Costello, E. J. (2002). Depression scale scores in 8-17-year-olds: effects of age and gender. *Journal of Child Psychology and Psychiatry, 43*(8), 1052-1063. https://doi.org/10.1111/1469-7610.00232

Asparouhov, T., & Muthén, B. (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*, 397 - 438.

Bergström, M., & Baviskar, S. (2020). A systematic review of some reliability and validity issues regarding the Strengths and Difficulties Questionnaire focusing on its use in out-of-home care. *Journal of Evidence-Based Social Work*, 1-31. https://doi.org/10.1080/26408066.2020.1788477

Birmaher, B., Khetarpal, S., Brent, D., Cully, M., Balach, L., Kaufman, J., & Neer, S. M. (1997). The Screen for Child Anxiety Related Emotional Disorders (SCARED): Scale Construction and Psychometric Characteristics. *Journal of the American Academy of Child and Adolescent Psychiatry, 36*(4), 545-553. https://doi.org/10.1097/00004583-199704000-00018

Brodbeck, J., Abbott, R. A., Goodyer, I. M., & Croudace, T. J. (2011). General and specific components of depression and anxiety in an adolescent population [journal article]. *BMC Psychiatry, 11*(1), 191. https://doi.org/10.1186/1471-244x-11-191

Burleson Daviss, W., Birmaher, B., Melhem, N. A., Axelson, D. A., Michaels, S. M., & Brent, D. A. (2006). Criterion validity of the Mood and Feelings Questionnaire for depressive episodes in clinic and non-clinic subjects. *Journal of Child Psychology and Psychiatry, 47*(9), 927-934. https://doi.org/10.1111/j.1469-7610.2006.01646.x

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Routledge.

Canals, J., Hernández-Martínez, C., Cosi, S., & Domènech, E. (2012). Examination of a cutoff score for the Screen for Child Anxiety Related Emotional Disorders (SCARED) in a non-clinical Spanish population. *Journal of Anxiety Disorders, 26*(8), 785-791. https://doi.org/10.1016/j.janxdis.2012.07.008

Chorpita, B. F., Yim, L., Moffitt, C., Umemoto, L. A., & Francis, S. E. (2000). Assessment of symptoms of DSM-IV anxiety and depression in children: A revised child anxiety and depression scale. *Behaviour Research and Therapy, 38*(8), 835-855.

Cole, D. A., Truglio, R., & Peeke, L. (1997). Relation Between Symptoms of Anxiety and Depression in Children: A Multitrait-Multimethod-Multigroup Assessment. *Journal of Consulting and Clinical Psychology, 65*(1), 110-119. https://doi.org/10.1037/0022-006X.65.1.110

Cuijpers, P., Boluijt, P. R., & van Straten, A. (2008). Screening of depression in adolescents through the Internet: Sensitivity and specificity of two screening questionnaires.

European Child and Adolescent Psychiatry, *17*(1), 32-38. https://doi.org/10.1007/s00787-007-0631-2

DeSousa, D. A., Salum, G. A., Isolan, L. R., & Manfro, G. G. (2012). Sensitivity and Specificity of the Screen for Child Anxiety Related Emotional Disorders (SCARED): A Community-Based Study. *Child Psychiatry and Human Development*, *44*(3), 391-399. https://doi.org/10.1007/s10578-012-0333-y

Dierendonck, C., de Chambrier, A.-F., Fagnant, A., Luxembourger, C., Tinnes-Vigne, M., & Poncelet, D. (2021). Investigating the Dimensionality of Early Numeracy Using the Bifactor Exploratory Structural Equation Modeling Framework. *Frontiers in Psychology*, *12*, 680124-680124. https://doi.org/10.3389/fpsyg.2021.680124

Dueber, D. M. (2017). *Bifactor Indices Calculator: A Microsoft Excel-based tool to calculate various indices relevant to bifactor CFA models*, https://dx.doi.org/10.13023/edp.tool.01.

Ebesutani, C., Reise, S. P., Chorpita, B. F., Ale, C., Regan, J., Young, J., Charmaine, H.-M., & Weisz, J. R. (2012). The Revised Child Anxiety and Depression Scale-Short Version: Scale reduction via exploratory bifactor modeling of the broad anxiety factor. *Psychological Assessment*, *24*(4), 833-845. https://doi.org/10.1037/a0027283

Eckes, T. (2011). Item banking for C-tests: A polytomous Rasch modeling approach. *Psychology science*, *53*(4), 414.

Garrison, C. Z., Addy, C. L., Jackson, K. L., McKeown, R. E., & et al. (1991). The CES-D as a screen for depression and other psychiatric disorders in adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry*, *30*(4), 636-641.

Gomez, R., Stavropoulos, V., & Griffiths, M. D. (2020). Confirmatory factor analysis and exploratory structural equation modelling of the factor structure of the Depression Anxiety and Stress Scales-21. *PloS One*, *15*(6), e0233998. https://doi.org/10.1371/journal.pone.0233998

Goodman, R., Ford, T., Simmons, H., Gatward, R., & Meltzer, H. (2009). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *International Review of Psychiatry*, *15*(1-2), 166-172. https://doi.org/10.1080/09540260210000046128

Hankin, B. L., Snyder, H. R., Gulley, L. D., Schweizer, T. H., Bijttebier, P., Nelis, S., Toh, G., & Vasey, M. W. (2016). Understanding comorbidity among internalizing problems: Integrating latent structural models of psychopathology and risk mechanisms. *Development and Psychopathology*, *28*(4pt1), 987-1012. https://doi.org/10.1017/S0954579416000663

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression, third edition* (3rd ed.). John Wiley and Sons. https://doi.org/10.1002/9781118548387

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Iani, L., Lauriola, M., & Costantini, M. (2014). A confirmatory bifactor analysis of the hospital anxiety and depression scale in an Italian community sample. *Health and Quality of Life Outcomes*, *12*(1), 84-84. https://doi.org/10.1186/1477-7525-12-84

Johnson, J. G., Harris, E. S., Spitzer, R. L., & Williams, J. B. W. (2002). The patient health questionnaire for adolescents: Validation of an instrument for the assessment of mental disorders among adolescent primary care patients. *Journal of Adolescent Health*, *30*(3), 196-204. https://doi.org/10.1016/S1054-139X(01)00333-0

Knepley, M. J., Kendall, P. C., & Carper, M. M. (2019). An analysis of the child behavior checklist anxiety problems scale's predictive capabilities. *Journal of Psychopathology*

and Behavioral Assessment, *41*(2), 249-256. https://doi.org/10.1007/s10862-019-09722-5

Lawrence, D., Johnson, S., Hafekost , J., Boterhoven De Haan, K., Sawyer, M., Ainley, J., & Zubrick , S. R. (2015). *The mental health of children and adolescents: Report on the second Australian child and adolescent survey of mental health and wellbeing*. Department of Health.

Linacre, J. M. (2012). Winsteps® Rasch measurement computer program user's guide. . In. Beaverton, OR: Winsteps.

March, S., Spence, S. H., Donovan, C. L., & Kenardy, J. A. (2018). Large-scale dissemination of internet-based cognitive behavioral therapy for youth anxiety: Feasibility and acceptability study. *Journal of Medical Internet Research Vol 20(7), 2018, ArtID e234*, *20*(7). https://doi.org/doi: 10.2196/jmir.9211

Meyer, J. P. (2014). *Applied measurement with jMetrik*. Routledge. https://doi.org/10.4324/9780203115190

Ortuño-Sierra, J., Chocarro, E., Fonseca-Pedrero, E., Riba, S. S. i., & Muñiz, J. (2015). The assessment of emotional and behavioural problems: Internal structure of The Strengths and Difficulties Questionnaire. *International Journal of Clinical and Health Psychology*, *15*(3), 265-273. https://doi.org/https://doi.org/10.1016/j.ijchp.2015.05.005

Radez, J., Waite, P., Chorpita, B., Creswell, C., Orchard, F., Percy, R., Spence, S. H., & Reardon, T. (2021). Using the 11-item Version of the RCADS to Identify Anxiety and Depressive Disorders in Adolescents. *Research on child and adolescent psychopathology*. https://doi.org/10.1007/s10802-021-00817-w

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385-401.

Read, K. L., Settipani, C. A., Peterman, J., Kendall, P. C., Compton, S., Piacentini, J., McCracken, J., Bergman, L., Walkup, J., Sakolsky, D., Birmaher, B., Albano, A. M., Rynn, M., Ginsburg, G., Keeton, C., Gosch, E., Suveg, C., Sherrill, J., & March, J. (2015). Predicting anxiety diagnoses and severity with the CBCL-A: Improvement relative to other CBCL scales? *Journal of Psychopathology and Behavioral Assessment*, *37*(1), 100-111. https://doi.org/10.1007/s10862-014-9439-9

Reardon, T., Spence, S. H., Hesse, J., Shakir, A., & Creswell, C. (2018). Identifying children with anxiety disorders using brief versions of the Spence Children's Anxiety Scale for children, parents, and teachers. *Psychological Assessment*. https://doi.org/doi: 10.1037/pas0000570

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*(2), 129-140. .

Reise, S. P., Ventura, J., Keefe, R. S. E., Baade, L. E., Gold, J. M., Green, M. F., Kern, R. S., Mesholam-Gately, R., Nuechterlein, K. H., Seidman, L. J., & Bilder, R. (2011). Bifactor and Item Response Theory Analyses of Interviewer Report Scales of Cognitive Impairment in Schizophrenia. *Psychological Assessment*, *23*(1), 245-261. https://doi.org/10.1037/a0021501

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3), 1-15. https://doi.org/10.1080/00223891.2015.1089249

Shaw, T., Campbell, M. A., Runions, K. C., & Zubrick, S. R. (2017). Properties of the DASS- 21 in an Australian community adolescent population. *Journal of Clinical Psychology*, *73*(7), 879-892. https://doi.org/10.1002/jclp.22376

Spence, S. H. (1997). Structure of anxiety symptoms among children: A confirmatory factor-analytic study. *Journal of Abnormal Psychology*, *106*(2), 280-297.

Spence, S. H. (1998). A measure of anxiety symptoms among children. *Behaviour Research and Therapy*, *36*(5), 545-566.

Spence, S. H. (2018). Review- Measurement Issues: Assessing anxiety disorders in children and adolescents. *Child and Adolescent Mental Health*, *23*(3), 266-282. https://doi.org/10.1111/camh.12251

Spence, S. H., Sawyer, M. G., Sheffield, J., Patton, G., Bond, L., Graetz, B., & Kay, D. (2014). Does the absence of a supportive family environment influence the outcome of a universal intervention for the prevention of depression? *International Journal of Environmental Research and Public Health*, *11*(5), 5113-5132.

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092-1097. https://doi.org/10.1001/archinte.166.10.1092

Stockings, E., Degenhardt, L., Lee, Y. Y., Mihalopoulos, C., Liu, A., Hobbs, M., & Patton, G. (2015). Symptom screening scales for detecting major depressive disorder in children and adolescents: A systematic review and meta-analysis of reliability, validity and diagnostic utility. *Journal of Affective Disorders*, *174*, 447-463. https://doi.org/10.1016/j.jad.2014.11.061

Stucky, B. D., & Edelen, M. O. (2015). Using hierarchical IRT models to create unidimensional measures from multidimensional data. . In S. P. Reise. & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment.* (pp. 183–206). Routledge/Taylor & Francis Group.

Thapar, A., & McGuffin, P. (1998). Validity of the shortened Mood and Feelings Questionnaire in a community sample of children and adolescents: a preliminary research note. *Psychiatry Research*, *81*(2), 259-268. https://doi.org/10.1016/s0165-1781(98)00073-0

Tran, T. D., Kaligis, F., Wiguna, T., Willenberg, L., Nguyen, H. T. M., Luchters, S., Azzopardi, P., & Fisher, J. (2019). Screening for depressive and anxiety disorders among adolescents in Indonesia: Formal validation of the Centre for Epidemiologic Studies Depression Scale-Revised and the Kessler Psychological Distress Scale [Clinical Psychological Testing 2224]. *Journal of Affective Disorders*, *246*, 189-194. https://doi.org/http://dx.doi.org/10.1016/j.jad.2018.12.042

Turner, N., Joinson, C., Peters, T. J., Wiles, N., & Lewis, G. (2014). Validity of the Short Mood and Feelings Questionnaire in late adolescence. *Psychological Assessment*, *26*(3), 752-762. https://doi.org/10.1037/a0036572

Vilagut, G., Forero, C. G., Barbaglia, G., & Alonso, J. (2016). Screening for depression in the general population with the Center for Epidemiologic Studies Depression (CES-D): A systematic review with meta-analysis. *PloS One*, *11*(5), 10.1371/journal.pone.0155431. https://doi.org/10.1371/journal.pone.0155431

Yang, W., Xiong, G., Garrido, L. E., Zhang, J. X., Wang, M.-C., & Wang, C. (2018). Factor structure and criterion validity across the full scale and ten short forms of the CES-D among Chinese adolescents [Clinical Psychological Testing 2224]. *Psychological Assessment*, *30*(9), 1186-1198. https://doi.org/http://dx.doi.org/10.1037/pas0000559 (Psychological Assessment: A Journal of Consulting and Clinical Psychology)

Zanon, C., Brenner, R. E., Baptista, M. N., Vogel, D. L., Rubin, M., Al-Darmaki, F. R., Gonçalves, M., Heath, P. J., Liao, H.-Y., Mackenzie, C. S., Topkaya, N., Wade, N. G., & Zlati, A. (2021). Examining the dimensionality, reliability, and invariance of the Depression, Anxiety, and Stress Scale–21 (DASS-21) across eight countries. *Assessment*, *28*(6), 1531-1544. https://doi.org/10.1177/1073191119887449

Zinbarg, R., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ωh. *Applied Psychological Measurement*, *30*(2), 121-144.

Table 1

Factor loadings of items for general and specific factors in the1-factor, 2-correlated-factor, and bi-factor models (Exploratory factor analyses using structural equation modelling)

	1-factor Model	2-correlated-factor Model		Bi-factor Model		
		Depression	Anxiety	General factor	Depression	Anxiety
I feel really sad	0.89	0.75	0.20	0.91	0.02	-0.13
I feel really alone	0.87	0.82	0.09	0.88	0.07	-0.19
I feel like there is nothing to look forward to	0.83	0.87	0.00	0.80	0.35	-0.13
I just don't enjoy things anymore	0.85	0.91	-0.03	0.82	0.45	-0.16
I worry that something bad will happen to me	0.65	0.06	0.67	0.61	-0.09	0.37
I feel nervous	0.70	0.00	0.79	0.66	-0.16	0.42
All of a sudden I feel really scared for no reason at all	0.79	0.20	0.68	0.76	-0.09	0.32
I feel afraid	0.67	-0.04	0.79	0.63	-0.17	0.43
Model Fit						
Chi-Square (df)	1149.88 (20)		183.44 (13)		40.00 (7)	
RMSEA	0.14		0.069		0.042	
CFI	0.951		0.993		0.999	
TLI	0.932		0.984		0.994	

Note: the χ^2 provided cannot be used in the regular way for difference testing, and the DIFFTEST function is used with WLSMV estimation for nested models (Muthén & Muthén, 2017). These results are reported in the text of the paper.

Table 2
Item analysis of the final 8 items with the community sample (unless otherwise specified)

	Results from Classical Test Theory							Results from IRT and Rasch Model						
ITEMS	Mean	SD	Skew-ness	Kurt-osis	Item-Total Correlation	Exp (B) Anxious vs Community Samples	Exp (B) Depressed vs Non-depressed Community Samples	Difficulty	Infit WMS	Outfit UMS	DIF Rating for Gender	DIF Rating for Age Group	ICC Crossovers between Theta -2 to +2	TIF Criterion
I feel really sad	0.60	0.80	1.28	1.01	0.75	2.71	12.43	-0.10	0.75	0.75	AA	AA	3	YES
I feel really alone	0.46	0.77	1.72	2.26	0.70	2.29	9.65	0.22	0.88	0.82	AA	AA	3	YES
I feel like there is nothing to look forward to	0.44	0.75	1.76	2.45	0.67	2.56	6.94	0.32	0.97	0.91	AA	AA	3	YES
I just don't enjoy things anymore	0.41	0.72	1.80	2.71	0.68	2.53	10.02	0.47	0.89	0.87	AA	AA	3	YES
I worry that something bad will happen to me	0.76	0.77	0.87	0.53	0.55	3.11	3.28	-0.43	1.22	1.18	AA	AA	3	YES
I feel nervous	0.97	0.74	0.60	0.41	0.58	.64	3.59	-0.97	1.11	1.09	AA	AA	2	YES
All of a sudden I feel really scared for no reason at all	0.46	0.74	1.62	2.04	0.66	.64	2.68	0.34	0.99	0.85	AA	AA	3	YES
I feel afraid	0.72	0.65	0.61	0.49	0.55	.58	3.65	0.15	1.12	1.11	AA	AA	3	YES

Note: **Criteria for inclusion**
Skewness: -2 to +2
Kurtosis: -7 to +7
Item-Total Correlation: > . 55

IRT Criteria for Acceptability: Item-Total Correlation: > . 50; Rasch difficulty score -1.5 to +1.5; Infit WMS (Weighted Mean Squares) 0.5 to 1.5; Outfit UMS (Unweighted Mean Squares) 0.5 to 1.5

Item Characteristic Curve (ICC): $\geq$ 2 cross-overs in curves between Theta -2 to +2
DIF = Differential Item Functioning: an AA rating indicates no evidence of differential item responding between groups of individuals (eg. gender) assuming an equivalent level of the latent trait ($p<.05$).

Test Information Function (TIF): Maximum precision occurs between Theta 0 to +2

Table 3

Means and Standard Deviations for the 8 item BEDSY Total Score by Age and Gender for Anxious-Referred and Community (Depressed and Non-depressed) Samples

		Community Sample			Anxious-Referred			Community Sample Below SMFQ depression cut-off			Community Sample Above SMFQ depression cut-off		
Gender	Age Group	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N
Male	11 to 13	4.23	3.81	724	8.28	3.75	72	3.18	2.52	622	10.62	4.13	102
	14 to 17	3.85	3.92	576	10.98	5.47	40	2.65	2.32	483	10.08	4.61	93
	Total	4.06	3.87	1300	9.25	4.60	112	2.95	2.45	1105	10.36	4.36	195
Female	11 to 13	5.00	4.42	759	10.93	4.74	70	3.43	2.61	611	11.51	4.44	148
	14 to 17	6.21	4.99	595	13.89	5.10	96	3.77	2.83	406	11.43	4.62	189
	Total	5.53	4.72	1354	12.64	5.15	166	3.56	2.70	1017	11.47	4.54	337
Total	11 to 13	4.63	4.15	1483	9.59	4.45	142	3.30	2.56	1233	11.15	4.33	250
	14 to 17	5.05	4.65	1171	13.03	5.36	136	3.16	2.62	889	10.99	4.65	282
	Total	4.81	4.38	2654	11.27	5.20	278	3.24	2.59	2122	11.06	4.50	532

Note: BEDSY = Brief Emotional Distress Scale for Youth
SMFQ = Short Mood and Feelings Questionnaire (Clinical cut-off is total score $\geq$ 11).

Table 4
Summary of Sensitivity/Specificity at Cut-off Options for BEDSY Total Scores Between Samples by Gender, plus Area Under the Curve (AUC) values for ROC analyses

		Anxious Referred vs Community		Depressed vs Non-depressed Community		Depressed community and anxious-referred versus non-depressed community	
Boys and Girls							
	Cut-off Options	6	7	6	7	6	7
	Sensitivity/ Specificity	0.85/ 0.67	0.79/ 0.73	0.92/ 0.82	0.85/ 0.88	0.90/ 0.82	0.83/ 0.88
	AUC	0.84		0.94		0.94	
Girls							
	Cut-off Options	7	8	7	8	7	8
	Sensitivity/ Specificity	0.86/ 0.68	0.83/ 0.73	0.86/ 0.86	0.81/ 0.91	0.86/ 0.86	0.81/ 0.91
	AUC	0.85		0.94		0.94	
Boys							
	Cut-off Options	6	7	6	7	6	7
	Sensitivity/ Specificity	0.76/ 0.73	0.67/ 0.80	0.91/ 0.85	0.84/ 0.91	0.85/ 0.85	0.78/ 0.91
	AUC	0.83		0.94		0.93	

Notes: Sensitivity (correct identification of "caseness") and specificity (correct identification as not a "case") computed from cell counts in cross-tabulations for specific cut-offs.
A cut-off score refers to a score greater than or equal to the value.