Measuring Depression in Adults with Autism Spectrum Disorder: Psychometric Validation of the Beck Depression Inventory–II

Zachary J. Williams

Medical Scientist Training Program, Vanderbilt University School of Medicine, Nashville, TN, USA

Vanderbilt Brain Institute, Vanderbilt University Medical Center, Nashville, TN, USA

Frist Center for Autism and Innovation, Vanderbilt University, Nashville, TN, USA

Zachary.j.williams@vanderbilt.edu

(805)-729-6691

Jonas Everaert

Department of Experimental Clinical and Health Psychology, Ghent University, Ghent, Belgium

Katherine O. Gotham

Department of Psychology, Rowan University, Glassboro, NJ, USA

Abstract

Depressive disorders are common in adults with autism spectrum disorder (ASD), but few studies have examined the extent to which common depression questionnaires are psychometrically appropriate for use in this population. Using item response theory, this study examined the psychometric properties of the Beck Depression Inventory-II (BDI-II) in a sample of 947 adults with ASD. BDI-II latent trait scores exhibited strong reliability, construct validity, and moderate ability to discriminate between depressed and non-depressed adults with ASD (*AUC* = 0.796 [0.763, 0.826], sensitivity = 0.820 [0.785, 0.852], specificity = 0.653 [0.601, 0.699]). These results collectively indicate that the BDI-II is a valid measure of depressive symptoms in adults with ASD, appropriate for quantifying depression severity in research studies or screening for depressive disorders in clinical settings. A free online score calculator has been created to facilitate the use of BDI-II latent trait scores for clinical and research applications (available at https://asdmeasures.shinyapps.io/bdi_score/).

Key Words: Autism Spectrum Disorder, Depression, Psychometric, Beck Depression Inventory–II, Item Response Theory

Introduction

Autism spectrum disorder (ASD) is a lifelong neurodevelopmental condition characterized by persistent social communication impairment as well as the presence of restricted and repetitive patterns of behavior and interests (American Psychiatric Association, 2013). Although ASD is often thought of as a childhood disorder, the challenges faced by individuals on the autism spectrum continue and are often magnified in adulthood (Howlin & Magiati, 2017; Kraper et al., 2017). Notably, psychiatric comorbidities are quite common in this population, with the majority of adults with ASD meeting criteria for one or more additional psychiatric diagnoses (Bishop-Fitzpatrick & Rubenstein, 2019; Croen et al., 2015; Davignon et al., 2018; Griffiths et al., 2019; Hofvander et al., 2009; Hollocks et al., 2019; Howlin & Magiati, 2017; Lever & Geurts, 2016; Nylander et al., 2018; Supekar et al., 2017; Vohra et al., 2017). Among comorbid conditions in adults with ASD, major depressive disorder is exceedingly common, with an estimated 23% current prevalence and 37% lifetime prevalence in this population (Hollocks et al., 2019). The functional impact of depression in autistic adults is substantial, with comorbid depressive symptoms predicting diminished quality of life, as well as increased rates of behavioral problems, self-injurious behaviors, and suicidality (Cassidy, Bradley, Shaw, et al., 2018; M.-H. Chen et al., 2017; Hedley et al., 2018; Hirvikoski et al., 2019; Licence et al., 2019; Mason et al., 2019; McConachie et al., 2018; Moseley et al., 2019; Pezzimenti et al., 2019).

Despite the large burden of depression in this population, few studies have attempted to establish the psychometric properties of common depression symptom measures in adults with ASD (for a review, see Cassidy, Bradley, Bowen, et al., 2018). Studies measuring depression in ASD have previously used a number of measures validated in the general population, including the Beck Depression Inventory-Second Edition (BDI-II; Moss et al., 2015), Depression Anxiety

Stress Scales (Maddox & White, 2015; Nah et al., 2018), Hamilton Depression Rating Scale (Buchsbaum et al., 2001), Hospital Anxiety and Depression Scale (Powell & Acker, 2014), Montgomery–Åsberg Depression Rating Scale (Wentz et al., 2012), and Patient Health Questionnaire-9 (PHQ-9; Hedley et al., 2018) without assessing the validity of those measures in ASD. In recent years, several studies have attempted to fill this gap, reporting psychometric properties of the BDI-II (Gotham et al., 2015), Hospital Anxiety and Depression Scale (Uljarević et al., 2018), and PHQ-9 (Arnold et al., 2019). Two of the aforementioned studies have compared the latent structures of depression questionnaires in ASD to those reported in non-ASD samples, finding similar structures across both groups (Arnold et al., 2019; Uljarević et al., 2018). Arnold and colleagues (2019) also reported the results of a bifactor model of the PHQ-9, which indicated the presence of a strong general factor and supported the use of PHQ-9 total scores as a measure of overall depressive symptomatology.

One major issue that has yet to be addressed in this literature is the comparability of item responses between adults with ASD and typically developing (TD) controls. Several authors have raised concerns that adults with ASD may answer questions about depressive symptoms in different ways than questionnaires originally intended. Adults with ASD may have systematic biases in item responses due to the overlapping clinical presentations of ASD and mood disorders (e.g., social withdrawal, noticeably slow motor response, difficulty concentrating), cognitive differences such as literal interpretation of items (e.g., "I wouldn't say I've *lost interest in daily activities* because I never felt particular *interest* in brushing my teeth"), or alexithymia that may limit individuals' insight into their own emotional experiences (Cassidy, Bradley, Bowen, et al., 2018; Gotham et al., 2015; Pezzimenti et al., 2019; Uljarević et al., 2018). However, no study to date has specifically tested whether ASD and TD adults exhibit differential

item functioning (DIF) on depression scales, and thus these claims remain purely speculative at this time. Formal tests of DIF between diagnostic groups are necessary to determine the presence and practical significance of differential item responses between diagnostic groups, which if severe enough may warrant the adoption of novel scales to assess depressive symptoms in adults with ASD.

In the current study, we sought to evaluate the psychometric properties of the Beck Depression Inventory-Second Edition (BDI-II; Beck et al., 1996) in adults with ASD, providing a comprehensive understanding of the measure's reliability, validity, and appropriateness for use in this population. The BDI-II has been utilized extensively over the last two decades, with many studies demonstrating sound psychometric properties and strong diagnostic performance in psychiatric, medical, and general population samples (for a review, see Wang & Gorenstein, 2013). This measure is also one of the most frequently used in the adult ASD population (Burns et al., 2019; Cederlund et al., 2010; Crane et al., 2013; Gotham et al., 2014, 2018; Han et al., 2019; Hill et al., 2004; Hillier et al., 2011; Limoges et al., 2005; Russell et al., 2017; Underwood et al., 2019; Unruh et al., 2018). Items on the BDI-II align well with the major depressive disorder criteria in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), and a recent study found that the BDI-II has the largest amount of symptom overlap with six other common depression questionnaires (Fried, 2017). Furthermore, the BDI-II was the only depression measure to have its psychometric properties examined in an ASD population at the time of a recent literature review (Cassidy, Bradley, Bowen, et al., 2018). However, as noted in this review, favorable evidence for use of the BDI-II in ASD came from a relatively small (N = 50) study by Gotham and colleagues (2015) that provided only weak evidence of the instrument's criterion validity. The current study further investigated the psychometric properties

of the BDI-II within the ASD population, examining its latent structure, reliability, nomological validity, and diagnostic test characteristics within an item response theory (IRT) framework (Embretson, 1996; Petrillo et al., 2015; Thomas, 2019).

IRT represents an alternative psychometric approach from the classical test theory (CTT) approach used to create the majority of scales for use in ASD research today (Petrillo et al., 2015). Although a full comparison of the two methodologies is beyond the scope of this paper, IRT models have several potential advantages over CTT methods in assessing self-reported health outcomes such as depressive symptoms (see Embretson, 1996; Hays et al., 2000; Reise & Henson, 2003 for reviews). Chief among these is the ability to calculate an estimated "latent trait score" from each unique combination of item responses (including missing data), replacing the unit-weighted raw score typically used in CTT applications. The use of this score allows each item to be weighted optimally according to model parameters, causing individuals with the same CTT total scores to be further discriminated based on the specific item scores comprising that composite. IRT-based latent trait scores are also associated with different standard error estimates at each point along the latent trait continuum, allowing score reliability to be estimated for each individual separately and providing more accurate score confidence intervals. Despite these and other advantages, IRT-based measurement tools are largely unused in autism research and clinical practice (though see Farmer et al., In Press). One major obstacle preventing the widespread use IRT-based scoring in these settings is that many clinicians and researchers lack the specific knowledge and expertise needed to calculate IRT-based latent trait scores from published item parameters. Thus, a secondary aim of the current study was to provide a free online scoring tool that allows non-experts to easily calculate latent trait scores from the BDI-II using the item parameters derived from our IRT model.

The IRT approach also provides an elegant framework in which to assess DIF, testing whether item slope and/or intercept parameters differ between specific subgroups of interest (Thomas, 2019). Using this framework, the we aimed to determine whether the items of the BDI-II function differentially between adults with ASD and TD controls, empirically testing the claim that individuals with ASD answer depression questionnaires in a qualitatively different manner from the general population (Cassidy, Bradley, Bowen, et al., 2018; Gotham et al., 2015; Pezzimenti et al., 2019; Uljarević et al., 2018). DIF of the BDI-II was also tested within the ASD sample in order to determine whether items function differentially in groups defined based on sociodemographic factors or common comorbid conditions. While the DIF null hypothesis of complete equivalence between groups is certainly false (Cohen, 1994), it remains to be determined whether there exist *practically significant* differences in item and test functioning between adults in these various groups. Thus, while we expected to detect some degree of DIF in our analyses, we hypothesized that these differences would not be practically significant at the level of test scores and would thus be small enough to be ignored in practice.

As a final goal of our study, we also sought to establish the nomological validity of BDI-II scores in adults with ASD by assessing the relationships of these scores with measures of anxiety, quality of life, ASD symptom severity, and demographic variables. As measures of depression and anxiety are highly correlated in the general population (Clark et al., 1994) and adults with ASD (R. Y. Cai et al., 2018; Griffiths et al., 2019; Nah et al., 2018), we examined the relationship between BDI-II and a measure of anxiety (GAD-7), expecting to find a correlations similar to previous studies in ASD (i.e., $r > 0.6$). Similarly, depressive symptoms are a strong predictor of lower quality of life in ASD (Arnold et al., 2019; McConachie et al., 2018), and thus a measure of global quality of life (WHOQOL-BREF) was used to assess the criterion validity of

BDI-II scores for this metric. In line with previous studies in the ASD population (Arnold et al., 2019; McConachie et al., 2018), we hypothesized that BDI-II scores have strong negative correlations ($r < -0.60$) with global quality of life. As depression and ASD have a number of overlapping features such as constricted affect and social withdrawal (Pezzimenti et al., 2019), we also examined the relationship of the BDI-II and a measure of ASD symptomatology (SRS-2) to establish divergent validity. While depressive symptoms are known to correlate moderately with self-reported ASD symptoms (Uljarević et al., 2019), we expected this relationship to be significantly smaller than the correlation between BDI-II scores and anxiety symptoms ($\Delta r > 0.2$). Lastly, we examined relationships between BDI-II scores and several demographic factors, including age, sex (male vs. female), race/ethnicity (non-Hispanic White vs. others), and level of education (at least some college vs. no college). Relationships with age, race/ethnicity, and education level were expected to be negligible (or $r < 0.1$ or $d < 0.2$), further establishing the discriminant validity of the BDI-II in this population. However, given the female predominance of depression in both the general population (Nolen-Hoeksema, 1987) and ASD (Lai et al., 2019), we expected BDI-II scores to be slightly higher in females with ASD compared to males ($d > 0.2$). By establishing the nomological network of the BDI-II in ASD (Cronbach & Meehl, 1955), we sought to establish the validity of this questionnaire as a bona fide measure of depressive symptoms suitable for general use in ASD research.

Methods

The current investigation was a secondary data analysis of BDI-II responses collected as a part of several laboratory and online studies (See “Participants” section for more details on each study). Participants with ASD were drawn primarily from the Simons Foundation Powering

Autism Research for Knowledge (SPARK) cohort, a U.S.-based online community that allows people with ASD and their families to participate in ASD research studies (Feliciano et al., 2018). These data were combined with a well-characterized community sample of adults with and without ASD who completed paper-and-pencil BDI-II forms as part of laboratory studies conducted at Vanderbilt University Medical Center (Gotham et al., 2018; Han et al., 2019; Unruh et al., 2018). To construct a sample of TD adults large enough for adequate DIF testing, BDI-II data from a general population comparison group were drawn from four online studies of cognitive biases and depressive symptoms that recruited participants using Amazon's Mechanical Turk (MTurk; Everaert et al., 2018, 2020; Everaert & Joormann, 2019). As participants from MTurk tend to report higher rates of depression than the general population (Ophir et al., 2020), we felt that these individuals would provide an adequate comparison group spanning the entire range of depressive symptoms. In addition to baseline levels of depression in this population, one of the MTurk samples used in the current study was enriched for participants with high levels of depressive symptoms on the PHQ-9 (Everaert et al., 2018). As these studies were not originally conducted with ASD in mind, participants were not screened for ASD diagnoses themselves. However, as the prevalence of self-reported ASD in unselected MTurk samples is approximately 1–2% (e.g., Mitchell & Locke, 2015; Skylark & Baron-Cohen, 2017), the number of "TD" adults with unrecognized ASD in our sample is unlikely to be large enough to mask the presence of DIF between diagnostic groups. Thus, the inclusion of this MTurk data provided us with an aggregate sample of approximately 1000 adults with ASD and 1000 TD controls, an ideal size for recovering IRT model parameters in both groups and testing the central hypothesis of DIF across diagnostic groups (Jiang et al., 2016).

Participants

SPARK (ASD) Sample. Adults diagnosed with ASD were invited to take part in our study via the SPARK research portal. To be eligible for the current study, participants had to be an adult with a professional diagnosis of ASD between the ages of 18 years and 45 years, 11 months at the time of survey distribution. Adults with ASD enrolled in the SPARK cohort must self-report a professional diagnosis of ASD, and although these diagnoses are not independently validated, the majority of SPARK participants are recruited from university autism clinics and thus have a very high likelihood of valid ASD diagnosis (Feliciano et al., 2018). Additionally, a study conducted in a previous version of this participant pool found that 98% of registry participants were able to produce documentation verifying a professional ASD diagnosis (Daniels et al., 2012). Participants completed the BDI-II, also providing information on demographics, comorbid psychiatric conditions, autism severity, quality of life, and a number of other clinical variables. Lifetime diagnoses of any depressive disorder was assessed with the following question: *Have you ever been diagnosed with Depression (such as major depressive disorder, seasonal affective disorder, postpartum depression, or some other kind of depression)?* Participants were able to respond (a) *Yes*, (b) *No*, or (c) *Diagnosis suspected by self or others, but never confirmed*. Those who answered "Yes" or "Suspected" were asked to answer the following question on current depressive symptoms: *Do you currently have Depression (symptoms present in the past 3 months, or receiving ongoing treatment)?* Individuals who answered "Yes" to this second question were classified as endorsing current depression, while those who answered "No" or were not presented the question were classified as not endorsing current depression.

All data used for the study were provided by self-report and were collected during Winter and Spring of 2019 as part of a wider study on repetitive thinking and its links to psychopathology in ASD. Participants received a total of $50 in Amazon gift cards for

completion of the study. A total of 1012 individuals enrolled in the study, 881 of whom were included in the final cohort. Participants were excluded if they (a) did not self-report a professional diagnosis of ASD, (b) did not complete the BDI-II, or (c) answered "Yes" or "Suspected" to a question regarding comorbid Alzheimer's disease (which given the age of participants in our study almost certainly indicated random or careless responding). All participants gave informed consent, and all study procedures were approved by the institutional review board at Vanderbilt University Medical Center.

MTurk Sample. BDI-II data from a general population comparison group were drawn from four online studies of cognitive biases and depressive symptoms that recruited participants using Amazon's Mechanical Turk (MTurk; Total N = 986; Everaert et al., 2018, 2020; Everaert & Joormann, 2019). In each of these studies, participants completed the BDI-II as part of a larger battery of online surveys, for which they were compensated. All included participants from these studies were between the ages of 18 and 46 years, resided in the United States, and had a history of providing good-quality responses on MTurk (i.e., an acceptance ratio of ≥ 95%). To be included in these samples, participants had to provide correct answers to 2–3 reading check questions (e.g., *To show that you are a human, please refuse to answer this question: How many fingers does a typical person have on each hand?*). Additional study-specific data quality measures were also undertaken, including the exclusion of participants who completed the surveys too quickly and those whose longitude/latitude were too close to those of a previous respondent (see original studies for more details). All participants gave informed consent in accordance with the institutional review board at Yale University.

Laboratory Sample. In addition to the online SPARK and MTurk cohorts, we also collected data from 182 individuals (66 ASD, 116 TD) who completed paper-and-pencil BDI-II

forms as part of laboratory studies conducted at Vanderbilt University Medical Center. Data from these individuals have been previously described in multiple reports (Gotham et al., 2018; Han et al., 2019; Unruh et al., 2018). Participants aged 18–46 years were recruited from three diagnostic cohorts: adults with ASD, TD adults with a current depressive disorder, or TD comparisons with no history of ASD or clinically significant depression or anxiety. Participants were recruited from national and local resources, including ResearchMatch, a state autism association, core recruitment services at the Vanderbilt Kennedy Center, and patient enrollment at Vanderbilt University Medical Center. Eligibility criteria included a verbal IQ of 70 or greater, verbal fluency per the Autism Diagnostic Observation Schedule – 2nd edition (ADOS-2; Lord et al., 2012), reading level $\geq 5^{th}$ grade, and no history or concerns of bipolar, psychotic, or substance use disorders. Diagnoses of ASD were confirmed using the ADOS-2 Module 4. The ADOS-2 was also used to rule out ASD in any TD participant who exceeded clinical cut-offs on the Social Responsiveness Scale (SRS-2; Constantino & Gruber, 2012) or Autism Spectrum Quotient. All participants were evaluated for depression using the Structured Clinical Interview for DSM-5 (SCID-5; First et al., 2015) depression module and/or the Mini International Neuropsychiatric Interview (MINI 5.0; Sheehan et al., 1998). Participants received research diagnoses of depression if they met criteria for Major Depressive Disorder or Persistent Depressive Disorder (Dysthymic Disorder) on the SCID-5 or MINI, each of which have algorithms that operationalize DSM criteria. Based on these criteria, 74 individuals (24 ASD, 50 TD) were diagnosed with a current depressive disorder. These rigorous diagnoses of depression in participants with ASD were further used as a diagnostic "gold-standard" to assess the sensitivity and specificity of SPARK sample-derived BDI-II cutoff scores (see "Statistical Analyses" section for more detail). All participants gave informed consent, and all study

procedures were approved by the institutional review board at Vanderbilt University Medical Center.

Measures

Beck Depression Inventory–II (BDI-II) The BDI-II (Beck et al., 1996) is a widely used 21-item self-report measure of depressive symptoms experienced over the past two weeks, with each item rated in severity from 0 to 3. Total scores range from 0 to 63, with scores of 14 or greater typically used to indicate clinically significant depression (Beck et al., 1996). Unlike most other depression questionnaires, the BDI-II does not use item stems and instead employs highly descriptive response options for each item, with higher point values assigned to statements representing more severe depressive symptoms (e.g., when assessing suicidality, 0 = *I don't have any thoughts of killing myself*, 1 = *I have thoughts of killing myself, but I would not carry them out*, 2 = *I would like to kill myself*, and 3 = *I would kill myself if I had the chance.*). This item format has theoretical advantages for use in ASD, as these more detailed items may be more easily interpreted by individuals who have difficulty with more ambiguous response options such as *Rarely* and *Often*.

The BDI-II has strong psychometric properties in the general population and patients drawn from psychiatric or medical settings (Wang & Gorenstein, 2013). Although many studies have disagreed on the factor structure of the BDI-II, a meta-analysis of studies has indicated that the BDI-II represents two highly correlated latent factors of cognitive-affective and somatic-vegetative symptoms (Huang & Chen, 2015). As an alternative to the two correlated-factor model, the BDI-II can be represented by a single general depression factor and two orthogonal group factors representing the cognitive-affective and somatic-vegetative symptom clusters (i.e., a bifactor model; Brouwer et al., 2013; de Miranda Azevedo et al., 2016). The bifactor model of

the BDI-II was utilized in the current study to calculate latent trait scores on the "general depression" factor. In the current study, the BDI-II demonstrated strong model-based reliability and general factor saturation coefficients (Green & Yang, 2009; Rodriguez et al., 2016a, 2016b; Zinbarg et al., 2005) in both the ASD (ω_t = 0.952, ω_H = 0.881) and TD (ω_t = 0.963, ω_H = 0.903) groups.

Generalized Anxiety Disorder–7 (GAD-7). The GAD-7 (Spitzer et al., 2006) is a self-report questionnaire assessing the symptoms of generalized anxiety disorder experienced over the previous two weeks. Participants indicate the frequency of seven anxiety symptoms on a Likert scale ranging from 0 (*not at all*), to 3 (*nearly every day*). Scores range from 0 to 21, with scores of 10 or greater indicating clinically significant anxiety. The psychometric properties of the GAD-7 have been examined extensively in the general population (Kroenke et al., 2010), but its use in the ASD population has been limited (Hull et al., 2020; Russell et al., 2017). The GAD-7 had strong reliability (ω = 0.916) in our SPARK sample (n = 874).

Social Responsiveness Scale–Second Edition (SRS-2). The SRS-2 (Constantino & Gruber, 2012) is a widely used 65-item measure of quantitative autistic traits in both the general population and individuals with ASD. Items are scored on a 4-point Likert scale, with 0 = *not true*, 1 = *sometimes true*, 2 = *often true*, and 3 = *almost always true*. Total scores on the SRS-2 range from 0–195, with higher scores indicating higher levels of autistic symptomatology. T-scores (M = 50, SD = 10) are also available for individuals based on sex and the specific form used. In the current study, the SRS-2 adult self-report form was used in the SPARK cohort as a measure of quantitative autistic traits, from which overall T-scores were derived.

Quality of Life Composite In order to measure global quality of life, we administered items from the World Health Organization Quality of Life – Brief Version (WHOQOL-BREF;

The WHOQOL Group, 1998), a widely-used quality of life measure that has previously been validated in the adult ASD population (McConachie et al., 2018). Items are rated on a 5-point Likert scale with varying response options. The full WHOQOL-BREF contains 26 items: 2 global quality of life items and 24 additional items organized into four domains of physical health, mental health, social relationships and environment. In general population samples, the WHOQOL-BREF can be fit by a bifactor model, which has demonstrated complete factorial invariance across genders (Perera et al., 2018). In the current study, we employed a 5-item global QOL composite, consisting of WHOQOL-BREF items 1 (*How would you rate your quality of life?*), 5 (*How much do you enjoy life?*), 6 (*To what extent do you feel your life to be meaningful?*), 17 (*How satisfied are you with your ability to perform your daily living activities?*), and 19 (*How satisfied are you with yourself?*). Item 1 is one of the form's two "Global QoL" items, and the other four items were good indicators of the general QoL factor in the bifactor model (Mean Item Explained Common Variance [I-ECV] = 0.76, range = 0.69–0.88; Perera et al., 2018). In our SPARK sample (n = 872), these items exhibited adequate fit to a unidimensional factor model (WLSMV estimation; CFI = 0.995, TLI = 0.989, SRMR = 0.035), and reliability for this five-item composite was good (ω = 0.897).

Statistical Analyses

All data analysis was performed in the *R* statistical computing environment (R Core Team, 2020). The BDI-II item responses from all ASD participants (n = 947) were fit to a confirmatory bifactor graded response model (L. Cai, 2010; Samejima, 1969; Toland et al., 2017) based on the factor model of Brouwer and colleagues (2013). This model includes a general factor onto which all items load, along with two specific factors representing the cognitive-affective (CA) and somatic-vegetative (SV) symptoms of depression. The model was

fit using maximum marginal likelihood estimation via the Bock–Aitkin EM algorithm (Bock & Aitkin, 1981), as implemented in the *mirt* R package (Chalmers, 2012). Model fit was assessed using the limited-information C_2 statistic (L. Cai & Monroe, 2014; Monroe & Cai, 2015) as well as C_2-based approximate fit indices. The guidelines for adequate fit (i.e., $RMSEA_2 < 0.089$ and $SRMR < 0.05$) proposed by Maydeu-Olivares & Joe (2014) were used to judge the fit of the IRT model. The assumption of local independence was tested using the standardized local dependency (LD) χ^2 statistic (W.-H. Chen & Thissen, 1997), with χ^2 values greater than 10 indicative of significant local dependence (Toland et al., 2017).

Items were evaluated for DIF in the ASD sample across groups based on sex, gender, age (>30 vs. ≤ 30 years), race (non-Hispanic White vs. Other), level of education (any higher education vs. no higher education), current depression, comorbid anxiety disorder, and lifetime diagnosis of ADHD. Additionally, a multi-group model was fit to the combined ASD and TD sample to test DIF by diagnostic group. DIF was tested using the iterative Wald test procedure proposed by Cao et al. (2017), with p-values < 0.05 (FDR-corrected; Benjamini & Hochberg, 1995) used to flag items for DIF. Significant omnibus Wald tests were followed up with tests of individual item parameters to determine which parameters significantly differed between groups (Stover et al., 2019). The effect sizes proposed by Meade (2010) were used to determine the practical significance of DIF and differential test functioning (DTF) on score comparisons. These effect sizes indices indicate the expected absolute difference in manifest item (UIDS) or test (UETSDS) scores between individuals of different groups possessing the same underlying trait level. As interpretive guidelines for UIDS and UETSDS have not been established, we additionally calculated the expected score standardized difference (ESSD) and expected test score standardized difference (ETSSD), which represent the standardized mean difference in

item or test scores between groups (i.e., DIF/DTF effect sizes in Cohen's *d* metric). As ESSD/ETSSD values of 0.2 are considered "small" (Cohen, 1988; Meade, 2010), we defined practically significant DIF as |ESSD| > 0.2 and practically significant DTF as |ETSSD| > 0.2. DIF testing and effect size calculations were carried out using custom R functions written by the first author (Williams, 2020).

To further test the validity of the BDI-II in ASD, expected a priori (EAP) latent trait scores (Bock & Mislevy, 1982) were calculated for all adults in the ASD sample. Using the *pROC* R package (Robin et al., 2011), we constructed Receiver Operating Characteristic (ROC) curves to evaluate the ability of the BDI-II latent trait score to predict self-reported depression in the SPARK cohort, comparing its performance to that of the BDI-II total score. The area under the ROC curve (*AUC*) was used to quantify the test's discrimination ability, and 95% confidence intervals for *AUC* were constructed using a stratified percentile bootstrap approach. Based on published guidelines for clinical psychological testing, *AUC* values of 0.7–0.8 are considered "fair," values of 0.8–0.9 are considered "good," and values ≥ 0.9 are considered "excellent" (Youngstrom, 2014). Based on the ROC constructed using SPARK data, an optimal diagnostic cutoff for the latent trait score was determined using Youden's *J* index (Youden, 1950). As the BDI-II is most likely to be used clinically to screen individuals with ASD for depressive disorders, we sought to maximize the sensitivity of the test rather than its specificity (Lalkhen & McCluskey, 2008). At minimum, we desired a cutoff score with a sensitivity value of 80% and specificity value of 50% in the SPARK sample. The diagnostic performance of this cutoff was then tested in the sample of 66 ASD individuals who were assessed for depressive disorders in person using structured clinical interviews. Sensitivity, specificity, and positive/negative likelihood ratios (Youngstrom, 2014) were presented for both the latent trait score and BDI-II

total score in both ASD samples, and positive/negative predictive values were also presented for both the observed sample prevalence and the 23% prevalence of current depression derived from meta-analytic methods (Hollocks et al., 2019).

The construct validity of BDI-II scores in this population was assessed by examining relationships between BDI-II scores and a number of clinical and demographic variables. Zero-order Spearman correlations were calculated to quantify the relationships between the BDI-II latent trait scores and the GAD-7 total score, WHOQOL 5-item composite, SRS-2 total T-score, and chronological. Group mean comparisons were undertaken by computing the standardized mean difference (*d*) in latent trait scores by sex (male vs. female), race/ethnicity (non-Hispanic White vs. others), and level of education (at least some college vs. no college). Specific hypotheses regarding effect magnitudes are presented in the Introduction.

Results

Demographics

In total, our sample included BDI-II data from 2049 individuals across the six data sources (Table 1). Participants recruited from SPARK (n = 881, age = 30.94±7.10 years) were predominantly White (78.7%), female (52.8%), and college-educated (71.6% with at least some college). A sizable portion of this sample (9.2%) also identified as a non-binary gender, reflecting the known increase in gender variance seen in individuals with ASD (Cooper et al., 2018). Eighty-two percent of the SPARK sample reported at least one current professionally diagnosed psychiatric condition other than ASD (i.e., they had experienced symptoms of the condition within the last three months or were receiving ongoing treatment for that condition), with a median of 2 current comorbidities (*IQR* = [1, 4]). As would be expected, the most common comorbidities reported were anxiety disorders (64%), depressive disorders (53%), and

ADHD (36%), followed by PTSD (24%) and OCD (18%). The combined MTurk sample (n = 986, age = 32.60±6.85 years) had similar demographics to the SPARK sample, with a slightly higher portion of the MTurk participants reporting at least some higher education (84.7%). Compared to the online samples, the ASD and TD participants recruited from Vanderbilt tended to be younger and more highly educated than the SPARK and MTurk samples, respectively (Table 1). Both diagnostic groups exhibited relatively high mean scores on the BDI-II (combined ASD groups: 17.18±12.85; combined TD groups: 15.55±13.29; d = 0.125, 95% CI [0.037, 0.212]), with 55% and 49% of the combined ASD and TD samples screening positive for depression on the BDI-II, respectively.

[Table 1 around here]

IRT Model

The bifactor graded response model fit the item responses of the ASD sample well ($C_2(168) = 528.59$, $p < 0.001$, $CFI_{C2} = 0.990$, $TLI_{C2} = 0.987$, $RMSEA_{C2} = 0.048$ [0.044, 0.053], SRMR = 0.037). Given the adequate global model fit statistics, item-level fit statistics were not examined. All items loaded strongly on the general factor ($\lambda_{Mean} = 0.71$; $\lambda_{range} = 0.56–.87$; Table 2), with a large proportion of common variance explained by this factor (ECV = 0.83, $I\text{-}ECV$ range = 0.68–1.00). Reliability of the general factor score was good ($\rho_{Mean} = 0.895$, bootstrapped 95% CI = [0.888, 0.902], $\rho_{range} = 0.676–0.995$), with the only reliability values less than 0.70 exhibited by participants answering "0" to all 21 questions of the BDI-II. Of note, the cognitive-affective and somatic-vegetative group factors exhibited poor reliability ($\rho_{Mean} = 0.546$ [0.521, 0.570] and 0.530 [0.506, 554], respectively), signifying that latent trait scores on these BDI-II factors are difficult to interpret. Furthermore, subscale-level omega-hierarchical values derived from the bifactor structure (ω_{HS}; Rodriguez et al., 2016a, 2016b) were very low (0.180 and 0.048

respectively), indicating that the BDI-II cognitive-affective and somatic-vegetative subscales do not represent meaningfully different constructs from the measure's total score or general factor. Thus, when considering the construct validity of the BDI-II IRT score, we restricted our analysis to only include latent scores on the general factor (θ_G). Item response category characteristic curves (conditional on $\theta_{CA} = \theta_{SV} = 0$) for the 21 BDI-II items are presented in Supplemental Figure S1.

Significant local dependence was found for one pair of items (4: "*Loss of Pleasure"* and 12: "*Loss of Interest"*; standardized LD-χ^2 = 13.42), likely reflecting the conceptual overlap of these two items. Notably, Yen's (1984) Q_3 residual correlation for this item pair was -0.008, a value that is typically not indicative of significant LD. Given that that combined criterion "loss of interest or pleasure" is one of two symptom options necessary for a major depressive disorder diagnosis (the other being "depressed mood"), we did not modify the scale by dropping either of those items. We did, however re-fit the IRT model, combining scores on items 4 and 12 into a single 7-point polytomous super-item reflecting the diagnostic criterion. As the latent general factor scores estimated by this model were nearly identical to the original model's scores (r = 0.994), we chose to retain the original IRT model for further analyses.

[Table 2 around here]

Differential Item and Test Functioning

DIF analyses within the ASD group indicated that all items functioned similarly in groups based on sex at birth, race/ethnicity, level of education, self-reported lifetime ADHD diagnosis, and self-reported current anxiety. Item 10 (*Crying*) exhibited small but practically significant DIF by gender (UIDS = 0.167, ESSD = -0.274). However, this single DIF item was not large enough to result in a practically significant amount of DTF (UETSDS = 0.167, ETSSD

= -0.011). In addition, items 8 (*Self-Criticalness*; UIDS = 0.181, ESSD = 0.236), and 21 (*Loss of Interest in Sex*; UIDS = 0.240, ESSD = -0.506) demonstrated practically significant amounts of DIF by age group. The DIF from these items canceled somewhat at the test level, and thus the overall impact of age on DTF was negligible (UETDS = 0.194, ETSSD = -0.004). Lastly, self-reported current depression was associated with DIF in items 1 (*Sadness*: UIDS = 0.272, ESSD = 0.601), 13 (*Indecisiveness*; UIDS = 0.319, ESSD = -0.448), and 19 (*Concentration Difficulty*; UIDS = 0.218, -0.351), but the overall effect on DTF remained practically insignificant (UETSDS = 0.279, ETSSD = -0.022). Full results of the DIF analyses are presented in Supplemental Table S3.

DIF analysis between the ASD and TD groups revealed that 18 of the 21 BDI-II items (all but items 4: *Loss of Pleasure*, 5: *Guilty Feelings*, and 16: *Changes in Sleeping Pattern*) exhibited significant DIF by diagnostic group (Table 3). However, expected score differences on nearly all items were too small to be of practical significance. Items that did exhibit practically significant DIF included 9 (*Suicidal Thoughts or Wishes*: UIDS = 0.093, ESSD = -0.220), 17 (*Irritability*: UIDS = 0.130, ESSD = 0.219), 19 (*Concentration Difficulty*: UIDS = 0.133, ESSD = -0.205), and 21 (*Loss of Interest in Sex*: UIDS = 0.117, ESSD = 0.233), with effects being small in each case. Moreover, the total effect of all 18 items on DTF between the diagnostic groups was practically negligible, with expected BDI-II score differences of only 0.524 points between ASD and TD respondents of the same latent trait levels (ETSSD = -0.039).

Although the ASD and TD samples used to examine DIF by diagnostic group were relatively well-matched on demographic variables, these samples were both majority female and thus poorly representative of the overall ASD population (i.e., a 3:1 male to female ratio; Loomes et al., 2017). Thus, in order to determine whether our conclusions about DIF/DTF of the

BDI-II would be similarly valid in male-predominant ASD samples, we repeated our DIF analyses in the subsample of male participants (n_{ASD} = 350, n_{TD} = 406). In male participants, we found evidence of significant DIF by diagnostic group in six of the 21 items, only one of which reached the threshold for practical significance (item 6: *Punishment Feelings*: UIDS = 0.148, ESSD = -0.248*;* Supplementary Table S3). As with the full sample, the combined effect of these DIF items on DTF between diagnostic groups was small and practically insignificant (UETSDS = 0.333, ETSSD = -0.025).

[Table 3 around here]

Diagnostic Performance

Using the BDI-II latent trait scores, we constructed receiver operating characteristic (ROC) curves to predict self-reported current depression in the SPARK sample. Of 868 participants responding to this question, 499 (57.5%) indicated that they had experienced depression symptoms [either professionally diagnosed or suspected] in the past three months or were currently undergoing depression treatment. BDI-II latent trait scores demonstrated fair-to-good ability to discriminate between those with and without current depressive symptoms (*AUC* = 0.796, 95% CI [0.763, 0.826]; Figure 1). Youden's *J* index indicated an optimal cutpoint of θ_G = -0.0893, resulting in a sensitivity and specificity above our a priori 80% and 50% threshold (Table 4). Notably, when excluding individuals with "suspected" depression from the ROC analyses, the results were essentially unchanged (*AUC* = 0.796, 95% CI [0.764, 0.826]), and Youden's *J* indicated an identical optimal cutpoint (θ_G = -0.0893, sensitivity = 0.823 [0.787, 0.857], specificity = 0.648 [0.597, 0.699]). With the high prevalence of current depression in our SPARK sample, the positive and negative predictive values of this score cutoff were both in the 0.7–0.8 range. However, when adjusting these values for the 23% estimated prevalence of

current depression in adults with ASD (Hollocks et al., 2019), positive predictive value decreased (0.414 [0.382, 0.451]) and negative predictive value increased (0.924 [0.909, 0.938]). In the full SPARK sample, the BDI-II total score performed similarly to the IRT score in terms of *AUC*, but the standard total score cutoff of 14 points or greater (Beck et al., 1996) demonstrated a lower sensitivity and higher specificity than the IRT score.

The discrimination ability of the BDI-II IRT and total scores were then examined in the clinical sample of 66 ASD adults (36.4% depressed) diagnosed with structured clinical interviews (either the SCID-5 or MINI). In this sample, the *AUC* of the latent trait score was somewhat lower than in the SPARK sample, although still deemed "fair" (Table 4.) Notably, due to the small sample size, 95% confidence intervals were very wide for all diagnostic efficiency statistics and these data were thus compatible with population *AUC* values ranging from "poor" to "good" (Youngstrom, 2014). Similarly, while the point estimate of sensitivity was slightly below the *a priori* 80% threshold, the confidence interval on this estimate was not able to exclude the possibility that sensitivity was above 80% in the population. As the prevalence of depression in this sample was lower than the SPARK sample, the positive predictive value of this cutoff was lower than in the online sample, whereas negative predictive value was higher. However, when adjusting for the population prevalence of depression in ASD, positive and negative predictive values were both slightly lower than those in the SPARK sample (i.e., a 4–7% decrease; Table 4). The *AUC* value for the BDI-II total score was again similar to that of the IRT score in this sample, with slightly higher values for the IRT score in both cohorts. However, in the Vanderbilt sample, a BDI-II score of 14 points or more had values of sensitivity and specificity both between 60% and 70%, indicating that this cutoff was likely not appropriate for screening purposes in adults with ASD.

[Figure 1 and Table 4 around here]

Overall, the BDI-II latent trait scores demonstrated a pattern of correlations consistent with our hypotheses, suggesting that the nomological network for the BDI-II in ASD is similar to that in the general population and consistent with prior correlational studies in ASD. As expected, the BDI-II scores had strong positive correlations with GAD-7 scores (r_s = 0.739, 95% CI [0.705, 0.770]) and strong negative correlations with WHOQOL composite scores (r_s = -0.719 [-0.752, -0.683]), supporting the criterion validity of the measure. A smaller but still substantial correlation was seen with SRS-2 T-scores (r_s = 0.497 [0.440, 0.551]), and the difference in correlations between GAD-7 and SRS-2 scores was greater than our 0.2 threshold for discriminant validity (Δr_s = 0.242). As hypothesized, females had higher mean BDI-II IRT scores than males (d = 0.348 [0.210, 0.486]), further confirming the ability of the BDI-II to capture known sex differences in depression prevalence in ASD (Lai et al., 2019). To further support the discriminant validity of the measure, no significant correlation was noted between BDI-II scores and age (r_s = 0.061 [-0.005, 0.127]), and no statistically significant differences were found between groups defined by race/ethnicity (d = 0.129 [-0.032, 0.291]) or education level (d = -0.093 [-0.240, 0.053]).

Discussion

Depressive disorders remain a major source of disability in the population of adults with ASD, and substantial future work is necessary to better understand and treat these highly comorbid conditions. Despite the scope of this problem, few studies have systematically assessed the psychometric properties of depression measures in ASD samples, and the suitability of many of these measures for clinical or research applications remains largely unknown (Cassidy,

Bradley, Bowen, et al., 2018). This study investigated the psychometric properties of the BDI-II in a large sample of adults with ASD, and our findings support both the reliability and validity of the BDI-II in this population. The bifactor structure of the BDI-II proposed by Brouwer and colleagues (2013) fit the item responses in both diagnostic groups well, and model-based reliability indices supported the interpretation that the BDI-II is essentially unidimensional (i.e., strongly saturated with a general factor; Rodriguez et al., 2016a, 2016b). Furthermore, examination of DIF across many demographic and clinical variables indicated that these items are largely endorsed in a similar manner by all subsets of adults with ASD. Practically significant DIF was present in a minority of items, but the test score differences resulting from this DIF were small enough to be practically ignorable. Finally, the relationships between BDI-II general factor scores and other clinical and demographic variables suggests that the construct validity of the BDI-II is similar in adults with ASD and the general population. These results as a whole provide strong empirical support for the use of the BDI-II as a dimensional measure of depression symptoms in adults with ASD.

In addition to testing the psychometric properties of the BDI-II, we sought to address the hypothesis that the cognitive differences of adults with ASD create substantial differences in the ways that this population answers questions about affective symptoms (Cassidy, Bradley, Bowen, et al., 2018; Gotham et al., 2015; Pezzimenti et al., 2019; Uljarević et al., 2018). Contrary to this belief, our differential test functioning analyses did not find evidence for meaningful test score differences on the BDI-II. This finding was not dependent on the gender breakdown of our sample, as a DIF sensitivity analysis on only male participants came to similar conclusions. Although the majority of BDI-II items did exhibit statistically significant DIF across diagnostic groups, the effect sizes of these differences were trivially small and unlikely to

have a meaningful effect on observed scores. However, practically significant DIF was observed in item 9 (*Suicidal Thoughts or Wishes*), with higher levels of depression required for individuals in the ASD group to endorse the statement "*I have thoughts of killing myself, but I would not carry them out*." Interestingly, this finding runs counter to previous results suggesting that adults with ASD may endorse suicidal ideation at a relatively high rate even when not reporting depression (Cassidy et al., 2014). Practically significant DIF was also found in items 17 (*Irritability*) and 19 (*Concentration Difficulty*), and 21 (*Loss of Interest in Sex*). Individuals with ASD endorsed the statements "*I am more irritable than usual*" and "*I am less interested in sex than I used to be*" more easily than their TD counterparts, whereas the statement "*It's hard to keep my mind on anything for very long*" required a higher level of depression in the ASD group to be endorsed. Although reasons for these differences cannot be determined without further study, differential responses to item 21 are consistent with prior reports of lower libido and sexual desire in some adults with ASD (Bejerot & Eriksson, 2014; Byers et al., 2013). As the combined effects of the 18 items with "significant" DIF on overall DTF was quite minimal (0.524 points, a standardized difference of d = -0.039), we contend that scores on the BDI-II can be thought of as equivalent in adults both with and without ASD. Although large and practically significant DIF/DTF may exist in ASD for other measures of depressive symptomatology, these findings indicate that the interpretation of BDI-II items is not meaningfully affected by the cognitive differences associated with ASD.

Although other studies have assessed the latent structure, reliability, and construct validity of depression measures in ASD (Arnold et al., 2019; Uljarević et al., 2018), this study additionally sought to determine how well the BDI-II total and IRT scores discriminated between depressed and non-depressed adults with ASD. In the SPARK sample, both the BDI-II general

factor score (*AUC* = 0.796) and BDI-II total score (*AUC* = 0.791) demonstrated a fair-to-good ability to discriminate between those reporting current depression and those who did not. These values are similar to the approximate *AUC* value calculated from the standardized mean difference in PHQ-9 scores between non-depressed and depressed adults with ASD in the study of Arnold and colleagues (*d* = 1.262, approximate *AUC* = 0.814). With regard to the newly derived latent trait score, Youden's *J* suggested a cutpoint with relatively good sensitivity (82%) and relatively poor specificity (65%). In contrast, a BDI-II score at the typical cutoff of 14 or greater demonstrated somewhat reduced sensitivity (74%) and increased specificity (69%) compared to the latent trait score. These cutoffs were then used to predict gold-standard depression diagnoses in a sample of 66 rigorously-phenotyped adults with ASD. In this sample, neither BDI-II score performed as well, with 75% sensitivity and 55% specificity for the latent trait score and 63% sensitivity and 67% specificity for the total score. However, this sample was much smaller, and the wide confidence intervals around the diagnostic efficiency statistics were not able to exclude either the point estimates from the SPARK sample or our *a priori* cutoff values of 80% sensitivity and 50% specificity. Future work in larger samples of adults with ASD with gold-standard mood disorder diagnoses is thus required to better estimate the true diagnostic efficiency of the BDI-II in adults with ASD.

Although the sensitivity and specificity of the BDI-II scores in the Vanderbilt cohort were lower than expected, these figures do not preclude the scale's usefulness for clinical practice. The BDI-II latent trait score demonstrated moderate sensitivity in both of the tested samples, and thus this measure has the potential to serve as a screening measure for depression in individuals on the autism spectrum. Notably, when using the meta-analytically estimated prevalence of current depression in adults with ASD (23%; Hollocks et al., 2019), estimates of

negative predictive value were relatively high (0.884–0.924), supporting the use of the BDI-II to screen out depression in clinical settings.

Although total scores discriminated nearly as well as latent trait scores as measured by the AUC, the total score cutoffs that achieved similar levels of sensitivity captured more false positives than the corresponding latent trait scores. In addition to its marginally improved specificity over the equivalent total score cutoff, the IRT-derived latent trait score possesses several other advantageous properties, including the accommodation of missing data, more realistic score confidence intervals, and the ability to discriminate between individuals whose total scores on the questionnaire are equal. Thus, until another measure of depression is shown to have higher diagnostic accuracy in this population, we recommend that the BDI-II latent trait score be utilized to screen for depression in adults with ASD. Nevertheless, given the low specificity and positive predictive values found in our samples, we caution against the use of the BDI-II alone to characterize individuals with ASD as being depressed or not. In line with the recommendations of Pezzimenti and colleagues (2019), we suggest that depression is best diagnosed by clinical interview and by employing information from multiple informants, including a self-report measure such as the BDI-II. Additional research will be needed to determine which combination of symptoms can best be utilized to screen for depression in this population with high sensitivity and specificity.

Although projects to create better clinical tools for depression assessment in ASD are ongoing (e.g., https://gtr.ukri.org/projects?ref=ES/N000501/1), our hope is that the use of psychometrically validated instruments such as the BDI-II can improve the scientific study and clinical management of depression in ASD until these measures have been fully developed. One major obstacle preventing the widespread use of the BDI-II in clinical or research settings is the

knowledge and expertise needed to calculate IRT-based latent trait scores from published item parameters. In order to overcome this barrier, we have developed a free online calculator (available at https://asdmeasures.shinyapps.io/bdi_score/) that will take BDI-II item scores as input and calculate (a) latent trait scores, (b) score confidence intervals, (c) individual score reliability, (d) an indication as to whether the individual screened positive for depression. The calculator can also generate individual printable score reports, which can easily be stored within a patient/participant file or uploaded to a medical record. We hope that the availability of this calculator can facilitate the use of evidence-based depression assessment in adults with ASD and improve the overall quality of research and clinical care involving this population.

Strengths and Limitations

This study had a number of strengths, including a large, geographically-diverse sample of adults with ASD, a broad range of measures to establish the nomological network of depression symptoms in this population, the inclusion of a large TD group with similar demographics and depressive symptom severity, and a smaller sample of individuals in which the BDI-II and structured interview-based clinical diagnoses of depression could be compared. Furthermore, by conducting analyses within an IRT framework, we were able to calculate latent trait scores, which in addition to their theoretical benefits were marginally better at discriminating between depressed and non-depressed ASD adults than did BDI-II total scores. We also provide an easy-to-use online calculator that allows these trait scores to be easily employed by clinicians and researchers. Lastly, the DIF/DTF analyses performed in this study allowed us to demonstrate that adults with and without ASD respond in a similar manner to questions on the BDI-II.

However, the study was not without its limitations. For one, the data utilized in this study was drawn from a number of different experiments, each with its own inclusion/exclusion

criteria, data quality assurance methods, and battery of measures administered. By far the largest limitation was the fact that, the MTurk samples were not properly screened for ASD, and there were likely individuals in the TD cohort with ASD diagnoses. However, given the low prevalence of ASD in unselected samples recruited from MTurk (e.g., Mitchell & Locke, 2015; Skylark & Baron-Cohen, 2017), the number of "TD" adults with unrecognized ASD in our sample was likely too few to meaningfully affect any of the conducted DIF analyses. Other limitations had to do with the ways in which diagnostic categories were assigned. As with many large-scale survey studies, we used self-report rather than clinical interviews to confirm autism and depression diagnoses in the SPARK cohort. Additionally, the ASD sample diagnosed with structured interviews was relatively small (n = 66), causing our estimates of sensitivity and specificity in this sample to be quite imprecise.

Another limitation of this study is the representativeness of the ASD sample, which was overwhelmingly female and college educated. Despite ASD being more prevalent in males at a ratio of at least 3:1 (Loomes et al., 2017), only 40% of our sample was male, and 72% had enrolled in at least some higher education, substantially higher than the 43% figure reported in the National Longitudinal Transition Study-2 (Newman et al., 2011). Notably, one strength of IRT is the ability to derive unbiased estimates of item parameters from samples that are not representative of the population of interest (Embretson, 1996). DIF by sex and education level was also found to be minimal, and thus it is unlikely that substantially different conclusions would be generated in a more representative sample. Nevertheless, we performed a sensitivity analysis of gender by testing DIF in the subset of male participants, finding once again that the expected total score differences across groups were not meaningfully different. One final limitation concerned the cross-sectional nature of this study, which did not allow us to estimate

the temporal stability, DIF over multiple administrations, or sensitivity to change of BDI-II IRT scores in the ASD sample. Future work including repeated BDI-II administration will be necessary to determine whether this measure is appropriate for tracking depression symptoms in ASD over the course of clinical trials or longitudinal observational studies.

Conclusion

This study built on previous work (Cassidy, Bradley, Bowen, et al., 2018; Gotham et al., 2015) to investigate the psychometric properties of the BDI-II in adults with ASD. Employing an IRT framework, we were able to determine that the BDI-II represents the same latent constructs in ASD and TD samples, and both groups respond to items of the measure in much the same manner. Moreover, the pattern of relationships between BDI-II scores and other variables is similar in adults with and without diagnosed ASD. Overall, our findings indicate that the BDI-II possesses the appropriate psychometric properties to serve as a dimensional measure of depressive symptoms that is comparable between individuals with ASD and the general population.

We also examined the diagnostic efficiency of the BDI-II, finding support for the use of the BDI-II as a clinical screening tool. The latent trait score calculated from the IRT model discriminates moderately between depressed and non-depressed adults with ASD, possessing appropriate sensitivity and specificity values for use in screening adults with ASD for depression. To facilitate the use of BDI-II IRT scores in research and clinical care, we have developed an easy-to-use online calculator that is freely available to clinicians and researchers (https://asdmeasures.shinyapps.io/bdi_score/). Although more work is needed, for example to develop symptom measures that capture the unique presentations of depression in ASD, we

believe that the BDI-II can provide clinicians and researchers with an evidence-based option for depression assessment until validated autism-specific tools with enhanced predictive validity become available.

Data Availability

Approved researchers can obtain the SPARK population dataset described in this study by applying at https://base.sfari.org. Data from the MTurk samples included in this study are available at https://osf.io/677jr/.The remainder of the data and materials used in this study are available from the first author upon reasonable request.

References:

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

Arnold, S. R. C., Uljarević, M., Hwang, Y. I., Richdale, A. L., Trollor, J. N., & Lawson, L. P. (2019). Brief report: Psychometric properties of the atient health questionaire-9 (PHQ-9) in autistic adults. *Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s10803-019-03947-9

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *BDI-II, Beck Depression Inventory: Manual* (2nd ed). Psychological Corporation.

Bejerot, S., & Eriksson, J. M. (2014). Sexuality and gender role in autism spectrum disorder: A case control study. *PLoS One, 9*(1), e87961. https://doi.org/10.1371/journal.pone.0087961

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *57*(1), 289–300. https://doi.org/10.2307/2346101

Bishop-Fitzpatrick, L., & Rubenstein, E. (2019). The physical and mental health of middle aged and older adults on the autism spectrum and the impact of intellectual disability. *Research in Autism Spectrum Disorders*, *63*, 34–41. https://doi.org/10.1016/j.rasd.2019.01.001

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/bf02293801

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). On the factor structure of the Beck Depression Inventory-II: G is the key. *Psychological Assessment*, *25*(1), 136–145. https://doi.org/10.1037/a0029228

Buchsbaum, M. S., Hollander, E., Haznedar, M. M., Tang, C., Spiegel-Cohen, J., Wei, T. C., Solimando, A., Buchsbaum, B. R., Robins, D., Bienstock, C., Cartwright, C., & Mosovich, S. (2001). Effect of fluoxetine on regional cerebral metabolism in autistic spectrum disorders: A pilot study. *International Journal of Neuropsychopharmacology*, *4*(2), 119 – 125. https://doi.org/10.1017/s1461145701002280

Burns, A., Irvine, M., & Woodcock, K. (2019). Self-Focused attention and depressive symptoms in adults with autistic spectrum disorder (ASD). *Journal of Autism and Developmental Disorders*, *49*(2), 692–703. https://doi.org/10.1007/s10803-018-3732-5

Byers, E. S., Nichols, S., & Voyer, S. D. (2013). Challenging stereotypes: Sexual functioning of single adults with high functioning autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *43*(11), 2617–2627. https://doi.org/10.1007/s10803-013-1813-z

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581–612. https://doi.org/10.1007/s11336-010-9178-0

Cai, L., & Monroe, S. (2014). *A new statistic for evaluating item response theory models for ordinal data* (CRESST Report 839; pp. 1–28). University of California, National Center

for Research on Evaluation, Standards, and Student Testing (CRESST). https://eric.ed.gov/?id=ED555726

Cai, R. Y., Richdale, A. L., Dissanayake, C., & Uljarević, M. (2018). Brief report: Inter-relationship between emotion regulation, intolerance of uncertainty, anxiety, and depression in youth with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *48*(1), 316–325. https://doi.org/10.1007/s10803-017-3318-7

Cao, M., Tay, L., & Liu, Y. (2017). A Monte Carlo study of an iterative Wald test procedure for DIF analysis. *Educational and Psychological Measurement*, *77*(1), 104–118. https://doi.org/10.1177/0013164416637104

Cassidy, S. A., Bradley, L., Bowen, E., Wigham, S., & Rodgers, J. (2018). Measurement properties of tools used to assess suicidality in autistic and general population adults: A systematic review. *Clinical Psychology Review*, *62*, 56–70. https://doi.org/10.1016/j.cpr.2018.05.002

Cassidy, S. A., Bradley, L., Shaw, R., & Baron-Cohen, S. (2018). Risk markers for suicidality in autistic adults. *Molecular Autism*, *9*(1), 42. https://doi.org/10.1186/s13229-018-0226-4

Cassidy, S. A., Bradley, P., Robinson, J., Allison, C., McHugh, M., & Baron-Cohen, S. (2014). Suicidal ideation and suicide plans or attempts in adults with Asperger's syndrome attending a specialist diagnostic clinic: A clinical cohort study. *The Lancet Psychiatry*, *1*(2), 142–147. https://doi.org/10.1016/s2215-0366(14)70248-2

Cederlund, M., Hagberg, B., & Gillberg, C. (2010). Asperger syndrome in adolescent and young adult males. Interview, self- and parent assessment of social, emotional, and cognitive problems. *Research in Developmental Disabilities*, *31*(2), 287–298. https://doi.org/10.1016/j.ridd.2009.09.006

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, M.-H., Pan, T.-L., Lan, W.-H., Hsu, J.-W., Huang, K.-L., Su, T.-P., Li, C.-T., Lin, W.-C., Wei, H.-T., Chen, T.-J., & Bai, Y.-M. (2017). Risk of suicide attempts among adolescents and young adults with autism spectrum disorder: A nationwide longitudinal follow-up study. *The Journal of Clinical Psychiatry*, *78*(9), e1174–e1179. https://doi.org/10.4088/jcp.16m11100

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.3102/10769986022003265

Clark, D. A., Steer, R. A., & Beck, A. T. (1994). Common and specific dimensions of self-reported anxiety and depression: Implications for the cognitive and tripartite models. *Journal of Abnormal Psychology*, *103*(4), 645–654. https://doi.org/10.1037/0021-843x.103.4.645

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037//0003-066x.49.12.997

Constantino, J. N., & Gruber, C. P. (2012). *Social Responsiveness Scale–Second Edition (SRS-2): Manual* (2nd ed.). Western Psychological Services.

Cooper, K., Smith, L. G. E., & Russell, A. J. (2018). Gender identity in autism: Sex differences in social affiliation with gender groups. *Journal of Autism and Developmental Disorders, 48*(12), 3995–4006. https://doi.org/10.1007/s10803-018-3590-1

Crane, L., Goddard, L., & Pring, L. (2013). Autobiographical memory in adults with autism spectrum disorder: The role of depressed mood, rumination, working memory and theory of mind. *Autism, 17*(2), 205–219. https://doi.org/10.1177/1362361311418690

Croen, L. A., Zerbo, O., Qian, Y., Massolo, M. L., Rich, S., Sidney, S., & Kripke, C. (2015). The health status of adults on the autism spectrum. *Autism, 19*(7), 814–823. https://doi.org/10.1177/1362361315577517

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Daniels, A. M., Rosenberg, R. E., Anderson, C., Law, J. K., Marvin, A. R., & Law, P. A. (2012). Verification of parent-report of child autism spectrum disorder diagnosis to a web-based autism registry. *Journal of Autism and Developmental Disorders, 42*(2), 257–265. https://doi.org/10.1007/s10803-011-1236-7

Davignon, M. N., Qian, Y., Massolo, M., & Croen, L. A. (2018). Psychiatric and medical conditions in transition-aged individuals With ASD. *Pediatrics, 141*(Suppl 4), S335–S345. https://doi.org/10.1542/peds.2016-4300k

de Miranda Azevedo, R., Roest, A. M., Carney, R. M., Denollet, J., Freedland, K. E., Grace, S. L., Hosseini, S. H., Lane, D. A., Parakh, K., Pilote, L., & Jonge, P. de. (2016). A bifactor model of the Beck Depression Inventory and its association with medical prognosis after myocardial infarction. *Health Psychology, 35*(6), 614–624. https://doi.org/10.1037/hea0000316

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341–349. https://doi.org/10.1037/1040-3590.8.4.341

Everaert, J., Bronstein, M. V., Cannon, T. D., & Joormann, J. (2018). Looking through tinted glasses: Depression and social anxiety are related to both interpretation biases and inflexible negative Interpretations. *Clinical Psychological Science*, *6*(4), 517–528. https://doi.org/10.1177/2167702617747968

Everaert, J., Bronstein, M. V., Castro, A. A., Cannon, T. D., & Joormann, J. (2020). When negative interpretations persist, positive emotions don't! Inflexible negative interpretations encourage depression and social anxiety by dampening positive emotions. *Behaviour Research and Therapy*, *124*, 103510. https://doi.org/10.1016/j.brat.2019.103510

Everaert, J., & Joormann, J. (2019). Emotion regulation difficulties related to depression and anxiety: A network approach to model relations among symptoms, positive reappraisal, and repetitive negative thinking. *Clinical Psychological Science*, *7*(6), 1304–1318. https://doi.org/10.1177/2167702619859342

Farmer, C. A., Kaat, A., Thurm, A., Anselm, I., Akshoomoff, N., Bennett, A., Berry, L., Bruchey, A., Barshop, B. A., Berry-Kravis, E., Bianconi, S., Cecil, K. M., Davis, R. J., Ficicioglu, C., Porter, F. D., Wainer, A., Goin-Kochel, R. P., Leonczyk, C., Guthrie, W., … Miller, J. S. (In Press). Person ability scores as an alternative to norm-referenced scores as outcome measures in studies of neurodevelopmental disorders. *American Journal on Intellectual and Developmental Disabilities*. https://pdfs.semanticscholar.org/03e8/3a4febaa232b202008ee2f4049ab7705a7b2.pdf

Feliciano, P., Daniels, A. M., Snyder, L. G., Beaumont, A., Camba, A., Esler, A., Gulsrud, A. G., Mason, A., Gutierrez, A., Nicholson, A., Paolicelli, A. M., McKenzie, A. P., Rachubinski, A. L., Stephens, A. N., Simon, A. R., Stedman, A., Shocklee, A. D., Swanson, A., Finucane, B., … Chung, W. K. (2018). SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. *Neuron*, *97*(3), 488–493. https://doi.org/10.1016/j.neuron.2018.01.015

First, M. B., Williams, J. B. W., Karg, R. S., & Spitzer, R. L. (2015). *Structured clinical interview for DSM-5—Research version (SCID-5 for DSM-5, research version; SCID-5-RV)*. American Psychiatric Association Publishing.

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*(15), 191–197. https://doi.org/10.1016/j.jad.2016.10.019

Gotham, K. O., Bishop, S. L., Brunwasser, S., & Lord, C. (2014). Rumination and perceived impairment associated with depressive symptoms in a verbal adolescent-adult ASD sample. *Autism Research*, *7*(3), 381–391. https://doi.org/10.1002/aur.1377

Gotham, K. O., Siegle, G. J., Han, G. T., Tomarken, A. J., Crist, R. N., Simon, D. M., & Bodfish, J. W. (2018). Pupil response to social-emotional material is associated with rumination and depressive symptoms in adults with autism spectrum disorder. *PloS One*, *13*(8), e0200340. https://doi.org/10.1371/journal.pone.0200340

Gotham, K. O., Unruh, K., & Lord, C. (2015). Depression and its measurement in verbal adolescents and adults with autism spectrum disorder. *Autism*, *19*(4), 491–504. https://doi.org/10.1177/1362361314536625

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*(1), 155–167. https://doi.org/10.1007/s11336-008-9099-3

Griffiths, S., Allison, C., Kenny, R., Holt, R., Smith, P., & Baron-Cohen, S. (2019). The Vulnerability Experiences Quotient (VEQ): A study of vulnerability, mental health and life satisfaction in autistic adults. *Autism Research, 12*(10), 1516 – 1528. https://doi.org/10.1002/aur.2162

Han, G. T., Tomarken, A. J., & Gotham, K. O. (2019). Social and nonsocial reward moderate the relation between autism symptoms and loneliness in adults with ASD, depression, and controls. *Autism Research, 12*(6), 884 – 896. https://doi.org/10.1002/aur.2088

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38*(9 Suppl), II28–II42.

Hedley, D., Uljarević, M., Wilmot, M., Richdale, A., & Dissanayake, C. (2018). Understanding depression and thoughts of self-harm in autism: A potential mechanism involving loneliness. *Research in Autism Spectrum Disorders, 46*, 1–7. https://doi.org/10.1016/j.rasd.2017.11.003

Hill, E., Berthoz, S., & Frith, U. (2004). Brief report: Cognitive processing of own emotions in individuals with autistic spectrum disorder and in their relatives. *Journal of Autism and Developmental Disorders, 34*(2), 229–235. https://doi.org/10.1023/b:jadd.0000022613.41399.14

Hillier, A. J., Fish, T., Siegel, J. H., & Beversdorf, D. Q. (2011). Social and vocational skills training reduces self-reported anxiety and depression among toung adults on the autism

spectrum. *Journal of Developmental and Physical Disabilities*, *23*(3), 267–276. https://doi.org/10.1007/s10882-011-9226-4

Hirvikoski, T., Boman, M., Chen, Q., D'Onofrio, B. M., Mittendorfer-Rutz, E., Lichtenstein, P., Bölte, S., & Larsson, H. (2019). Individual risk and familial liability for suicide attempt and suicide in autism: A population-based study. *Psychological Medicine*, 1–12. https://doi.org/10.1017/S0033291719001405

Hofvander, B., Delorme, R., Chaste, P., Nydén, A., Wentz, E., Stahlberg, O., Herbrecht, E., Stopin, A., Anckarsater, H., Gillberg, C., Rastam, M., & Leboyer, M. (2009). Psychiatric and psychosocial problems in adults with normal-intelligence autism spectrum disorders. *BMC Psychiatry*, *9*(1), 35. https://doi.org/10.1186/1471-244x-9-35

Hollocks, M. J., Lerh, J. W., Magiati, I., Meiser-Stedman, R., & Brugha, T. S. (2019). Anxiety and depression in adults with autism spectrum disorder: A systematic review and meta-analysis. *Psychological Medicine*, *49*(4), 559–572. https://doi.org/10.1017/s0033291718002283

Howlin, P., & Magiati, I. (2017). Autism spectrum disorder: Outcomes in adulthood. *Current Opinion in Psychiatry*, *30*(2), 69–76. https://doi.org/10.1097/yco.0000000000000308

Huang, C., & Chen, J.-H. (2015). Meta-analysis of the factor structures of the Beck Depression Inventory-II. *Assessment*, *22*(4), 459–472. https://doi.org/10.1177/1073191114548873

Hull, L., Lai, M.-C., Baron-Cohen, S., Allison, C., Smith, P., Petrides, K. V., & Mandy, W. (2020). Gender differences in self-reported camouflaging in autistic and non-autistic adults. *Autism*, *24*(2), 352–363. https://doi.org/10.1177/1362361319864804

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00109

Kraper, C. K., Kenworthy, L., Popal, H., Martin, A., & Wallace, G. L. (2017). The gap between adaptive behavior and intelligence in autism persists into young adulthood and is linked to psychiatric co-morbidities. *Journal of Autism and Developmental Disorders*, *47*(10), 3007–3017. https://doi.org/10.1007/s10803-017-3213-2

Kroenke, K., Spitzer, R. L., Williams, J. B. W., & Löwe, B. (2010). The Patient Health Questionnaire somatic, anxiety, and depressive symptom scales: A systematic review. *General Hospital Psychiatry*, *32*(4), 345–359. https://doi.org/10.1016/j.genhosppsych.2010.03.006

Lai, M.-C., Kassee, C., Besney, R., Bonato, S., Hull, L., Mandy, W., Szatmari, P., & Ameis, S. H. (2019). Prevalence of co-occurring mental health diagnoses in the autism population: A systematic review and meta-analysis. *The Lancet Psychiatry*, *6*(10), 819–829. https://doi.org/10.1016/s2215-0366(19)30289-5

Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: Sensitivity and specificity. *Continuing Education in Anaesthesia Critical Care & Pain*, *8*(6), 221–223. https://doi.org/10.1093/bjaceaccp/mkn041

Lever, A. G., & Geurts, H. M. (2016). Psychiatric co-occurring symptoms and disorders in young, middle-aged, and older adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *46*(6), 1916–1930. https://doi.org/10.1007/s10803-016-2722-8

Licence, L., Oliver, C., Moss, J., & Richards, C. (2019). Prevalence and risk-markers of self-harm in autistic children and adults. *Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s10803-019-04260-1

Limoges, É., Mottron, L., Bolduc, C., Berthiaume, C., & Godbout, R. (2005). Atypical sleep architecture and the autism phenotype. *Brain*, *128*(5), 1049–1061. https://doi.org/10.1093/brain/awh425

Loomes, R., Hull, L., & Mandy, W. P. L. (2017). What Is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. *Journal of the American Academy of Child and Adolescent Psychiatry*, *56*(6), 466–474. https://doi.org/10.1016/j.jaac.2017.03.013

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K. O., & Bishop, S. (2012). *Autism Diagnostic Observation Schedule, Second Edition (ADOS-2)*. Western Psychological Services.

Maddox, B. B., & White, S. W. (2015). Comorbid social anxiety disorder in adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *45*(12), 3949–3960. https://doi.org/10.1007/s10803-015-2531-5

Mason, D., Mackintosh, J., McConachie, H., Rodgers, J., Finch, T., & Parr, J. R. (2019). Quality of life for older autistic people: The impact of mental health difficulties. *Research in Autism Spectrum Disorders*, *63*, 13–22. https://doi.org/10.1016/j.rasd.2019.02.007

Maydeu-Olivares, A., & Joe, H. (2014). Assessing Approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*(4), 305 – 328. https://doi.org/10.1080/00273171.2014.911075

McConachie, H., Mason, D., Parr, J. R., Garland, D., Wilson, C., & Rodgers, J. (2018). Enhancing the validity of a quality of life measure for autistic people. *Journal of Autism and Developmental Disorders*, *48*(5), 1596–1611. https://doi.org/10.1007/s10803-017-3402-z

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *The Journal of Applied Psychology*, *95*(4), 728–743. https://doi.org/10.1037/a0018966

Mitchell, G. E., & Locke, K. D. (2015). Lay beliefs about autism spectrum disorder among the general public and childcare providers. *Autism*, *19*(5), 553–561. https://doi.org/10.1177/1362361314533839

Monroe, S., & Cai, L. (2015). Evaluating structural equation models for categorical outcomes: A new test statistic and a practical challenge of interpretation. *Multivariate Behavioral Research*, *50*(6), 569–583. https://doi.org/10.1080/00273171.2015.1032398

Moseley, R. L., Gregory, N. J., Smith, P., Allison, C., & Baron-Cohen, S. (2019). A 'choice', an 'addiction', a way 'out of the lost': Exploring self-injury in autistic people without intellectual disability. *Molecular Autism*, *10*(1), 339. https://doi.org/10.1186/s13229-019-0267-3

Moss, P., Howlin, P., Savage, S., Bolton, P., & Rutter, M. (2015). Self and informant reports of mental health difficulties among adults with autism findings from a long-term follow-up study. *Autism*, *19*(7), 832–841. https://doi.org/10.1177/1362361315585916

Nah, Y.-H., Brewer, N., Young, R. L., & Flower, R. (2018). Brief report: Screening adults with autism spectrum disorder for anxiety and depression. *Journal of Autism and Developmental Disorders*, *48*(5), 1841–1846. https://doi.org/10.1007/s10803-017-3427-3

Newman, L., Wagner, M., Knokey, A.-M., Marder, C., Nagle, K., Shaver, D., & Wei, X. (2011). *The post-high school outcomes of young adults with disabilities Up to 8 years after high School: A report from the national longitudinal transition study-2 (NLTS2)* (NCSER 2011-3005). SRI International.

Nolen-Hoeksema, S. (1987). Sex differences in unipolar depression: Evidence and theory. *Psychological Bulletin*, *101*(2), 259–282. https://doi.org/10.1037/0033-2909.101.2.259

Nylander, L., Axmon, A., Björne, P., Ahlström, G., & Gillberg, C. (2018). Older adults with autism spectrum disorders in Sweden: A register study of diagnoses, psychiatric care utilization and psychotropic medication of 601 individuals. *Journal of Autism and Developmental Disorders*, *48*(9), 3076–3085. https://doi.org/10.1007/s10803-018-3567-0

Ophir, Y., Sisso, I., Asterhan, C. S. C., Tikochinski, R., & Reichart, R. (2020). The Turker blues: Hidden factors behind increased depression rates among Amazon's Mechanical Turkers. *Clinical Psychological Science*, *8*(1), 65–83. https://doi.org/10.1177/2167702619865973

Perera, H. N., Izadikhah, Z., O'Connor, P., & McIlveen, P. (2018). Resolving dimensionality problems with WHOQOL-BREF item responses. *Assessment*, *25*(8), 1014–1025. https://doi.org/10.1177/1073191116678925

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples. *Value in Health*, *18*(1), 25–34. https://doi.org/10.1016/j.jval.2014.10.005

Pezzimenti, F., Han, G. T., Vasa, R. A., & Gotham, K. O. (2019). Depression in youth with autism spectrum disorder. *Child and Adolescent Psychiatric Clinics of North America*, *28*(3), 397–409. https://doi.org/10.1016/j.chc.2019.02.009

Powell, T., & Acker, L. (2014). Adults' experience of an Asperger syndrome diagnosis. *Focus on Autism and Other Developmental Disabilities, 31*(1), 72–80. https://doi.org/10.1177/1088357615588516

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.0.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment*, *81*(2), 93–103. https://doi.org/10.1207/S15327752JPA8102_01

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). PROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3), 223–237. https://doi.org/10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

Russell, A., Cooper, K., Barton, S., Ensum, I., Gaunt, D., Horwood, J., Ingham, B., Kessler, D., Metcalfe, C., Parr, J., Rai, D., & Wiles, N. (2017). Protocol for a feasibility study and randomised pilot trial of a low-intensity psychological intervention for depression in adults with autism: The Autism Depression Trial (ADEPT). *BMJ Open*, *7*(12), e019545. https://doi.org/10.1136/bmjopen-2017-019545

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(1), 1–97. https://doi.org/10.1007/bf03372160

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry*, *59*(Suppl 20), 22–33.

Skylark, W. J., & Baron-Cohen, S. (2017). Initial evidence that non-clinical autistic traits are associated with lower income. *Molecular Autism*, *8*(1), 61. https://doi.org/10.1186/s13229-017-0179-z

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

Stover, A. M., McLeod, L. D., Langer, M. M., Chen, W.-H., & Reeve, B. B. (2019). State of the psychometric methods: Patient-reported outcome measure development and refinement using item response theory. *Journal of Patient-Reported Outcomes*, *3*(1), 50. https://doi.org/10.1186/s41687-019-0130-5

Supekar, K., Iyer, T., & Menon, V. (2017). The influence of sex and age on prevalence rates of comorbid conditions in autism. *Autism Research*, *10*(5), 778–789. https://doi.org/10.1002/aur.1741

The WHOQOL Group. (1998). Development of the World Health Organization WHOQOL-BREF quality of life assessment. *Psychological Medicine*, *28*(3), 551–558. https://doi.org/10.1017/s0033291798006667

Thomas, M. L. (2019). Advances in applications of item response theory to clinical assessment. *Psychological Assessment*, *31*(12), 1442–1455. https://doi.org/10.1037/pas0000597

Toland, M. D., Sulis, I., Giambona, F., Porcu, M., & Campbell, J. M. (2017). Introduction to bifactor polytomous item response theory analysis. *Journal of School Psychology*, *60*, 41–63. https://doi.org/10.1016/j.jsp.2016.11.001

Uljarević, M., Hedley, D., Rose-Foley, K., Magiati, I., Cai, R. Y., Dissanayake, C., Richdale, A., & Trollor, J. (2019). Anxiety and depression from adolescence to old age in autism spectrum disorder. *Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s10803-019-04084-z

Uljarević, M., Richdale, A. L., McConachie, H., Hedley, D., Cai, R. Y., Merrick, H., Parr, J. R., & Couteur, A. L. (2018). The Hospital Anxiety and Depression scale: Factor structure and psychometric properties in older adolescents and young adults with autism spectrum disorder. *Autism Research*, *11*(2), 258–269. https://doi.org/10.1002/aur.1872

Underwood, J. F. G., Kendall, K. M., Berrett, J., Lewis, C., Anney, R., van den Bree, M. B. M., & Hall, J. (2019). Autism spectrum disorder diagnosis in adults: Phenotype and genotype findings from a clinically derived cohort. *British Journal of Psychiatry*, *215*(5), 647–653. https://doi.org/10.1192/bjp.2019.30

Unruh, K. E., Bodfish, J. W., & Gotham, K. O. (2018). Adults with autism and adults with depression show similar attentional biases to social-affective images. *Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s10803-018-3627-5

Vohra, R., Madhavan, S., & Sambamoorthi, U. (2017). Comorbidity prevalence, healthcare utilization, and expenditures of Medicaid enrolled adults with autism spectrum disorders. *Autism*, *21*(8), 995–1009. https://doi.org/10.1177/1362361316665222

Wang, Y.-P., & Gorenstein, C. (2013). Psychometric properties of the Beck Depression Inventory-II: A comprehensive review. *Brazilian Journal of Psychiatry*, *35*(4), 416–431. https://doi.org/10.1590/1516-4446-2012-1048

Wentz, E., Nydén, A., & Krevers, B. (2012). Development of an internet-based support and coaching model for adolescents and young adults with ADHD and autism spectrum disorders: A pilot study. *European Child & Adolescent Psychiatry*, *21*(11), 611 – 622. https://doi.org/10.1007/s00787-012-0297-2

Williams, Z. J. (2020). *irt_extra: Additional functions to supplement the mirt R package* [R]. https://doi.org/10.13140/RG.2.2.10226.04803/1

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3

Youngstrom, E. A. (2014). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology*, *39*(2), 204–221. https://doi.org/10.1093/jpepsy/jst062

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ωH: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133. https://doi.org/10.1007/s11336-003-0974-7

Figure 1.

Receiver Operating Characteristic Curves for BDI-II Total and IRT scores in Adults with ASD

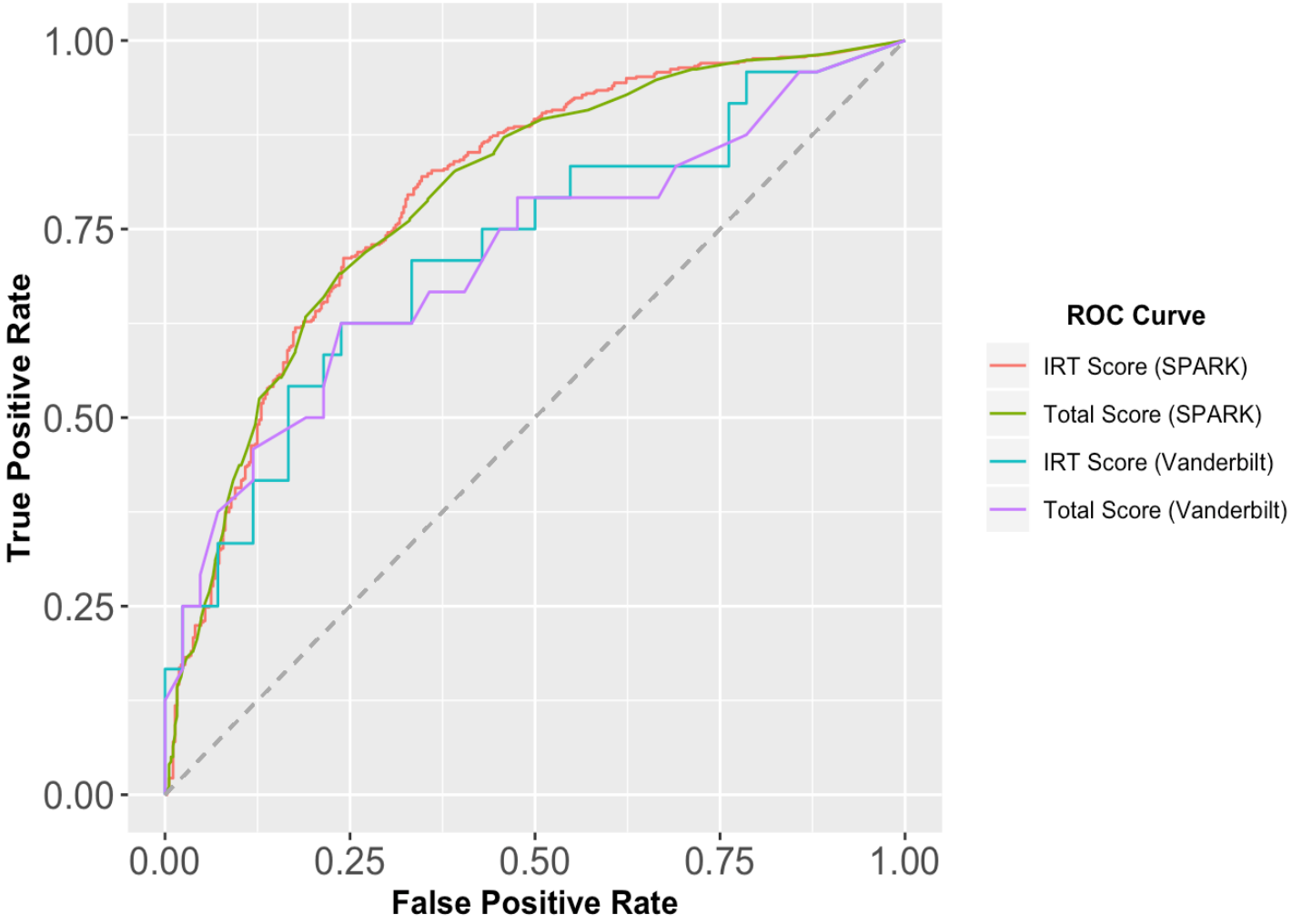

Table 1.

Participant demographics and BDI-II scores

	SPARK (ASD)	MTurk (TD)	Vanderbilt (ASD)	Vanderbilt (TD)
Total *N*	881	986	66	116
Age in Years (*M* [*SD*])	30.94 (7.10)	32.60 (6.85)	24.09 (5.60)	27.83 (6.75)
Non-Hispanic White (*N* [%])	693 (78.7%)	735 (74.5%)	57 (86.4%)	83 (71.5%)
Gender (*N* [%])				
Male	332 (37.7%)	368 (37.3%)	37 (56.1%)	38 (32.8%)
Female	466 (52.9%)	616 (62.5%)	26 (39.4%)	76 (65.5%)
Other/Non-binary	81 (9.2%)	2 (0.002%)	3 (4.5%)	2 (1.7%)
Education (*N* [%])				
Less than High School	4 (0.5%)	8 (0.8%)	2 (3.0%)	0 (0%)
High School Diploma[a]	223 (25.3%)	142 (14.4%)	15 (22.7%)	3 (2.6%)
Some College	233 (26.4%)	204 (20.7%)	16 (24.2%)	18 (15.5%)
2-year College Degree	88 (10.0%)	133 (13.5%)	6 (9.1%)	5 (4.3%)
4-year College Degree	198 (22.5%)	370 (37.5%)	21 (31.8%)	49 (42.2%)
Graduate/Professional Degree	112 (12.7%)	129 (13.1%)	4 (6.1%)	41 (35.3%)
BDI-II				
Total Score (*M* [*SD*])	17.48 (12.98)	15.75 (13.29)	13.20 (10.34)	13.90 (13.19)
Above Clinical Cutoff (*N* [%])[b]	492 (55.8%)	489 (49.6%)	29 (43.9%)	53 (45.7%)

Note. Samples included (a) 881 adults with ASD recruited from the Simons Foundation SPARK cohort (SPARK), (b) 986 general population adults recruited through Amazon's Mechanical Turk (MTurk), (c) 182 adults (66 with ASD) recruited through laboratory experiments performed at Vanderbilt University Medical Center (Vanderbilt)

[a] Includes individuals who received a GED or completed trade school/vocational programs that granted certificates/licenses but no degree.

[b] Based on BDI-II total score of 14 or greater; missing items imputed using mean of remaining items.

Table 2.

Bifactor Loadings and Model-based Statistics for Combined ASD Group

Item	Endorsed[a]	λ_G	λ_{CA}	λ_{SV}	h^2	*I-ECV*
1. Sadness	52.9%	0.79	0.23	—	0.67	0.92
2. Pessimism	58.0%	0.69	0.33	—	0.59	0.81
3. Past Failure	65.4%	0.70	0.44	—	0.68	0.71
4. Loss of Pleasure	56.2%	0.85	—	-0.05	0.73	>0.99
5. Guilty Feelings	55.6%	0.65	0.41	—	0.58	0.72
6. Punishment Feelings	36.8%	0.56	0.39	—	0.47	0.68
7. Self-Dislike	52.6%	0.74	0.48	—	0.78	0.71
8. Self-Criticalness	56.8%	0.67	0.46	—	0.66	0.68
9. Suicidal Thoughts or Wishes	36.5%	0.68	0.26	—	0.53	0.87
10. Crying	35.0%	0.65	—	0.03	0.42	>0.99
11. Agitation	50.4%	0.71	—	—	0.51	>0.99
12. Loss of Interest	53.8%	0.87	—	-0.02	0.76	>0.99
13. Indecisiveness	50.1%	0.71	0.08	—	0.51	0.99
14. Worthlessness	47.7%	0.76	0.48	—	0.81	0.72
15. Loss of Energy	69.6%	0.79	—	0.47	0.84	0.74
16. Changes in Sleeping Pattern	68.4%	0.64	—	0.36	0.54	0.76
17. Irritability	47.2%	0.75	—	0.06	0.56	0.99
18. Changes in Appetite	53.6%	0.61	—	0.18	0.40	0.92
19. Concentration Difficulty	56.4%	0.74	—	0.17	0.58	0.95
20. Tiredness or Fatigue	67.1%	0.77	—	0.54	0.89	0.68
21. Loss of Interest in Sex	34.7%	0.58	—	0.13	0.35	0.95
		G	**CA**	**SV**		
	ω_t/ω_S	0.952	0.913	0.916	*ECV* = 0.834	
	ω_H/ω_{HS}	0.881	0.180	0.048	*PUC* = 52.38%	

Note. Loadings and model-based statistical indices are derived from a full-information maximum likelihood confirmatory factor analysis. The equivalent graded response model parameters can be found in supplemental table S2. G = general factor. CA = cognitive-affective factor; SV = somatic-vegetative factor; h^2 = communality; (*I*-)*ECV* = (Item-level) explained common variance*; PUC* = percentage of uncontaminated correlations.

[a] The percentage of respondents with a score of "1" or greater on a given item

Table 3.

Differential Item Functioning Results Comparing ASD and TD Groups

	$\chi 2(4)$	*p*-value	UIDS	ESSD	Parameters[a]
1. Sadness	21.59	< 0.001	0.056	0.078	—
2. Pessimism	17.10	0.003	0.030	-0.009	***d1***, *d2*, *d3*
3. Past Failure	25.46	< 0.001	0.130	-0.159	***a1***, *d2*, *d3*
6. Punishment Feelings	24.77	< 0.001	0.122	-0.192	***a1***, *d1*, *d2*, *d3*
7. Self-Dislike	13.11	0.011	0.019	-0.020	*d3*
8. Self-Criticalness	38.91	< 0.001	0.064	0.065	***d1***, *d3*
9. Suicidal Thoughts or Wishes	39.33	< 0.001	0.093	-0.220*	*d1*
10. Crying	18.84	0.002	0.042	-0.013	*d3*
11. Agitation	13.29	0.011	0.076	0.138	***d1***
12. Loss of Interest	15.84	0.004	0.072	0.013	***d1***
13. Indecisiveness	54.73	< 0.001	0.126	-0.183	*d2*, *d3*
14. Worthlessness	15.62	0.004	0.051	-0.041	***a1***, *d2*
15. Loss of Energy	17.64	0.002	0.112	-0.141	*d1*
17. Irritability	29.89	< 0.001	0.130	0.219*	***d1***, ***d2***
18. Changes in Appetite	16.94	0.003	0.101	-0.182	*d3*
19. Concentration Difficulty	38.73	< 0.001	0.133	-0.205*	*d2*
20. Tiredness or Fatigue	12.36	0.015	0.096	-0.106	—
21. Loss of Interest in Sex	25.15	< 0.001	0.117	0.233*	***d1***
Differential Test Functioning:	UETSDS = 0.524		ETSSD = -0.039		
Multi-group Model Fit:	$C_2(349) = 1241.4$		$CFI_{C2} = 0.990$		$RMSEA_{C2} = 0.036$

Note. Results indicate omnibus Wald DIF tests using the iterative anchor-selection method of Cao et al., (2017). *p*-values are corrected for a 5% false discovery rate. Parameters that were significantly different between groups when tested alone with follow-up Wald tests (FDR < 0.05) are indicated in the Parameters column. UIDS = Unsigned Expected Item Score Difference in the Sample; ESSD = Expected Score Standard Deviation (in Cohen's *d* metric); *a1* = general factor slope parameter; *d1–d3* = item intercept parameters; UETSDS = Unsigned Expected Test Score Difference in the Sample; ETSSD = Expected Test Score Standardized Difference (in Cohen's *d* metric).

[a] Parameters in bold are larger (i.e., more discriminating for *a* parameters and "easier" for *d* parameters) in the ASD group. Larger values of *a* indicate that the item is more strongly related to the latent trait in the ASD group, whereas larger values of *d* indicate that a given item response is endorsed at lower latent trait levels in the ASD group than the TD group.

* Practically significant DIF (i.e., |ESSD| > 0.2)

Table 4.

Diagnostic Efficiency Statistics of BDI-II Scores in Adults with ASD

	SPARK Sample (N_{ASD} = 868, N_{DEP} = 499)		Vanderbilt Sample (N_{ASD} = 66, N_{DEP} = 24)	
	IRT Score	**Total Score**	**IRT Score**	**Total Score**
AUC	0.796 [0.763, 0.826]	0.791 [0.759, 0.821]	0.718 [0.577, 0.849]	0.711 [0.572, 0.839]
Sensitivity	0.820 [0.786, 0.854]	0.743 [0.703, 0.782]	0.750 [0.583, 0.917]	0.625 [0.417, 0.792]
Specificity	0.653 [0.604, 0.699]	0.694 [0.648, 0.740]	0.571 [0.429, 0.714]	0.667 [0.524, 0.810]
LR+	2.363 [2.065, 2.751]	2.428 [2.084, 2.887]	1.750 [1.167, 2.800]	1.875 [1.105, 3.500]
LR-	0.276 [0.223, 0.333]	0.370 [0.311, 0.433]	0.438 [0.146, 0.824]	0.562 [0.280, 0.917]
PPV_{Sample}	0.762 [0.736, 0.788]	0.767 [0.738, 0.796]	0.500 [0.400, 0.615]	0.517 [0.387, 0.667]
NPV_{Sample}	0.728 [0.689, 0.768]	0.667 [0.631, 0.704]	0.800 [0.680, 0.923]	0.757 [0.656, 0.862]
PPV_{Pop}	0.414 [0.382, 0.451]	0.420 [0.384, 0.463]	0.343 [0.258, 0.455]	0.359 [0.248, 0.511]
NPV_{Pop}	0.924 [0.909, 0.938]	0.901 [0.886, 0.915]	0.884 [0.803, 0.958]	0.856 [0.785, 0.923]

Note. Statistics are presented with 95% bootstrapped confidence intervals. Values are based upon diagnostic cutoffs of -0.0893 for BDI-II IRT (latent trait) scores and 14 for BDI-II total scores. SPARK = Simons Powering Autism Research Knowledge; N_{ASD} = number of individuals with ASD and diagnostic outcome data in the sample; N_{DEP} = number of individuals with ASD who are diagnosed with a current depressive disorder (self-reported in SPARK sample, based on SCID-5 or MINI algorithm in Vanderbilt Sample); *AUC* = area under the receiver operating characteristic curve; LR+ = positive likelihood ratio; LR- = negative likelihood ratio; PPV_{Sample} = positive predictive value based on the prevalence of depression in the given sample; NPV_{Sample} = negative predictive value based on the prevalence of depression in the given sample; PPV_{Pop} = positive predictive value based on the estimated prevalence of current depression in adults with ASD (23%; Hollocks et al., 2019); NPV_{Pop} = negative predictive value based on the estimated prevalence of current depression in adults with ASD.