7-10-2017 2:00 PM

Development and Psychometric Validation of the interRAI Internalizing and Externalizing Subscales

Chloe Lau, *The University of Western Ontario*

Supervisor: Dr. Donald Saklofske, *The University of Western Ontario*
Joint Supervisor: Dr. Shannon Stewart, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Psychology

ABSTRACT

The purpose of this study was to develop and assess the validity of two brief measures drawing from the interRAI Child and Youth Mental Health (ChYMH) for screening internalizing and externalizing mental health. Data were collected from children/youths 4 to 18 years of age (*N*=3464). First, items relevant to internalizing and externalizing disturbances underwent expert content validation. Second, unrestricted factor analyses and item response theory parameterizations were conducted to test the validity of the measurement model. Finally, concurrent validity of these two measures were confirmed based on relationships with other established subscales from criterion measures (e.g., Child Behaviour Checklist). Two separate, 12-item scales were developed based on item representativeness (i.e., S-CVI/UA>0.80) and factor analyses. The final subscales showed high internal consistency and correlated strongly with the appropriate criterion measures. The development of two psychometrically sound brief scale subscales will provide useful information for triaging and prioritizing referrals to appropriate services for children/youths.

KEYWORDS: interRAI; internalizing; externalizing; mental health; children; psychometric

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my research advisors, Dr. Donald H. Saklofske and Dr. Shannon L. Stewart, for providing feedback and support throughout my graduate studies. I wanted to express sincere gratitude for Dr. Saklofske and Dr. Stewart for providing me not only the guidance, but opportunity of working on such an impactful project. I would also like to thank the panel of experts who took time out of their busy days to assist with content validation.

I also wanted to acknowledge my master's thesis committee members Dr. Leora Swartzman, Dr. Richard Goffin, and Dr. Claire Crooks. Your comments and recommendations were not only extremely helpful for this project, but generalizable and applicable for my future academic endeavors. Thank you to Dr. Paul F. Tremblay and Dr. Richard W.J. Neufeld for their advice and role modelling their passion for application of advanced statistics and mathematical models in psychology. I would also like to thank my incredible supervisors and mentors Dr. Randi McCabe, Dr. Jennifer Ford, Dr. Michael Van Ameringen, and Ms. Beth Patterson for the learning opportunities and continuous mentorship for the past three years. It was with your support and guidance that I had the opportunity to begin my graduate studies at this reputable clinical psychology graduate program at Western University.

I would like to thank my mom Isabella, dad Dennis, aunt Rodensa, uncle Leslie, and aunt Wendy for your endless support throughout this process. My aunt Rodensa has fought the most courageous battle with cancer and will always remain an inspiration for me to persevere despite the difficult circumstances. Thank you to my wonderful friends Victoria Vass, Adrienne Borrie, Vanessa Sinclair, Callista Forchuk, and Anita Feher for being the most wonderful support system during the difficult times throughout the past two years. Victoria was a very special friend throughout the difficult times in transitioning from undergraduate to graduate studies. I will always remember the kindness my family and friends have shown me.

The principal study from which the data for this thesis were obtained was supported by an Operating Grant awarded to Dr. Shannon L. Stewart. This work done for this thesis was supported by the Joseph Bombardier Canadian Graduate Scholarships Master's Award (2016-2017), the Western Graduate Research Scholarship (2015-2017), and the Faculty of Social Sciences Graduate Research Awards Fund (2016-2017) awarded to Chloe Lau.

TABLE OF CONTENTS

CHAPTER ONE: INTRODUCTION AND LITERATURE REVIEW

1.1 Mental Health Care for Children

In Canada, approximately 10 to 20% of children/youth are currently suffering from mental illness, and it is one of the most disabling group of disorders in the world (Canadian Mental Health Association, 2017). Most contemporary research and clinical settings utilizes criteria from the *Diagnostic and Statistical Manual of Mental Disorders* (DSM) as the gold-standard for operationalizing disordered mental health (Berenbaum, 2013; American Psychiatric Association, 2013). Using diagnostic criteria in the DSM-Fourth Edition (DSM-IV), Costello, Mustillo, Erkanli, Keeler, and Angold (2003) showed that the three-month prevalence in children aged 9 to 13 of any DSM-IV is 2.4%, 2.2% and 7.0 % for anxiety disorders, depressive disorders, and behavioural disorders respectively. Using these prevalence values, the researchers predicted that the expected cumulative prevalence of psychiatric disorders by age 16 for any anxiety, depressive, and behavioural disorder is 9.9%, 9.5%, and 23% respectively (Costello et al., 2003).

The importance of clinical identification of mental health disorders, as early as preschool, is emphasized in research demonstrating that childhood mental health disorders showed chronicity and recurrence (Luby, Si, Belden, Tandon, & Spitznagel, 2009). Bufferd, Dougherty, Carlson, Rose, & Klein (2012) conducted the first study of its kind utilizing a large, longitudinal representative sample and objective diagnostically-valid tools. Results of the study suggested about a quarter of children, as young as those in pre-school years, had diagnosed mental disorders in the previous 3 months at ages 3 and 6 (Bufferd et al., 2012). Not only did earlier diagnosis in pre-school children predict a later mental health diagnosis by 5-fold, both homotypic and heterotypic continuity of mental disorders were evident in these young children (Bufferd et al., 2012). Within internalizing disorders, there were significant heterotypic continuity from depression to anxiety and anxiety to depression. Within externalizing disorders, attention hyperactive deficit disorder predicted later oppositional defiant disorder and conduct disorder predicted later substance use (Bufferd et al., 2012). Given the chronicity and early vulnerability to mental illness to different disorders across the life span, it is important to

assess for childhood mental health problems and illnesses early in life (Luby et al., 2009; Griffith, 2010; Government of Canada, 2006).

With many mental health disorders onsetting in early childhood and adolescence (e.g., Kessler et al., 2005), an accurate and seamless assessment of childhood psychiatric symptoms that allows the continuity of care is imperative to decrease burden on children/youth, families, and the economy (Government of Canada, 2006). Despite that only 1 in 5 children who need mental health care get access, waitlists for assessment and intervention can span from as long as six months to a year (Children's Mental Health Ontario, 2016; Office of the Auditor General Ontario, 2016). To increase efficiency, the public sector needs a low cost, easily transferable, and accessible system where needs and preferences of children and families can be appropriately identified and prescribed (Hébert et al., 2003). Such a system should also provide efficient care transitions to reduce redundant assessment on patients who received repeated intake across service providers (Hébert et al., 2003). Thus, it is important to invest in mental health initiatives that could provide assessment for much-needed access and intervention early in the life span for long-lasting impact (Heckman, Gray, & Hirdes, 2013).

1.2 The InterRAI Assessment System

The international Resident Assessment Instrument (interRAI) was developed by a non-profit international organization network of 50 members from over 30 countries who intended to create standardization of items across assessment tools in different sectors (Bernabei, Landi, Onder, Liperoti, & Gambassi, 2008). The interRAI assessment instruments are used along the service continuum (e.g., homecare, palliative care, emergency department) designed to assess and monitor symptoms and treatment outcomes across age groups and vulnerable populations (Hirdes et al., 2003). Items in every interRAI instrument were developed for comprehensive and accurate assessment through extensive literature review of relevant symptoms and construct and consultation with expert researchers and clinicians (Hirdes et al., 2011). New items and scales have been developed through detailed reviews and clinical expertise with expert working groups and collaborations (Hirdes et al., 2002; Hirdes et al., 1999; Hirdes et al., 2011). Items with country-specific content are flagged within the instruments, such that the instrument is designed to evaluate the needs of patients in different care settings

(Heckman, 2013; Hirdes et al., 1999). There are more than 20 instruments built with a core set of items relevant for all health care sectors with identical definitions, time frame observation period, and method of scoring (Hirdes et al., 2011). This compatibility of assessment allows for decreased assessment burden and outcome tracking for smooth transitioning across multiple treatment settings (Stewart et al., 2015). While the collected data allows every agency to conduct quality assurance to warrant accuracy of assessments, data could also be compared across agencies thereby permitting population comparisons and resource allocations across programs (Stewart et al., 2015).

Multiple reliability and validity studies have demonstrated strong psychometric properties and criterion validity for interRAI instruments in adult and geriatric samples (e.g., Martin, Hirdes, Fries, & Smith, 2007; Morris, Carpenter, Berg, & Jones, 2000) and in children/youth samples (e.g., Stewart & Hamza, 2017; Lau, Stewart, Saklofske, Tremblay, & Hirdes, 2017). Five of the interRAI family of instruments have demonstrated substantial reliability, with an overall kappa mean of 0.75 that evaluated over 160 shared items in 2 or more instruments, and more than 60% of items on the interRAI family suites had a kappa mean above 0.70 (Hirdes et al., 2008). Multiple validation studies on interRAI measures have been released since, including but not limited to the aggressive behaviour scale, pressure ulcer risk scale, cognitive performance scale, pain scale, delirium screener, and depressive symptoms scale (see Perlman & Hirdes, 2008; Poss et al., 2010; Travers, Byrne, Pachana, Klein, & Gray, 2013; Fries, Simon, Morris, Flodstrom, & Bookstein, 2001; Salih, Klein, Lakhan, & Gray, 2012 respectively for the aforementioned studies).

Strong reliability and validity was also found across different languages for the Chinese version of interRAI Mental Health (e.g., Chan, Lai, Chi, 2017), the Korean version (e.g., Kim et al., 2015), and the Hong Kong version of interRAI (e.g., Liu, Chi, Chan, Lai, & Leung, 2015). InterRAI can also be used for large-scale research studies, such that the large dataset of diverse range of variables and large, multi-site sample going through standardized assessment can be used to answer certain research questions, and appropriate confounding variables can be generated and controlled for (e.g., Foebel et al., 2013; Nishtala & Jamieson, 2017; Ribbe, Jonsson, & Bernabei, 2012). Overall, interRAI systems have the potential to promote better communication between researchers and

practitioners across service sectors and nations to promote a seamless transition across sectors in healthcare. The large dataset collected using the standardized assessment also allows for health outcome tracking and clinical care research at a local and population level.

1.3 InterRAI Child and Youth Mental Health Suite

In recent years, the mental health care sector has largely adopted the use of interRAI measures for mental health needs, including the incorporation of the Resident Assessment Instrument-Mental Health (RAI-MH; Hirdes et al., 2000; Hirdes et al., 2011) and interRAI Child and Youth Mental Health (ChYMH; Stewart et al., 2015) as part of standardized assessment. The interRAI ChYMH was developed for efficient and effective care-planning, as well as promoting streamlined access to services across service sectors for vulnerable populations (Stewart et al., 2015). Assessment is conducted through communicating with the child/youth and primary caregiver, observation, communication with healthcare providers, and review of medical records (Stewart et al., 2015). The interRAI Child and Youth Mental Health (ChYMH) consists of over 400 items assessing psychiatric, functional, medical, and social issues (Stewart et al., 2015). The tool has a shared set of items allowing informational transfer with other child/youth interRAI instruments, such as ChYMH-Education (ChYMH-EDU), ChYMH-Developmental Disabilities (ChYMH-DD), the Youth Justice Custodial Facilities instrument (YJCF), Homecare Pediatrics, and the adult version interRAI-Mental Health (Hirdes et al., 2011).

The interRAI ChYMH assessment is based on a semi-structured interview format, such that trained assessors complete the instrument using all sources of information, including contact with the child/youth and the family and their child or youth, as well as other service providers and records (Stewart et al., 2015). The form solicits information on history (e.g., living situation), mental state indicators (e.g., responsiveness and adherence to treatment regimens), and indicators of behaviour (e.g., behaviours monitored during observation by the clinician). Subsequent sections assess for indicators of emotional distress (e.g., mood disturbance, anxiety), behavioural disturbance (e.g., hyperactivity), substance use or excessive behaviour (e.g., drugs and alcohol, problematic video gaming), harm to self and others (e.g., indicators of violence), and other relevant indicators (e.g., psychosis, sleep problems). It employs a specific assessment period of

three days in order to provide reliable and valid measures of symptomatology. Like other measures in the interRAI suites, the ChYMH has multiple subscales and collaborative action plans embedded in the instrument to define measures of status or functioning. Of note, there is an inpatient version for children/youth placed in a residential or psychiatric facility and the community-based form for community referrals. These forms are compatible and coordinate seamless service provision for longitudinal tracking of clinical status across the continuum of care. An adolescent supplement is provided for part of standardized assessment for age-appropriate risk behaviours, function, mental and physical health, social support, patterns of service use (Stewart et al., 2015). Data collected from the interRAI ChYMH has been used in publications of clinically relevant findings (e.g., Baiden, Stewart, & Fallon, 2017a, Baiden, Stewart, & Fallon, 2017b; Armiento, Hamza, Stewart, & Leschied, 2016).

Rather than using diagnostic categories, the ChYMH conceptualizes mental health conditions using a dimensional and holistic approach, in which the clinician understands the child's broader individual and environmental context for a symptomatically coherent pattern. At the individual level, the interRAI Child and Youth suite promotes streamlining of the assessment process for access to local services. For service settings, the interRAI ChYMH provides information to assist with triaging, wait-list management, program evaluation, and quality improvement (Stewart et al., 2015).

In the manual, each item holds explicit information regarding intent of items, supplementary definition of the item, explicit instructions for assessment, and coding of response (Stewart et al., 2015). Under intent, the focus of the item in identifying the problem and relations to care-planning is provided. The definition for each item is provided for universal interpretations of key terms used. Each item also includes the process of assessment, which includes sources of information and methods to determine the correct response from a variety of the item sources, whether it is through interview, observation, discussion with child/youth, members of care planning team, and/or review of any clinical documents. The coding instructions provide standardized procedures for proper codes to score and report each of the responses. Publications by the interRAI authors further detail the format and utility of instrument (Stewart and Hirdes, 2015;

Stewart et al., 2015a; Stewart et al., 2015b), and is also available from interRAI's website (The InterRAI Organization, 2017).

1.4 Childhood Internalizing and Externalizing Symptoms

Emotional and behavioural dysregulation in childhood and adolescence are commonly categorized into the taxonomy of broadband and narrow-band conceptualization of psychopathology (Achenbach, 2016). Within the broadband psychopathology are internalizing and externalizing disorders, which researchers have characterized as the dichotomy between (1) overcontrolled and undercontrolled, (2) emotional and conduct problems, or (3) inhibition and aggression (e.g., Caspi et al. 1995; Southam-Gerow & Kendall 2002; Achenbach. 1995). Many contemporary childhood mental health measures, such as the Social Skills Improvement System, interpret competing problem behaviours as externalizing (e.g., noncompliance, aggression, or coercive behaviors) or internalizing behavior patterns (e.g., social withdrawal, anxiety, or depression; Gresham, Elliot, & Kettler, 2010; Gresham & Elliott, 1990, 2008). Epidemiological research suggested that stability in developmental psychopathology can be accounted for by broad internalizing and externalizing domains (e.g., Mesman & Koot, 2000). It is imperative to assess for both internalizing and externalizing conditions, as youth with mixed symptom profiles (i.e., both internalizing and externalizing behaviours) have maladaptive developmental outcomes, and may need longer-term treatments and greater monitoring of treatment responses than those with only one of the symptom profiles (Capaldi and Stoolmiller, 1999).

Internalizing problems are characterized by negative affectivity and inner-directed intensity of emotions (Robins et al., 1996; Asendorpf & van Aken, 1999; Eisenberg et al., 2001). The internalizing spectrum has been broadly conceptualized as symptoms related to anxiety, depression, and psychosomatic complaints, along with behaviours related to social withdrawal (Blumberg & Izard, 1985; Keltner et al., 1995; Eisenberg et al., 2001). Homotypic continuity is most prominent at the broader internalizing level, and less stable in specific, narrow-band disorders, even in children as young as 2 to 3 years of age (Mesman & Koot, 2000). High levels of behavioural inhibition in children are predictive of development of internalizing symptoms later in life, including phobias and anxiety (e.g., Kagan, Snidman, Zentner, & Peterson, 1999). In the Diagnostic and Statistical

Manual-5th Edition (DSM-5; American Psychiatric Association, 2013), internalizing disorders are largely captured through diagnostic categories in depression and anxiety. Depression is characterized as low mood and/or melancholic symptoms (i.e., lack of motivation, anhedonia). Although it was previously believed that young children experienced atypical symptoms of "masked depression" (e.g., Lesse, 1983), Luby et al. (2003) found that typical symptoms are the most sensitive clinical markers of depression, even amongst young children. Traditional markers of depression found in adults were most frequently endorsed in depressed children compared to healthy and externalizing psychiatric comparison groups (Luby et al., 2003). Although irritability and sadness assessed jointly were the most sensitive symptoms of early depression, anhedonia was found to be the most specific symptom for the depressed group of young children (Luby et al., 2003).

Externalizing problems is characterized by outer-directed distress and "undercontrolled" regulation, such that inhibition of controlling cognitive processing and attention span is evident (e.g., Olson, Schilling, & Bates, 1999; Oosterlaan & Sergeant, 1996; Rothbart, Posner, & Hershey, 1995; Eisenberg et al. 2001). With externalizing disorders being outer-directed and undercontrolled, manifestation of symptoms (e.g., anger, temper tantrum) tend to generate discomfort and conflict in the child's surrounding environment (Bornstein, Hahn, & Suwalsky, 2013; Weeks et al., 2016). In the DSM-5, two of the most chronic and debilitating externalizing mental health disorders in the DSM include oppositional defiant disorder (ODD) and conduct disorder (CD), which are categorized as disruptive behaviour disorders (DBDs) in the DSM (American Psychiatric Association, 2013). DBDs are categorized by patterns of ongoing uncooperative, argumentative, and defiant behaviors toward authority figures (American Psychiatric Association, 2013). Symptoms of externalizing disorders tend to cluster, which may be a result of shared temperamental risk factors and the common heritable factor of impulsivity, a trait that accounts for a large amount of covariation between ADHD and DBDs (American Psychiatric Association, 2013).

While trait impulsivity may predispose children/youth to early hyperactive behaviours, there is evidence of increased vulnerability to other externalizing disorders across the life span (Beauchaine, Hinshaw, & Pang, 2010). The diagnosis of ODD is

often comorbid with attention-deficit hyperactivity disorder (ADHD), a neurodevelopmental disorder characterized by early manifestations of more-than-normal activity and inattention (APA, 2013; Costello et al., 2013). Moreover, a great number of boys diagnosed with ODD will progress to conduct disorder in adolescence, which is characterized by repetitive and persistent patterns of antisocial behaviours, violation of age-appropriate norms and rules, destructive behaviour, and deceitful behaviour (Lahey, Waldman, & McBurnett, 1999). Thus, it is important to identify this cluster of externalizing behaviours at an early stage as these disorders impact social, occupational, or academic functioning of the child/youth (Mannuzza, Klein, Bessler, Malloy, & LaPadula, 1993; Weiss, Hechtman, Milroy, & Perlman, 1985). Moreover, children and teens diagnosed with CD are at great risk of continuity to antisocial personality disorder and substance abuse in adolescence and adulthood (Faraone, Biederman, Jetton, & Tsuang, 1997; Faraone, Biederman, Mennin, Russel, & Tsuang, 1998). Early aggression is predictive of criminality, even when controlling for variables such as lower parental education and lower IQ (Huesmann, Eron, & Dubow, 2002). As such, it is important to accurately measure and capture externalizing symptoms for early intervention and prevention.

1.5 Need for the Present Study

In recent years, researchers and clinicians have attempted to move away from single disorder classification as their primary focus of concern and instead, conceptualize the internalizing and externalizing spectrum as a normally distributed, continuous assessment of psychopathology (Forbes, Tackett, Markon, & Krueger, 2016). While individual diagnoses assess different facets related to the broader constructs, shared variance in the assessment of broader levels of dimensional risk are often neglected (Forbes et al., 2016). Conceptualizing internalizing and externalizing psychopathology as a continuous symptom-level phenomenon allows for analysis in both shared and unique variance in the hierarchy (Lahey et al., 2004). Moreover, unlike traditional categorization of disordered presentation as present or absent, this dimensional assessment allows frequency and severity of signs and symptoms of well-validated broadband psychopathology to be adequately captured (Krueger & DeYoung, 2016; Krueger & Eaton, 2012). For example, Krueger et al. (2007) argued that conceptualizing symptoms

of conduct disorder as part of the externalizing spectrum would present greater longitudinal tracking of symptoms as well as greater clinical utility. As such, lead agencies using the interRAI ChYMH assessment requested the incorporation of broadband measures into the instrument to assist in examining placement and treatment progress. These broadband measures, together with the narrowband measures, would provide more useful information for the clinician.

Given that reliability and validity sets psychological measurement tools apart from other clinical information, this aim of this study was to explore the dimensionality, reliability, and validity of the internalizing and externalizing subscales in the interRAI ChYMH (Meyer et al., 2001). The development of a brief, subscale into the interRAI ChYMH scale summary reports would suit the needs of a large network of Canadian mental health treatment facilities, but also provide precise measurement and empirically quantified information on internalizing and externalizing mental health symptoms.

CHAPTER 2: METHODS

2.0 Introduction to the Study

The current study utilizes existing items on the interRAI Child and Youth Mental Health (interRAI ChYMH) to develop two separate brief subscales aimed at capturing a broad spectrum of internalizing and externalizing mental health symptoms in children and adolescence. This chapter delineates the scale development process and addresses the following objectives: (1) establishing content validity through rational expert judgments, (2) an analysis of the underlying factor structure of the internalizing and externalizing mental health subscales, and (3) establishing criterion validity of the subscales through total scale score correlations with existing gold standard child/youth mental health measures. The expert content data was collected to validate the scales in this study, but the dataset used for the latter steps were retrospective analyses of existing datasets. Refer to Figure 1 for a flow diagram of the study design.

To establish measurement adequacy, Lynn (1986) proposed two stages in assessment tool development, which includes (1) the developmental stage (e.g., item generation) and (2) judgement quantification (i.e., at the item and scale level). Given that the goal of the present study is to take existing items from the interRAI ChYMH to develop an internalizing mental health subscale, the first stage included selecting relevant content from the large pool of items embedded in the instrument.

2.1.1 Item Selection for the Internalizing Scale

Prior to expert content validation, the items embedded in the interRAI ChYMH were reviewed by the authors of this study. Each of these items were evaluated on a 5-point ordinal scale (0 = not present to 4= exhibited daily in last 3 days, 3 or more episodes or continuously; Stewart et al., 2015). Items from the existing subscales related to emotional disturbance, including the social disengagement scale, depressive symptoms scale, and anxiety scale, were considered for inclusion in the internalizing mental health subscale:

*Social Disengagement Sca*le. This scale has 4 items (i.e., anhedonia, withdrawal from activities of interest, lack of motivation, lack of interest in social interaction) with scores ranging from 0 to 16. Higher scores indicate higher levels of social disengagement.

Depression Symptoms Scale. This scale has 9 items (i.e., made negative statements, sad, pained, worried facial expression, crying or tearfulness, self-deprecation, guilt/shame, hopelessness, irritability, lack of motivation, withdrawal from activities of interest) with scores ranging from 0 to 36. Higher scores indicate greater frequency of the indicators of depression. Two items on this scale (i.e., lack of motivation, withdrawal from activities of interest) are duplicates of items in the social disengagement scale.

Anxiety Scale. This scale has 7 items (i.e., repetitive anxious concerns, unrealistic fears, obsessive thoughts, intrusive thoughts or flashbacks, episodes of panic, hypervigilance, nightmares) with scores ranging between 0 to 28. Higher scores indicate greater frequency and severity of anxiety.

The authors of the study included four additional items (i.e., decreased energy, repetitive health complaints, irritability, intrusive thoughts or flashbacks) from the interRAI ChYMH assessment into the item pool to be considered for the internalizing scale. Taken together, a total of 22 items were selected to undergo expert content validation.

2.1.2 Defining Expertise

Expertise is defined as "displayed behavior within a specialized domain and / or related domain in the form of consistently demonstrated actions of an individual that are both optimally efficient in their execution and effective in their results." (Swanson & Holton, 2001,p. 241; as cited in Germain, 2006). The authors specified the basis for expertise as relevant training, experience, and qualifications (American Educational Research Association, 1999):

1. *Qualifications.* An advanced degree in a mental health related profession (e.g., Doctor of Philosophy in clinical psychology).

2. *Experience.* Current or prior experience working with children and adolescents demonstrating internalizing mental health concerns (i.e., depression or anxiety) in a clinical setting. All expert judges had experience working with children/youth to various degrees and would have the analytical capabilities to make accurate and informed clinical judgment (Patton, 2002).

3. *Relevant training.* Experts received formal training and coursework in mental health from a recognized academic institute.

2.1.3 Judgement Quantification for the Internalizing Scale

To facilitate the evaluation process, Grant and Davis (1997) proposed the panel of experts should address representativeness, comprehensiveness, and clarity of all items in a measure. The interRAI ChYMH was created with the goals of creating a comprehensive and broad assessment of childhood outcomes, and the clarity of each individual item included in the instrument has previously been reviewed (Stewart et al., 2015). Hence, the following steps ensure that items collectively capture representation of the internalizing mental health construct (Haynes, Richard, & Kubany 1995).

Experts were recruited on a voluntary basis to participate in the content validation process. An email was sent out to invited experts that linked each participant to an online survey (www.limesurvey.com) and participants were asked to indicate the extent to which each individual item was representative of the internalizing mental health construct. All survey respondents were notified that their name would be linked to survey responses, as the authors needed to verify all participants responding in the expert panel were, in fact, invited. The following definition was provided in the survey for "internalizing":

> "Internalizing difficulties in children and adolescents refers to "conditions whose central feature is disordered mood or emotion" (Wilkinson, 2009). The terminology commonly used as ‘‘emotional’’ disorders versus ‘‘behavior’’ difficulties are synonymous with ‘‘internalizing’’ versus ‘‘externalizing’’ difficulties. Internalizing conditions are characterized by symptoms of depressed mood, anxiety, and anhedonia."

Experts assigned numerical ratings for the representativeness of each of the 22 items on a 4-point ordinal scale (1= not representative, 2=minimally representative, 3=moderately representative, and 4= strongly representative) for children (specified to be between the ages of 4 and 11), and adolescence (specified to be between the ages of 12 and 18) separately (Lynn, 1986). The 22 items were presented in random order to avoid an ordering effect (https://www.random.org/lists/), and the description provided for each item on the interRAI ChYMH assessment form was provided for the experts. Refer to Appendix A for the invitation email and design of the survey).

2.1.4 Analysis of Expert Panel Judgements

One advantage of using an online survey, compared to group discussion and consensus, is that each individual experts' respective judgments can be evaluated as statistically independent variables (American Educational Research Association, 1985). Consensus estimates were calculated to determine the extent experts shared interpretation regarding whether the item and the overall scale is representative of a latent construct (Hinkin, Tracey, & Enz, 1997; Polit et al., 2007). The content validity index (CVI; Waltz & Bausell, 1981) is a commonly reported consensus estimate, with expert ratings of relevance calculated at the item level (i.e., I-CVI) and at the scale level (i.e., S-CVI; Polit et al., 2007). I-CVI is the most suitable method in evaluating agreement compared to other consistency estimates, such as Cronbach's alpha, as the latter evaluates relative ordering of the score and internal consistency, rather than agreement in relevance of an item to a construct. In other words, even when agreement that item-to-construct relevance is low, results could yield high consistency estimates if the relative ordering of item scores are similar across experts (Polit et al., 2007). I-CVI for each of the 22 items was calculated through dividing the proportion of experts who rated the item as content valid (i.e., a rating of 3 or 4) by the total number of experts (Polit et al., 2006; Waltz & Bausell, 1981). While I-CVI is easy to calculate, and has the advantage of focus on item-to-construct relevance, it fails to adjust for chance agreement with expert panels typically consisting of small sample sizes (Lynn, 1986; Polit et al., 2007). Hence, P_C and modified Kappa (*K**) were calculated based on mathematical equations provided by Polit and colleagues (2007) in order to adjust for change agreement amongst experts with I-CVI values:

(1) Equation for Pc:

$$p_c = \left[\frac{N!}{A!(N-A)!}\right].5^N$$

(2) Equation for modified kappa (*K**)

$$k^* = \frac{\text{I-CVI} - p_c}{1 - p_c}$$

Pc, the probability of *A* endorsements of good relevance by chance, produces the binomial probability of content endorsement of *N* judges and this chance endorsement of agreement is then multiplied by .5 to the power of *N* judges. Modified kappa (k*) adjusts for convergence and mutual agreement of *A* endorsements of good relevance between 2 or more judges with endorsements by chance, is defined as the binomial chance endorsement of content relevance. The interpretation of *k** values were detailed in Fleiss (1981) and Cicchetti and Sparrow (1981), with values bounded between .40 to .59 rated as fair, .60 to .74 rated as good, and .75 to 1.00 rated as excellent. Items that received low modified kappa statistics (i.e., k* value of .74 or lower) were flagged for removal (Polit et al., 2007).

There are two commonly-used methods for calculating the overall scale content validity index (S-CVI), which include the universal agreement method (S-CVI/UA) and the averaging method (S-CVI/AVE; Polit & Beck, 2006; Polit et al., 2007). The authors made an a priori decision to calculate the S-CVI/UA after the set of items are finalized through both expert content validation and exploratory factor analyses. S-CVI/UA is the proportion of items that are rated content valid (i.e., rating of 3 or 4) by all the experts. Based on previous literature, a lower limit of acceptability for S-CVI is set at 0.80 (Polit et al., 2007). S-CVI/UA for the internalizing scale were calculated separately for children (ages 4 to 11) and adolescence (ages 12 to 18), as there may be developmental differences whether an item is content valid, such that an item could be characterized as representative to a construct for one age group, but non-representative for the other.

2.1.5 Expert Panel Ethics

Refer to Appendix D for the personal communication and correspondence with the first author and Dr. Riley Hinson, Chair of the University of Western Ontario Research Ethics board for Non-Medical Research Involving Human subjects and approval for secondary data analysis.

2.2.0 Construct Validity of the Internalizing Scale

Archival data was analyzed to determine whether the items of the internalizing scale deemed content valid by the experts possessed desirable psychometric properties, this section aims to evaluate the factor structure of the scale, test for reliability, and assess model fit for the final set of items. The procedure for data collection and statistical analyses are carefully documented in this section.

2.2.1 Participants and Procedure

The present study examined archival data collected between 2012 and 2016 using the interRAI Child and Youth Mental Health (ChYMH) from 39 mental health agencies in the Province of Ontario, Canada. The total sample comprised of 3464 clinically referred children and youth (60.3% male) between the ages of 4-18 years (M_{age} =11.85, SD = 3.58) who completed the interRAI ChYMH across 39 mental health services sites within the province of Ontario, Canada.

As described in Chapter 1, the interRAI ChYMH is a semi-structured interview comprising 400 items to assess the child/youth's strengths, needs, functioning and areas of risk to inform intervention for mental health needs. Assessors had at least 2 years of experience working in a mental health setting and were trained over a two full-day interRAI ChYMH training workshop. Every clinician was provided a users' manual for the interRAI ChYMH, which provided the intent, definition, suggested process to obtain the information, and coding for every item in the instrument (Stewart et al., 2015). Clinicians provided ratings on the instrument based on all sources of information available to them, including interview with the child and family, consultation with service providers who could provide information about the child (e.g., psychologists, teachers), and documentations on file. All ratings are required to be gathered within a three-day window (Stewart et al., 2015).

2.2.2 Item Level Analyses

Preliminary analysis of the data included evaluation of the mean, standard deviation, and variance of each individual item (refer to Table 1). Any item with a variance value of less than 1.0 were flagged as potential items for removal, as the restricted range of responses would suggest that the item is not useful is discriminating between clients with or without the latent trait (Jackson, 1970). In total, 3 items from the

internalizing scale were flagged for potential deletion due to low variance values. Moreover, item-total correlations were computed between each item and the overall scale, as low item-total correlations indicate that items may be gathered from inappropriate domains and produce measurement error and unreliability within a scale (Hinkin, Tracey, & Enz, 1997). The authors made an a priori decision to set >.20 as the acceptable value of item-total correlation (Dozois, Dobson, & Ahnberg, 1998).

2.2.3 Internalizing Scale– Unrestricted Factor Analysis

A series of unrestricted factor analyses were performed, after item-level psychometrics, to determine the dimensionality of the measure without restriction on the structural model (Tabachnick & Fidell, 1989; Child, 1990). Bootstrap sampling (number of bootstrap samples = 500) was conducted to compute robust factor analysis (Osbourne, 2014). Recent literature suggested that robust factor analysis resampling methods using multiple permutations can inform reliable and replicable summary statistics (e.g., average effect, 95% CI) and these methods generally reported acceptable model-data fit (Osbourne, 2014). Bartlett's statistic of sphericity and Kaiser-Mayer-Olkin (KMO) test were used to calculate the adequacy of the correlation matrix.

The Hull Method was conducted to determine factor retention (Lorenzo-Seva, Timmerman, & Kiers, 2011). The Hull method aims to optimize both number of parameters and goodness-of-fit in a series of factor solutions (Lorenzo-Seva et al., 2011). In a study using simulated data, the Hull method, with 85-94% success rate, outperformed Parallel Analysis of Horn (81% success rate), Minimum Average Partial Test (MAP; 51% success), and Bayesian Information Criterion (44% success rate). In the same study, out of four Hull method indices tested with varying performances, the Hull-comparative fit index (CFI) and Hull-Common part Accounted For (CAF) appeared to be the most successful indices (Lorenzo-Seva et al., 2011). Thus, the present study uses a combination of Hull-CFI, optimal implementation of parallel analysis, and examination of eigenvalues to extract factors.

Parameter estimates were generated from bootstrap sampling of estimated asymptotic polychoric correlation matrix algorithm using Bayes modal estimation, as the method is robust to normality violations (Choi, Kim, Chen, & Dannels, 2011; Flora & Curran, 2004). Polychoric correlation variance-covariance dispersion matrices were

chosen as a weighting element to reflect on the ordinal data in the scale with the assumption that an unobserved latent trait is being measured in an underlying joint, continuous normal distribution (Holgado–Tello, Chacón–Moscoso, Barbero–García, & Vila–Abad, 2010; Nunnally & Bernstein, 1994). Multiple simulation studies published in recent years (e.g., Holgado-Tello et al., 2010) suggested that the original measurement model is more accurately reproduced with polychoric correlations compared to Pearson's *r* correlations, as categorization reduces variability and thus, reduces the magnitude of loadings in Pearson's *r* obtained in the data (Garrido, Abad, & Ponsoda, 2013; Olsson, 1979). Hence, polychoric correlations would be preferred on ordinal variables (Garrido et al., 2013; Holgado–Tello et al., 2008; Timmerman & Lorenzo-Seva, 2011).

The diagonally weighted least squares (DWLS) estimator is designed for accurate parameter estimates for ordinal data, and recent Monte Carlo Simulation studies showed that DWLS produced more accurate and less biased factor loadings than robust maximum likelihood across different configurations of latent response distributions, numbers of categories, and sample size (Li, 2016). Moreover, an asymptotically distribution-free estimator is suitable when assumptions of normality and continuity of variables is violated (DiStefano & Morgan, 2014). Thus, first-order analyses were carried out by means of diagonally weighted least squares estimator with an oblique, Promin rotation (Lorenzo-Seva, 1999). An oblique rotation is most appropriate in practice, as previous research had shown that internalizing factors should have moderate correlations (Browne, 2001; Costello & Jason, 2005; Osbourne, 2014). A Promin, oblique Procrustean rotation method was used to compute a semi-specified target matrix to maximize factor simplicity (Lorenzo-Seva, 2013). Finally, model fit was reported using robust goodness of fit statistics including root mean square error of approximation (RMSEA), non-normed fit index, Comparative Fit Index (CFI), Schwarz's Bayesian Information Criterion (BIC), and Goodness of Index in Table 6 (Kline, 2011).

2.2.4 Internalizing Scale – Exploratory Bifactor Model

While the internalizing mental health subscale is predicted to have a multidimensional factor structure, an exploratory bifactor model is a useful procedure to gain additional insights into the factor structure as a single first-order factor that loads onto all primary factors specified into the model (Holzinger & Swineford, 1937; Reise,

Morizot, & Hays, 2007; Reise et al., 2010). It represents direct relationships between primary variables and higher order factors and could give insight into theoretical constructs between latent constructs and observed variables (Steer et al., 2001; Steer et al., 2005). Loadings of variables onto the general factor (i.e., GF) orthogonal to all other first order factors, and loadings onto the first-order factors (F1 to F3) and correlations between the GF and first-order factors are also depicted in this analysis (Wolff & Preising, 2005).

2.2.5 Internalizing Scale – Item Response Theory (IRT) Parameterization

The functioning of each individual item with respect to the internalizing mental health construct is further validated with Reckase's (1985) multidimensional item response theory (MIRT) model, a logistic function that models probability of success of an item for graded responses, was used to fit the final set of items. The FA modelling of the polychoric matrix is equivalent to fitting the normal-ogive version of Samejima's graded-response model (see e.g., Ferrando & Lorenzo-Seva, 2013, 2014; Mislevy, 1986; or Reckase, 2009). Eight parameters were estimated for each item for a three-factor model (i.e., one a-parameter item-discrimination value for each factor, four b-parameter item-difficulties between-category threshold values, one MDISC value for each value). Slope parameters of the a-parameter typically range from 0 to 3, with higher values representing a better indicator of the latent trait (i.e., θ) and less noise in measurement (i.e., factor weight is much greater than the residual standard deviation; Ferrando & Lorenzo-Seva, 2013). The b-parameter of category *k* represents the threshold of the item *j*, which is the point on θ, measured on a z-score between -3 and +3, where probability of scoring in category *k* or higher response is 0.5 (Reckase, 1989). MDISC is described as the analog of the a-parameter in unidimensional IRT, as there are many locations that exist in multidimensional space where an item is most discriminating (Reckase, 1985). All factor analyses, reliability estimates, and item response theory parameterizations were conducted on the FACTOR software 10.5.01 and 10.5.02 for Windows 64-bits (Baglin, 2014; Lorenzo-Seva & Ferrando, 2013). The FACTOR software is a free, stand-alone program capable of fitting exploratory factor analytic (EFA) models based on orthogonal and oblique rotation to a (partially) specified target.

2.2.6 Convergent-Discriminant Validity

After verifying the factor structure of the individual scales, Bayesian independent samples *t*-tests to discriminate between children/youth with a mood/anxiety disorder diagnosis using Diagnostic and Statistical Manual of Mental Disorders – 4th Edition DSM criteria and children/youth without a mood/anxiety disorder diagnosis. The interRAI ChYMH provides documentation for the four most impairing psychiatric diagnoses as determined by a psychiatrist, psychologist, or attending physician. The interRAI ChYMH assessors are not typically the same people who have given the child/youth a mental health diagnoses. These provisional diagnoses were obtained from the clinical record or completed by the psychiatrist, attending physician, or qualified psychologist at the time of assessment. The diagnoses were coded and ranked for their importance as factors contributing to the admission of the child/youth. For the purposes of this study, disorders were recoded as present (i.e., child/youth was assessed and met criteria) or absent (i.e., child/youth was assessed but did not meet criteria). Children/youth with comorbid mental disorders were included in the analysis.

2.2.7 Internalizing Scale– Creating a Developmentally Appropriate Scale

At present, all interRAI ChYMH subscales embedded in the instrument use a common set of items and metric for children and adolescence across the age span of 4 to 18. While there are benefits to using a common scale (e.g., simplicity of use with repeated administration, ease of tracking longitudinal change, simplicity of interpretation of scores across time), the authors acknowledge that emotional and behavioural disturbance would demonstrate substantive changes across development. Hence, a comparison analysis was conducted to determine the validity for (1) two separate scales deemed content valid for children (i.e., 15 items) and adolescence (i.e., 17 items) respectively and (2) a common scale content valid for both age groups (i.e., 12 items) in differentiating children/youth with/without a mood/anxiety disorder diagnosis to determine whether there were benefits to using separate developmentally-relevant internalizing scales for children and adolescents separately. Using multiple independent sample *t*-test, all three scales significantly differentiated ($p < 0.05$) between children/youth with and without depression and/or anxiety. However, results showed that the 12-item scale, common to both children and adolescence, differentiated groups with a

larger effect size compared to the separate scales. Hence, given the benefits of greater convergent-discriminant validity and scale brevity, a common set of 12 internalizing items were used for children and adolescence across the age span of 4 to 18. The analyses detailed in this section are not included in this document, but are available upon request.

2.3.1 Externalizing Scale – Content Validity

The expert content validation process for the externalizing mental health subscale followed the aforementioned protocol suggested by Lynn (1986). Prior to the expert content validation, the authors carefully reviewed all items relevant to problematic, externalizing behaviours embedded in the interRAI ChYMH. Items from existing subscales related to hyperactivity and disruptive behaviour, including the hyperactive/distraction scale and the disruptive/aggressive behaviour scale, were integrated for expert panel evaluation to be considered for items in the externalizing mental health subscale:

Hyperactive/Distraction Scale (HDS). The HDS consists of four items that assess the frequency of hyperactivity and distractibility (i.e., impulsivity, ease of distraction, hyperactivity, and disorganization). The frequency of each behaviour is assessed using a 5-point ordinal scale (0 = not present to 4 = exhibited daily in last 3 days, 3 or more episodes or continuously). The total score has a possible range of 0–16, and higher scores indicate higher levels of hyperactivity and distractibility (Stewart et al., 2015; Stewart & Hamza, 2017).

Disruptive/Aggression Behaviour Scale (DABS). The DABS has five items assessing the frequency and severity of aggressive and disruptive behavior (i.e., physical abuse, verbal abuse, socially inappropriate or disruptive behavior, destructive behavior toward property, outbursts of anger). The items are assessed on a 5-point ordinal scale (0 = not present to 4 = exhibited daily in last 3 days, 3 or more episodes or continuously), and the total DABS score consists of the total on the five items, with a possible range of 0–20 and higher scores indicating higher levels of aggressive and disruptive behavior (Stewart et al., 2015; Stewart & Hamza, 2017).

The authors included 14 additional items to be considered in the externalizing mental health subscale. These items included (1) repetitive lying, (2) elopement attempts/threats, (3) demonstrates limited understanding of consequences to behaviour,

(4) preoccupation of violence, (5) bullying peers, (6) fire-setting or misuse of ignition, (7) argumentativeness, (8) intimidation of others or threatened violence, (9) violent ideation, (10) violence to others, (11) cruelty to animals, (12) defiant behaviour, (13) stealing, and (14) expressions supportive of criminal activity. Refer to Appendix B for detailed descriptions and scoring criteria for each item. A total of 14 items were scored using a 6-point ordinal scale (0=not present to 5=present in the last 3 days) and 9 items were scored using a 5-point scale (0=not present to 4 = exhibited daily in last 3 days, 3 or more episodes or continuously). A total of 23 items were selected to undergo expert content validation.

2.3.2 Externalizing Scale – Defining Expertise

The authors reviewed every expert's credentials and deemed each participant as a knowledgeable individual with experience in the field. As mentioned previously, the authors specified the basis for expertise as relevant training, experience, and qualifications (American Educational Research Association, 1985):

1. *Qualifications.* Doctor of Philosophy degree in Clinical Psychology.
2. *Experience.* Current or prior experience working in a clinical setting with children exhibiting externalizing mental health concerns (e.g., disruptive behaviours). All expert judges had experience, as a licensed psychologist, working with children/youth to various degrees and would have the observational and analytical capabilities to make accurate, clinical judgment (Patton, 2002).
3. *Relevant training.* Experts should have received formal training and coursework in mental health from a recognized academic institute.

A total of 13 expert judges (6 males, 7 females) served as panel members. The final panel of expert judges consisted of all licensed doctoral-level psychologists in Canada. The authors of the present study reviewed every expert's credentials and deemed each participant as a knowledgeable individual with experience in the field, and would demonstrate the ability to analyze whether an item would be content valid as an externalizing mental health indicator in child and adolescent psychopathology (Schriesheim, Powers, Scandura, Gardiner, & Lankau, 1993; Minichiello, Aroni, Timewell, & Alexander, 1995).

2.3.3 Analysis of Expert Judgement

Experts were recruited on a voluntary basis to participate in the expert content validation process. The email linked each participant to an online survey (www.limesurvey.com) and participants were asked to indicate the extent to which individual items were representative as an externalizing mental health indicator. The following definition "externalizing" was provided:

> Externalizing" and "internalizing" difficulties in children and adolescents are synonymous with "behavioural" and "emotional" problems respectively. (Achenbach, 1978). Externalizing behaviour refers to a child/youth's problematic "outward behavior" as the child acts excessively negative towards the external environment (Campbell, Shaw, & Gilliom, 2000; Eisenberg et al., 2001). The externalizing behaviour construct includes aggression, conduct problems, delinquency, hyperactivity, and "undercontrolled behaviour" (Hinshaw, 1987; Liu, 2004).

The ratings of items and administration procedure were identical to the internalizing expert panel described in the earlier sections. Refer to Appendix C to view the invitation email for the externalizing scale expert panel and full questionnaire provided for the experts. Like the internalizing panel, I-CVI for each item were subsequently calculated for each item and items with low k^* values (<.74) were flagged for removal.

2.4.1 Externalizing Scale – Participants, Procedures, and Item Analyses

The same dataset and administration procedure was used for the internalizing scale. Refer to the earlier sections for more details.

2.4.2 Externalizing Scale– Unrestricted Factor Analysis

A series of unrestricted factor analyses were performed after item-level psychometrics. Bootstrap sampling (number of bootstrap samples = 500) was conducted to compute robust factor analysis, with the diagonally weighted-least squares estimator and an oblique, Promin rotation (Lorenzo-Seva, 1999). Bartlett's statistic, and Kaiser-Mayer-Olkin (KMO) test were used to calculate the adequacy of the correlation matrix. Because some items were measured using a 5-point ordinal scale while others were measured using a 6-point ordinal scale, Pearson's product-

moment correlation matrix was used for factor analysis, as it de facto entails the cross products of standardized scores. The Hull-CAF Method was conducted to determine factor retention (Lorenzo-Seva, Timmerman, & Kiers, 2011). Results of the unrestricted factor analysis were interpreted based on several fit indices, including root mean square error of approximation (RMSEA; Steiger, 1990), comparative fit index (CFI; Bentler, 1990), and goodness-of-fit index (GFI).

2.4.3 Externalizing Scale– Recoding Items for Summation and IRT

While a total composite could be computed through unweighted summing of scores, doing so potentially conflates separate dimensions of severity and confluence of dimensions when scores are interpreted (Sykes, Hou, Hanson, and Wang, 2002). Hence, to create a total score, variables on the externalizing scale are recoded as present (coded as 1) or absent (coded as 0) and then summed. A separate factor analysis was conducted with variables recoded as absent and present, and the factor structure and factor loadings replicated onto its original two dimension factor structure. MacDonald's and Reckase's two-parameter item response theory is applied to create item discrimination parameters based on the number of identified factors with a single threshold value (i.e., threshold between absent and present). All factor analyses, reliability estimates, and item response theory parameterizations were conducted on the FACTOR software 10.5.02 for Windows 64-bits (Baglin, 2014; Lorenzo-Seva & Ferrando, 2013).

2.5.0 Criterion Validity of Final Internalizing and Externalizing Subscales

The next step was to determine the degree to which the interRAI ChYMH internalizing and externalizing scales were correlated with other gold-standard childhood internalizing and externalizing childhood mental health measures.

2.5.1 Participants and Procedure

A small subset of participants (N = 48–53) from the larger interRAI ChYMH dataset completed additional criterion measures in the same time frame as the interRAI ChYMH assessment, including the Beck Youth Inventories, Social Skills Improvement System (SSIS), the Child and Adolescent Functional Assessment Scale (CAFAS), the Child Behavior Checklist (CBCL), and the Brief Child and Family Phone Interview (BCFPI).

2.5.2 Internalizing Scale – Scoring and Missing Data

The final internalizing scale consists of 12 items rated on a 5-point ordinal scale (0 = not present to 4= exhibited daily in last 3 days, 3 or more episodes or continuously). The criterion measures were completed with the interRAI ChYMH assessment before the item "hypervigilance" was introduced into the assessment instrument, and hence, the item is missing for all participants in this smaller subset. The missing "hypervigilance" value was replaced with the mean of the sum of the remaining three variables in the anxiety factor (i.e., repetitive anxious complaints, unrealistic fears, episodes of panic) on the anxiety factor to compute total scores. Scores in this measure range from 0 to 48, with higher scores revealing greater frequency and severity of internalizing symptoms.

2.5.3 Externalizing Scale – Scoring

The final externalizing scale consists of a total of 12 items, with 5 items that are rated on a 5-point ordinal scale (0 = not present to 4= exhibited daily in last 3 days, 3 or more episodes or continuously) and 7 items rated on a 6-point ordinal Scale (0=never to 5= in the last 3 days). Scores are recoded as 0, which equates to absent, or 1, which equates to present. Scores range from 0 to 12, with higher scores indicating greater frequency of externalizing behaviours.

2.5.4 Criterion Measures

Beck Youth Inventory. The Beck Youth Inventory measures the frequency of maladaptive thoughts, feelings, and behaviours that represent emotional and social impairment in children and adolescents 7 to 18 years old. The inventory comprises of 5 sub-scales (i.e., depression, anxiety, anger, disruptive behaviour, self concept) consisting of 20-self report items each (Beck, Beck, & Jolly, 2001).

Social Skills Improvement System (SSIS). The SSIS is a parent or caregiver-report measure of social skills and problematic behaviours. There are 10 subscales in the SSIS, which include hyperactivity/inattention, bullying, self-control, engagement, empathy, responsibility, assertion, cooperation, communication, internalizing behaviour, and externalizing behaviour (Gresham et al., 2010; Gresham & Elliott, 1990). The SSIS externalizing scale consists of 12 items (e.g., disobeys rules or requests. fights with others) and the internalizing scale consists of 7 items (e.g., withdraws from others, acts anxious with others).

Child Behavior Checklist (CBCL). The CBCL is a set of standardized measures for children and adolescents between the ages of 4 to 18 years. Several subscales are embedded in the instrument, which include destructive behavior, social withdrawal, somatic complaints, anxiety and depression, social problems, thought problems, attention problems, aggressive behavior, and delinquent behaviors (Achenbach & Edelbrock, 1991). This measure consists of two broadband measures: internalizing (i.e., anxious, depressive, and over-controlled) and externalizing (i.e., aggressive, hyperactive, noncompliant, and undercontrolled) behaviours.

Child and Adolescent Functional Assessment Scale (CAFAS). The CAFAS is a clinician-administered measure that assesses functioning across several domains (i.e., school, home, community behaviour towards others, moods/emotion, self-harm, substance use, and thinking). The CAFAS demonstrates reliability, concurrent validity, and discriminant validity (Hodges & Wong, 1996).

Brief Child and Family Phone Interview (BCFPI). The BCFPI is a parent- or caregiver-report standardized measure with 9 subscales embedded, including regulating attention, regulating impulsivity and activity level, regulating attention, impulsivity and activity level, cooperativeness, conduct, separating from parents, managing anxiety, managing mood, and managing mood and self-harm (Cunningham et al., 2009; Boyle et al., 2009). In the BCPFI subscales, the separation from parents, managing anxiety, managing mood subscale create the internalizing subscale. The externalizing behaviour subscale is an 18 item scale including items from the regulating attention, impulsivity, and activity level subscale, cooperativeness subscale, and conduct subscale.

2.5.5 Statistical Analysis

Bayesian correlations tests were conducted to compare internalizing and externalizing scale scores with various subscales on criterion measures. Jeffreys's Bayes Factor (1961) computes the probability of the observed data under the null hypothesis using a prior probability (i.e., probability hypothesis is true pre-data collection) and posterior probability (probability hypothesis is true post-data collection; Kass & Raftery, 1995). This technique, unlike commonly used frequentist tests, allows researchers to quantify evidence in favor of the null hypothesis (e.g., Wagenmakers, 2007; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wetzels et al., 2011). All Bayesian correlation

analyses were conducted using JASP 0.8.1.1, which uses correlation tests proposed by Jeffreys (1961), assuming bivariate normal distribution and a uniform, default prior on rho. Interpretation of Bayes Factors as evidence for alternative hypotheses with Bayes Factor of 1–3 as weak, 3–10 as substantial, 10–30 as strong, 30–100 as very strong and >100 as decisive (Jarosz & Wiley, 2014).

2.5.6 Ethics Approval

The interRAI ChYMH database used for research purposes assigns a randomly generated study-specific participant number for each participant and includes no identifying information of the child/youth or their families. This study was approved by University of Western Ontario Ethics Board (REB #106415). Parents or legal guardians of the children/youth gave informed consent prior to assessment. The data are only accessible for researchers supervised by an interRAI fellow and therefore, cannot be transferred or deposited in an open-access repository.

CHAPTER 3: RESULTS

3.1.0 Psychometric Validation of the Internalizing Subscale

The following section details the results of the expert content validation process and unrestricted, robust factor analysis for the internalizing subscale.

3.1.1 Expert Panel Results of the Internalizing Scale

The final panel consisted of 3 psychiatrists, 6 doctoral-level clinical psychologists, 5 PhD candidates (i.e., psychologists in training) with a Master's of Science in clinical psychology enrolled in an accredited clinical psychology program, and 1 psychiatric nurse who owns a private mental health practice. In total, 15 participating experts (i.e., 4 males, 11 females) examined 22 potential internalizing instrument items.

The results demonstrated that a total of 15 and 19 items remained quantitatively content valid (I-CVI>0.78 and $k^* > .74$) as an internalizing mental health indicator for children (aged 4 to 11) and adolescents (aged 12-18) respectively. Overall, 14 of 22 items (63.63%) demonstrated excellent content validity (I-CVI>0.78 and $k^* > .74$) across both age groups. Only items that were content relevant for both children and adolescent age groups were retained for subsequent factor analyses. Close examination of modified k^* scores showed that the "hypervigilance" item was content valid for adolescents and only .02 below the cutoff of .74 in children and therefore, this item was also retained for subsequent factor analyses. The 15 items (i.e., "episodes of panic", "expressions of hopelessness", "hypervigilance", "crying, tearfulness", "sad, pained, or worried facial expression", "self-deprecation", "anhedonia", "expressions of guilt or shame", "unrealistic fears", "made negative statements", "repetitive anxious complaints/concerns", "withdrawal from activities of interest", "decreased energy", "lack of motivation", and "repetitive health complaints") deemed content valid during the expert panel evaluation process were selected to undergo unrestricted factor analysis. Of note, each of these items and their content map respectively onto the clinical literature on the conceptualization of internalizing disorders. Refer to Table 3 and Table 4 for the I-CVI, Pc, and k* scores for children and adolescents respectively in the internalizing scale.

3.1.2 Sample Characteristics

Data collected between 2012 to 2016 using the interRAI ChYMH were complete for 3464 children/youth for all variables, with the exception of 929 missing "hypervigilance" variables as it was added as part of standardized assessment in 2015. There were no other missing variables in the dataset, with the exception of newly added variables. Only complete data were used for subsequent analyses, with 2536 children/youth (M_{age} =12.16, SD = 3.59; 57.6% males) assessed between 2015 to 2016 across 34 sites. In this sample, 2272 (89.6%) of children/youth lived with their parents or primary guardian, 24 (0.9%) lived alone, 72 (2.8%) lived with other relatives, 10 (0.4%) lived with their siblings with no parents or guardians, 92 (3.6%) lived with a foster family, 65 (2.6%) lived with a nonrelative (but not a foster family). Among those children and youth referred for assessment at time of intake into care, 641 (25.3%) had no contact with a community mental health agency or professional within the past year, 720 (28.4%) had contact within 31 days or more, and 1174 (46.3%) had contact within the last 30 days. Refer to Table 5 for more demographic information.

3.1.3 Item-Level Analyses

None of the 15 items selected for subsequent factor analyses were flagged for potential deletion for low item variance values (i.e., < 1.0). All items had acceptable values of corrected item-total correlations with values ranging from .408 to .581.

3.1.4 Unrestricted Factor Analysis

Initially, 15 items were subjected to a series of robust, unrestricted factor analysis (bootstrap sample = 500) using polychoric correlations dispersion matrices and the diagonally weighted least squares method of extraction. The Kaiser-Meyer-Olkin measure of sampling adequacy was 0.880, which indicated that variables within this correlation matrix demonstrated strong relationships. Bartlett's test of sphericity was significant (χ^2 (105) = 12097.2, $p < 0.001$), indicating the factor analytic model is appropriate given the overall significance of correlations within the matrix.

Based on inspection of eigenvalues greater than one and the Hull-Comparative Fit Index (CFI) method of extraction, the three-factor solution (i.e., anxiety, depression, anhedonia) had the best fit. In contrast, the optimal implementation of parallel analysis would suggest a one-factor solution (i.e., general internalizing factor). With the Hull-CFI

assessment and parallel analysis providing conflicting suggestions of a one- and three-factor structure, a closer examination of the assessment of dimensionality was conducted. The values obtained revealed a unidimensional congruence of 0.949 (BC Bootstrap 95% CI: 0.922, 0.965), a value of explained common variance (ECV) of 0.802 (BC Bootstrap 95% CI: 0.787, 0.819), and mean of item residual absolute loadings of 0.295 (BC Bootstrap 95% CI: 0.272, 0.310). Of note, these values collectively suggested that both unidimensional and multidimensional solutions were interpretable. Hence, the root mean square residual (RMSR) statistic was conducted to assess model fit based on residual correlations for the one- and three- factor solution. RMSR represents a quantitative index which describes the average size of residual correlations once predicted response frequencies has been fitted to correspond with observed response frequencies. The values obtained were 0.1274 and 0.0279 for the one- and three- factor solution models respectively. Given recommendations that RMSR indices less than 0.08 are indicative of a good-fitting model, this analysis would suggest the retention of a three-factor model (Hu & Bentler, 1998).

Next, a series of robust factor analyses (bootstrap sampling = 500) with a three-factor model specified were conducted with the full sample, and analyzed separately for children aged 4 to 11 years and adolescents aged 12 to 18 years to ensure the factor structure and factor loadings replicated across development. Factor loadings that had a difference of less than 0.200 between the primary and secondary factor were flagged for removal. In the full sample, all items loaded onto their respective factors without any significant cross loadings and the authors labelled the resulting factors based on their item content. However, the items "crying, tearfulness" and "sad, pained, or worried expression" cross-loaded onto two separate factors (i.e., the "depression" and "anxiety" factor) in the child-only sample and the item "repetitive health complaints" cross-loaded onto two separate factors (i.e., the "anxiety" and "anhedonia" factor) in the adolescent-only sample. Moreover, close examination of the item response theory discrimination (a_j) parameters calculated based on Reckase's Multidimensional Item Response Theory Model revealed low discrimination values for all 3 items, which is indicative of high levels of noise in latent trait measurement (e.g., Ferrando & Lorenzo-Seva, 2013). Hence,

the items “repetitive health complaints”, “sad, pained, or worried expressions”, and “crying, tearfulness” were removed from subsequent analyses.

The remaining 12 items were, once again, subjected to unrestricted factor analysis using a polychoric correlation dispersion matrix, and the diagonally weighted least squares estimator. The Kaiser-Meyer-Olkin measure of sampling adequacy was 0.91 with a significant Bartlett’s Statistic ($\chi 2$ (66)= 9737.3, $p < 0.001$), indicating that the data were suitable for factor analysis. When the Hull-CFI Method was, once again, used to extract factors (Lorenzo-Seva, Timmerman, & Kiers, 2011), the goodness-of-fit index using Robust comparative fit index (CFI) reached 0.991 (*df*= 33; Scree test values: 17.266) with three interpretable factors and a three-factor solution was advised. The first factor (i.e., anhedonia) explained 44.50% of the common variance between items, showing evidence of a dominant factor (e.g., accounts for 20% of variance; Hambleton, Swaminathan, & Rogers, 1991). The second factor (i.e., anxiety) accounted for 13.12% of variance (i.e., 57.51% cumulative variance), and the third factor (i.e., depression) accounted for 11.74% of variance (i.e., 69.36% cumulative variance). The analysis was conducted with the full sample, and analyzed separately for children aged 4 to 11 years and adolescents aged 12 to 18 years to ensure the factor structure and factor loadings replicated across development. The factor structure did, in fact, replicate across both age groups and all factor loadings were in the acceptable range (>0.300) onto the primary factor and no significant cross-loadings were observed. The rotated factor loadings demonstrated that the items clustered to load onto three interpretable factors of anhedonia, anxiety, and depression factors, as hypothesized. Refer to Table 7 for factor loadings with bias-corrected and accelerated (Bca) bootstrap 95% confidence intervals (CI) of the internalizing subscale.

3.1.5 Application of the Exploratory Bifactor Model

A dominant first dimension, as revealed by a large first to second eigenvalue difference (first eigenvalue = 4.97, second eigenvalue = 1.17), was evident in this measure. Hence, exploratory bifactor analysis was carried out by means of diagonally weighted least squares and the model specified included a three-factor structure along with a general factor (GF or “internalizing”) that loads directly onto the primary variables (Holzinger & Swineford, 1937; Schmid, 1957; Reise et al., 2010). RMSR for this matrix

is 0.022, which is proximate to the RMSR for the 3-factor model, suggesting both models fit the data equally well. To further assess model fit, comparison of the one- and three-factor model fit was presented in Table 6 (Kline, 2011). All values, with the exception of the goodness of fit index in which both models produced a value of 1.00, suggested that the three-factor model is a better-fitting model than the one-factor solution. Table 8 respectively for the fit indices of the bifactor model and factor loadings and correlations between the GF and first-order factors in the exploratory bifactor model.

3.1.6 Reckase's Multidimensional Item Response Theory Model

Standardized weighted root mean squares of item residuals (RMSR) can inform model-data fit in the conducted IRT analysis (Drasgow, Levine, Tsien, Williams, & Mead, 1995). Using Kelley's (1935) criterion, the expected mean value of root mean square of residuals is 0.0199. As mentioned above, the RMSR for this matrix is 0.0279 (BCA Bootstrap 95% CI: 0.025, 0.029), suggesting items showed negligible misfits and all items are closely related to the overall latent trait. The IRT discrimination parameter (a_j) for an item is conceptually similar to a factor loading and represents the slope of the item response expressed in probability on the latent dimension. A lower value of a_j would indicate the item is less related to the underlying trait. Value of item discrimination parameters for the anhedonia factor parameterization (a_1) were between 0.871 and 1.661. The item discrimination parameters (a_2) for the anxiety factor were between 0.850 and 1.527. The item discrimination parameters for the depression factor (a_3) were between 1.007 and 2.013. Consistent with the item discrimination parameters (a_j) values, MDISC for items ranged from 0.854 to 2.036.

The multidimensional graded-response model also provides threshold parameters, with the number of thresholds equal to the number of response options minus one (i.e., four thresholds in the present analyses referred to b_1 to b_4). Threshold values represent the latent trait level (i.e., θ) expressed in standardized z scores) at which 50% of respondents endorse or "cross over" into the higher item response category. The thresholds for the lowest internalizing item (b_1) ranged from -0.460 to 1.058, while the threshold for the highest category (b_4) ranged from 1.370 to 3.340 on a z-score scale. Inspection of these threshold values for each item shows that in general they are well dispersed across the item response options. Refer to Table 9 for Reckase's

multidimensional item response theory item discrimination, MDISC, and category threshold parameterization values.

3.1.7 Scale Distribution, Scoring, and Scale-CVI of the Internalizing Scale

The final internalizing subscale consisted of 12 items, with each item scored using an ordinal scale ranging from 0 and 4 and total scores ranging between 0 to 48 (M = 10.45; SD = 8.74; Cronbach's α = .88). A Shapiro-Wilk's test ($p < .001$; Shapiro & Wilk, 1965; Razali & Wah, 2011) and a visual inspection of their histograms, normal Q-Q plots and box plots revealed that internalizing scale scores were not normally distributed for both males and females with a skewness of 1.000 (SE= 0.064) and a kurtosis of .637 (SE=.128) for males and a skewness of .988 (SE= .075) and a kurtosis of .513 (SE=.149) for females (Cramer, 1998; Cramer & Howitt, 2004).

In children, the S-CVI/UA for the total scale, once low modified k^* and cross-loading items were removed, increased from 0.82 to 0.89 (Polit et al., 2007). In adolescents, the S-CVI/UA for the total scale, once low modified k^* and cross-loading items were removed, increased from 0.88 to 0.97 (Polit et al., 2007). The descriptive statistics of the final 12-item internalizing scale and internalizing scale scoring sheet is presented in Table 10 and Appendix E respectively.

3.1.8 Group Differences in Total Scores

Sex Differences. Consistent with this previous literature (e.g., Wang et al., 2015), girls had higher internalizing scale scores (M = 11.74, SD = 9.61, n = 1074) than boys (M = 9.50, SD = 7.91, n = 1461; $t(2036.272) = 6.246$, $p < 0.001$). Further, Cohen's effect size value obtained was .25, suggesting low effect size. The Bayesian t-test also estimated a Bayesian Information Criteria factor (Wagenmakers, 2007), comparing the fit of the data under the null hypothesis and the alternative hypothesis. An estimated Bayes factor (Bf_{10}) is obtained at 3.29e+07, suggesting that the alternative hypothesis predicts 3.29 e+07 times better than the null hypothesis.

Inpatient Facility and Outpatient Treatment Comparison. Symptom severity in the internalizing subscale was expected to differ between inpatients and outpatients. Surprisingly, outpatients had higher internalizing scale scores (M = 10.55, SD = 8.76, n=160) than inpatients (M = 8.91, SD = 8.76, n = 2375; $t(183.562) = 2.405$, $p < 0.05$, Bf_{10}

= 1.20). Further, Cohen's effect size value obtained was .19, suggesting very low effect size.

Age differences. Internalizing mental health scores were compared amongst children (i.e., aged 4 to 11) and adolescents (i.e., aged 12 to 18). Adolescents had higher internalizing scale scores (M = 11.43, SD = 9.18, n = 1471) than children (M = 9.09, SD = 7.89, n = 1064; $t(2459.770) = 6.246$, $p < 0.001$). Further, Cohen's effect size value obtained was .27, suggesting low effect size. An estimated Bayes factor (Bf_{10}) is valued at 2.22e+08, suggesting that the alternative hypothesis predicts 2.22 e+08 times better than the null hypothesis.

3.1.9 Concurrent Validity with DSM Diagnosis.

Mood Disorder. Children/youth with a mood disorder diagnosis had higher internalizing scale scores (M = 14.43, SD = 9.56, n=410) than those without the diagnosis (M = 9.22, SD = 8.06, n=1475; $t(580.390) = -10.077$, $p < 0.001$). Further, Cohen's effect size value obtained was .59, suggesting a moderate to large effect size. Close examination of the prior distribution and posterior distribution demonstrated that most of the posterior distribution falls on large values (median= -0.616, 95% Credible Interval = -0.727, -0.509) of the effect size. The graphical representation of the Bayes factors robustness check demonstrated that changes the values of the prior width does not substantially affect the BF_{10}.

Anxiety Disorder. Children/youth with an anxiety disorder diagnosis had higher internalizing scale scores (M = 13.45, SD = 8.09, n = 917) than those without the diagnosis (M = 8.09, SD = 7.83, n =1003; $t(1828.154) = -13.903$, $p < 0.001$). Further, Cohen's effect size value is .64, suggesting moderate to large effect size. An estimated Bayes factor (Bf_{10}) suggested that the alternative hypothesis predicts the current data 8.39e+38 times better than the null hypothesis. Close examination of the prior distribution and posterior distribution suggested that most of the posterior distribution falls on large values (median= -0.635, 95% Credible Interval =-0.727, -0.542) of the effect size. The graphical representation of the Bayes factors robustness check suggested that changes the values of the width of the prior does not affect the BF_{10}.

3.2.0 Psychometric Evaluation of the Externalizing Subscale

The following section details the results in the expert content validation process and unrestricted factor analysis for the externalizing subscale.

3.2.1 Expert Panel Results

A total of 14 and 19 items remained quantitatively valid as externalizing symptoms or behaviours for children (aged 4 to 11 years) and adolescents (aged 12-18 years) respectively. See **Table 11 and Table 12 for the I-CVI, Pc, and k* scores for children and adolescents respectively for the externalizing scale**. Overall, 14 of 23 items (60.87%) showed excellent content validity (I-CVI>0.78 and k* > .74) across both age groups. With discrepancies amongst ratings in expert judges, items that were deemed content relevant for either children and/or adolescent age groups were kept for subsequent analyses to further evaluate for fit in the measurement model.

3.2.2 Sample Characteristics

Data were complete for 3464 children/youth. At the time of assessment, 3117 (90.0%) of children/youth lived with their parents or primary guardian, 31 (0.90%) lived alone, 94 (2.7%) lived with other relatives, 12 (0.3%) lived with their siblings with no parents or guardians, 121 (3.5%) lived with a foster family, 88 (2.5%) lived with a nonrelative (but not a foster family), and 1 (0.0002%) unspecified. Among those children and youth referred for assessment at time of intake into care, 848 (24.5%) had no contact with a community mental health agency or professional with the past year, 979 (28.3%) had contact within 31 days or more, and 1636 (47.2%) had contact within the last 30 days. Refer to **Table 13** for more details on the demographic information.

3.2.3 Item Level Analyses and Exploratory Factor Analysis

Participant item endorsement frequencies was low (i.e., variance < 1.0) for 3 items (i.e., “fire-setting/misuse of ignition”, “cruelty to animals”, “expressions of support of criminal activity”) and these items were flagged for deletion. Moreover, item-total correlations were computed between each item and the overall scale. All items reached acceptable levels for corrected item-total correlations (i.e., >.200), with values ranging from .269 to .725.

The remaining items were subjected to a series of robust, unrestricted factor analyses using Pearson’s correlations matrices and the diagonally weighted least squares

method of extraction. The Kaiser-Meyer-Olkin measure of sampling adequacy was 0.941, which indicated that the variables within this correlation matrix demonstrated strong relationships suitable for this analysis. Bartlett's test of sphericity was significant (χ^2 (190) = 13153.5, $p < 0.001$), indicating the factor analytic model is appropriate given the overall significance of correlations within the matrix. Based on inspection of the eigenvalues greater than one and Hull-Robust Common Part Accounted For (CAF) to extract factor, the value reached 0.48 (*df*= 43) with two factors and the two-factor solution was suggested. In contrary, when the Hull-Robust Comparative Fit Index (CFI) method of factor extraction and optimal implementation of parallel analysis was used, a one-factor solution was advised. Further, assessment of dimensionality revealed a unidimensional congruence of 0.945 (BC Bootstrap 95% CI: 0.933, 0.955), a value of explained common variance (ECV) of 0.801 (BC Bootstrap 95% CI: 0.789, 0.814), and mean of item residual absolute loadings of 0.295 (0.278, 0.301). Taking all three values into consideration, both the unidimensional and multidimensional solution would be interpretable.

With the one- and two- factor solution both interpretable, residual correlations were conducted using the root mean square residual (RMSR) statistic; the values obtained were 0.095 and 0.0327 for the one- and two- factor solution respectively. This analysis led to the retention of two factors and model fit was further evaluated based on root mean square error of approximation (RMSEA), non-normed fit index, Comparative Fit Index (CFI), Schwarz's Bayesian Information Criterion (BIC), and Goodness of Index in Table 14 (Kline, 2011). Of note, all values suggested that the two-factor solution is a better fitting model.

Subsequently, a series of robust, unrestricted factor analyses (bootstrap sampling = 500) were conducted with the full sample, and analyzed separately for children aged 4 to 11 years and adolescents aged 12 to 18 years to ensure the factor structure and factor loadings replicates across age groups. In the full sample, "limited understanding of consequences of behaviour", "socially inappropriate/disruptive behaviour", and "repetitive lying" cross-loaded onto their respective factors. The items "physical abuse" and "destructive behaviour towards property" cross-loaded onto the two factors in the

adolescent-only sample. These items were consequently removed from subsequent analyses.

In the analysis with the remaining 12 items, the rotated factor loadings revealed that the items clustered to load onto two interpretable factors of proactive and reactive aggression. The first factor explained 41.4% of the common variance between items, showing evidence of a dominant factor (e.g., accounts for 20% of variance; Hambleton, Swaminathan, & Rogers, 1991). This first factor (i.e., proactive aggression or "cold" aggression) consisted of 7 items (i.e., stealing, elopement attempts/threats, bullying peers, preoccupation of violence, violence to others, intimidation of others or threatened violence, violent ideation) and demonstrated good internal consistency (α =0.77). The second factor (i.e., reactive aggression or "hot" aggression) consisted of 5 items (i.e., impulsivity, verbal abuse, outburst of anger, defiant behaviour, argumentativeness) that demonstrated high internal consistency (α = 0.88) and accounted for 12.8% of the variance. Acceptable factor loadings values were obtained ranging from 0.586 to 0.895 in the reactive aggression factor and 0.394 to 0.780 in the proactive aggression factor. The analysis was conducted again, with the sample split into children (4-11 years) and adolescents (12-18 years), and the factor structure replicated with no significant cross loadings. Refer to Table 15 for the rotated loading matrix of the 12-item externalizing subscale in the full sample.

Given the strong evidence of both unidimensional and multidimensional construct, an exploratory bifactor model was applied, and close examination of the factor loading values suggested that it was an equally good-fitting explanatory model compared to the two-factor solution. Refer to Table 14 and Table 16 respectively for the model fit statistics of the exploratory bifactor model and rotated factor loadings using the exploratory bifactor in the externalizing subscale. RMSR for this matrix is 0.0244, which is proximate to the RMSR for the two-factor model, suggesting both models fit the data equally well.

3.2.4 Item Response Theory Parameterization

The multidimensional graded-response model provided two item discrimination parameters (i.e., one per factor) and a single category threshold parameter as responses are coded as present (i.e., 1) or absent (i.e., 0) before scoring. Standardized weighted root

mean squares of item residuals (RMSR) obtained in this model is 0.043, suggesting items showed negligible misfits and all items are closely related to the overall latent trait. The IRT discrimination parameter (a_j) for the proactive aggression factor parameterization (a_1) and reactive aggression factor (a_2) are shown in Table 16. Threshold values represent the latent trait level (i.e., θ) expressed in standardized z scores) at which 50% of respondents endorse or "cross over" from absent to present. The category thresholds values ranged from -1.859 to 1.692. Refer to Table 16 for McDonald-Reckase's multidimensional two-parameter normal ogive model of item discrimination, MDISC, multidimensional difficulty (i.e., MDIFF), and category threshold parameterization values (Reckase, 1997).

3.2.5 Scale Distribution, Scoring, and Scale-CVI in the Externalizing Scale

Descriptive statistics of the 12-item externalizing scale are presented in Table 10. Scores range from 0 to 12, with higher scores indicating greater frequency of externalizing symptoms. A Shapiro-Wilk's test ($p > .05$) (Shapiro & Wilk, 1965; Razali & Wah, 2011) and a visual inspection of their histograms, normal Q-Q plots, and box plots showed that scale scores were not normally distributed for both males and females with a skewness of -.100 (*Std. Error*= .054) and a kurtosis of -.701(*Std. Error* =.107) for males and a skewness of .315 (*Std. Error*= .066) and a kurtosis of -.960 (*Std. Error* =.132) for females (Cramer, 1998; Cramer & Howitt, 2004).

In children, the final S-CVI for the total scale, items with low variance, items with low modified *k** values, and cross-loading items were removed, increased from 0.79 to 0.83 (Polit et al., 2007). In adolescents, the final S-CVI for the total scale, once items with low variance, items with low modified *k** values, and cross-loading items were removed, increased from 0.87 to 0.94. The final 12-item externalizing scale scoring template is presented in the Appendix E.

3.2.6 Group Differences in Externalizing Subscale Scores

Sex Differences. The Bayesians *t*-test suggested that boys had higher externalizing scale scores ($M = 5.71$, $SD = 3.07$, $n = 2090$) than girls ($M = 4.20$, $SD = 3.25$, $n = 1374$; $t(2821.497) = 13.721$, $p < 0.001$, Bf_{10}=1.411e+39). Further, Cohen's effect size value is .48, suggesting a moderate effect size.

Inpatient Facility and Outpatient Treatment Comparison. Symptom severity in the externalizing scale was expected to differ between inpatients and outpatients. In this sample, inpatients had higher externalizing scale scores (M = 7.17, SD = 3.03, n = 331) than outpatients (M = 4.90, SD = 3.17, n = 3133; t(410.218) = 12.900, $p < 0.001$, Bf_{10}=2.28e+31). Further, Cohen's effect size value is .73, suggesting a moderate to large effect size.

Age differences. In this sample, children aged 4 to 11 years had higher internalizing scale scores (M = 5.57, SD = 2.92, n = 1586) than adolescents (M = 4.73, SD = 3.43, n= 1878; t(3461.763) = 7.758, $p < 0.001$, Bf_{10}=1.41e+31). Further, Cohen's effect size value is .26, suggesting a low effect size.

3.2.7. Concurrent Validity of Externalizing Subscale and DSM Diagnosis.

Attention Deficit Hyperactivity Disorder (ADHD) Diagnosis. Children/youth with an ADHD diagnosis had higher externalizing scale scores (M = 6.65, SD = 2.75) than those without the diagnosis (M = 3.70, SD = 3.01; t(2730.001) = -27.217, $p < .001$). Further, Cohen's effect size value is 1.02, suggesting a very large effect size. The obtained Bayes factor is 3.23e+142, and close examination of the distributions shows that the posterior distribution falls on large values of the effect size (median=-1.026; 95% CI: -1.102, -0.948). The Bayes factor is large across changes of the prior width, suggesting the results are robust regardless of the default prior width.

Disruptive Behaviour disorder. Children/youth with a disruptive behaviour disorder diagnosis had higher externalizing scale scores (M = 7.55, SD = 2.43, n = 796) than those without the diagnosis (M = 4.27, SD = 3.04, n = 1991; t(1815.726) = -29.917, $p < 0.001$). Further, Cohen's effect size (d) value is 1.19, suggesting a very large effect size. The obtained Bayes factor is 6.63e+140, and close examination of the distributions shows that the posterior distribution falls on large values of the effect size. The Bayes factor is large across changes of the prior width, suggesting the results are robust regardless of the default prior width.

3.3.0 Criterion validity of the Internalizing and Externalizing Subscales

Results from the expert panel and factor analyses provided initial support for the content validity and measurement model respectively in the internalizing and externalizing subscales. Bayesian Pearson's ρ correlations, with the prior stretched beta

width of 1, were conducted between ChYMH internalizing and externalizing subscales and various criterion measures. With the hypothesis clearly directional, a hypothesis of whether the ChYMH subscale and criterion subscale would correlate positively or negatively was set a priori for each criterion subscale. Consequently, a one-sided Bayesian model specification (BF_{0+}) was warranted, given that every value of the correlation ρ was not equally likely a priori.

3.3.1 Participants and Missing Data

Community-based data was collected between 2013 to 2014. Children/youth aged 6 to 18 years (*M*= 11.60, *SD*=2.87; 75.6% males) were assessed using the interRAI ChYMH and additional criterion measures within a three-day window, including the Social Skills Improvement System (SSIS), the Child and Adolescent Functional Assessment Scale (CAFAS), Beck Youth Inventory (Beck), the Child Behaviour Checklist (CBCL), and the Brief Child and Family Phone Interview (BCFPI).

Of note, the item "hypervigilance" from the internalizing scale was not part of standardized assessment between 2013 and 2014. Hence, the mean of the remaining 3 items from the anxiety factor (i.e., "episodes of panic", "repetitive anxious complaints", "unrealistic fears") was computed to replace the hypervigilance item to generate a total score. Using the dataset from the full sample (*N*=3464), the "true" anxiety factor of the internalizing subscale was correlated with the "modified" anxiety factor (i.e., the mean of anxiety factor replacing the hypervigilance variable). The Pearson's *r* value obtained was 0.968, suggesting the replacement of the "hypervigilance" item with the mean of the remaining items from the anxiety factor should not substantially affect the criterion validity correlations.

3.3.2 Correlation of the ChYMH Subscales with Validated Childhood Measures

As predicted, results demonstrated that the interRAI ChYMH internalizing and externalizing subscales correlated well in the expected direction with the criterion measures. The strongest correlations were between the ChYMH internalizing subscale and the criterion scales of the CBCL internalizing scale (Pearson's ρ = .624), in which the alternative hypothesis (H_1) predicted the data 30352 times better than the null hypothesis (H_0), and CBCL internalizing scale (Pearson's ρ = 0.61), in which the alternative hypothesis (H_1) predicted the data 15310 times better than the null hypothesis. In

addition, the BCFPI managing mood and self-harm scale correlated highly with the interRAI ChYMH internalizing subscale (Pearson's ρ = .55), such that the observed data were 1396 times more likely under H_1 than under H_0. Lower correlations were found between the ChYMH internalizing subscale and the CAFAS mood/emotions subscale (Pearson's ρ = 0.212), Beck depression subscale (ρ = 0.231), and Beck anxiety subscale (ρ = 0.231), with $Bf_{+0}(d)$ ranging from 0.65 to 1.18. Refer to Table 18 for Pearson's rho Bayesian correlations, Jeffreys's Bayes Factor, and evidence for H_1 based on Jeffreys's criterion for the ChYMH internalizing subscale and criterion measures of SSIS, BCFPI, CBCL, Beck, and CAFAS.

The strongest correlations were found between the ChYMH externalizing subscale and the criterion scales of the SSIS externalizing scale (Pearson's ρ = 0.648), in which the alternative hypothesis (H_1) predicted the data 216,652 times better than the null hypothesis, and BCFPI externalizing scale (Pearson's ρ = 0.632), in which the alternative hypothesis (H_1) predicted the data 199,157 times better than the null hypothesis. In addition, the ChYMH externalizing subscale correlated strongly with the criterion measure of the CBCL externalizing scale (Pearson's ρ = .645), in which the alternative hypothesis (H_1) predicts the data 65257 times better than the null hypothesis. Strong evidence for criterion validity were also found in the Beck anger subscale (Pearson's ρ = 0.486), Beck disruptive behaviour subscale (Pearson's ρ = 0.433) and the CAFAS behaviour towards others subscale (Pearson's ρ = 0.442). Refer to Table 19 for Pearson's rho correlations, Jeffreys's Bayes Factor, and evidence for H_1 between the ChYMH Externalizing Subscale and Criterion Measures of SSIS, BCFPI, CBCL, Beck, and CAFAS.

CHAPTER 4: GENERAL DISCUSSION

The primary purpose of this study was to examine the internalizing and externalizing subscales of the interRAI ChYMH in a large sample of clinically-referred Canadian children and adolescents and to make any needed modifications to improve both the psychometric and clinical utility of these measures. Previous research has acknowledged the importance of measuring both broadband and narrow-band symptoms in childhood mental health (Achenbach et al., 2016). While the interRAI ChYMH is adequately equipped with the narrow-band scales (e.g., Stewart & Hamza, 2017), the findings of this study demonstrated that the current modifications to both the internalizing and externalizing scales result in acceptable psychometric properties and have strong criterion validity.

Following expert content validation, theoretically incoherent items were flagged for deletion if items were deemed not representative of the overall internalizing/externalizing latent construct. The overall scale-content validity index (universal agreement; S-CVI/UA) deduced from results of the expert panel evaluation demonstrated that the final internalizing and externalizing scales included items that adequately captured the conceptualization and operationalization of internalizing and externalizing mental health conditions (Hinkin, 1995; Hinkin & Schriesheim, 1989). Item-level analyses and unrestricted factor analyses further reduced the number of items on both the internalizing and externalizing scales. Within both scales, the resulting three- and two- factor solution respectively each demonstrated acceptable levels of internal consistency within each factor. While the items within the externalizing scale had larger overlap reflected in the inter-item correlations and inter-factor correlations, this finding is consistent with previous research that suggested proactive and reactive aggression are highly correlated constructs (e.g., Smeets et al., 2017). In terms of known-groups validity, group discrimination demonstrated medium to large effect sizes for mood disorders and anxiety disorders, and very large effect sizes for externalizing behaviours. These results suggested that higher scores on these ChYMH measures may be predictive of a DSM diagnosis. Moreover, almost all reported Bayes factors from the Bayesian correlations tests indicated that the observed data offered overwhelming support for the existence of the expected relationship between the internalizing/externalizing scales and

criterion measures. Traditionally, support in favour of the alternative hypothesis, using a p-value, only provided researchers information when the null hypothesis is true, and ignores provision of evidence if the alternative hypothesis were true (Berkson, 1938). Overall, these results provided supporting evidence that the internalizing and externalizing subscales are a reliable and valid measure of the broadband constructs of mental health.

4.1 Discussion of the Measurement Model

Based on results from the robust goodness-of-fit indices in the measurement model, the present study demonstrated that the two- and three-factor models were determined to be the most parsimonious, well-fitting model for the externalizing and internalizing scales respectively. The psychometric properties were validated through unrestricted factor analysis, as the partial correlations between the variables and the measured constructs were deemed excellent with no significant item cross-loadings between factors.

The individual contribution of specific items and person parameters were further investigated using multidimensional item response theory parameterizations. The item response theory analyses conducted in this paper further supported the validity of the measurement model. Values of a_j ranged from 0.850 to 2.013 in the internalizing scale, which were in the acceptable range, suggesting that all items adequately reflect the latent trait measured. Within the externalizing scale, item discrimination values were lower for the items “bullying peers”, “stealing”, and “elopement attempts/threats”, indicating greater level of noise in measurement. These results may be indicative of the more diverse manifestations of externalizing behaviours in children/adolescents. For example, even if a child/youth has high trait levels of the externalizing latent trait, stealing is not a certain behaviour. Future studies should consider the inclusion of a “d” parameter in the item response theory model (i.e., changing item parameter upper asymptote). Adding a “d” parameter would suggest that respondents very high on the latent trait are not guaranteed (i.e. having less than 1.00 probability) to endorse an item.

Interestingly, measures in clinical psychological measurement tend to report high discrimination a_j values (i.e., $a_j > 4$) as measured constructs tend to be conceptually narrow (e.g., fatigue) and include homogeneous item content throughout the scale (Reise

& Waller, 2009). Given that item slopes would increase with excessive item intercorrelations, these results would suggest that the internalizing and externalizing measures have item intercorrelations within an acceptable range (Hambleton, Swaminathan, & Rogers, 1991; Edelen & Reeve, 2007).

Moreover, the category difficulties found in this study span across an acceptable trait range. The metric of a calibration sample should be considered when interpreting the item parameters of the IRT model. It is important to note that the internalizing or externalizing trait level measured reflects the mean of a clinically-referred youth sample, rather than a general Canadian youth norm. In clinical measures, researchers have noted item total scores tended to be positively skewed and threshold parameters are clustered in a restricted range of location parameters, with individuals 2 standard deviations below the mean highly likely to endorse the highest response (Reise & Waller, 2009). Interestingly, threshold parameters in this measure spread across θ and were not located in extreme values, which may reflect the larger sample size used in this study (*N*=3464) compared to other clinical samples and/or items are well-spread across the continuum to adequately capture the trait. For example, within the internalizing scale, the present findings revealed that in terms of individual items, that item "made negative comments" is the "easiest" item because the probability of endorsing the higher categories for the internalizing trait tends to be higher compared to other items. For example, a child/youth with a trait level of -0.460 has a 50% probability of responding "crossing over" from the 0 to 1 category. "Hypervigilance" is the "hardest" item because the probability of "crossing over" from the 0 to 1 category (b_1) for a given trait level tends to be lower (b_1= 1.058).

While one could argue that the measure could be improved by adding items that measure the lower end of θ (i.e., 1 or 2 standard deviations below the mean), the latent traits of interest (i.e., internalizing and externalizing mental health) in this study is a quasi-trait, which is a unipolar construct measuring presence or absence of a trait (e.g., depressed vs. non-depressed; Reise & Waller, 2009). This is in contrast with a bipolar trait, in which both extremes on opposite ends represent variations in two meaningful entities (e.g., depressed vs. happy). Higher standard error (SE) for quasi-trait estimates exist for respondents who are 2 standard deviations (SD) below the mean, while SE for trait estimates are relatively small for respondents 2 SD above the mean, with SE four

times larger for low-trait individuals (Reise & Waller, 2009). Reise and Waller (2009) suggested that this discrepancy can be explained as the low end of a quasi-trait is not a meaningful construct, and as such, there is no need to measure the extreme end of the spectrum in a quasi-trait. One limitation of these measures may be that of any quasi-trait scales, such that evaluating change may be especially difficult with different precision for individuals at different trait levels, as it is difficult to create items that span across the continuum relative to a bipolar trait (Reise & Waller, 2009).

Exploratory bifactor structures were also implemented in this study to allow items free to load onto the general and specified number of group factors. Given that the bifactor model and the two- and three- factor model in the externalizing and internalizing subscale respectively have relatively similar robust goodness-of-fit indices, closer examination of the factor loadings are required to determine fit. In the externalizing measure, there were items that displayed cross-loadings on the group factors that were incoherent, suggesting item parameter estimation distortion and model misfits within the unrestricted models (Reise et al., 2010). Hence, the current results would suggest that the externalizing scale is best interpreted as a two-factor model.

In the internalizing subscale, robust goodness-of-fit indices were approximately equal for the exploratory bifactor model and three-factor model, suggesting both models were equally interpretable. In these data, the results of the bifactor model were quite similar with the exception that the loadings on the group factors were consistently higher in the three-factor model than the bifactor model pattern matrix. Hence, the current results would suggest that the externalizing scale is best interpreted as a bifactor model and a three-factor model.

4.2 Discussion on Criterion Validity

The interRAI ChYMH internalizing and externalizing scales were developed to assist with problem identification as well as monitor changes in clinical status over time. The strong criterion validity suggested that the internalizing and externalizing scales have promising clinical utility. For instance, strong correlations were found between the ChYMH internalizing subscale and the criterion scales of the CBCL internalizing scale, SSIS internalizing scale, and BCFPI managing mood and self-harm scale. Group differences in this study were also similar to previous findings in the literature. Consistent

with previous findings from the CBCL, externalizing scores declined with age, while internalizing scores increased with age (Crijnen, Achenbach, & Verhulst, 1997). Moreover, boys obtained higher externalizing scores, but lower internalizing scores compared to girls (Crijnen et al., 1997). In future studies, researchers are also encouraged to analyze the factorial invariance of the internalizing and externalizing subscales across age (i.e., children and adolescence) and gender groups (i.e., boys and girls).

Achenbach and colleagues (2016) suggested several recommendations for research on and clinical use of measures for internalizing or externalizing problems. First and foremost, even if only one of internalizing or externalizing problems were the construct of interest, these broadband measures are moderately correlated in most samples and should be controlled for in statistical analyses where either scale scores act as a predictor or outcome variable (Achenbach et al., 2016). Total scale scores can provide information for clinical decisions about triage for specialized treatment targeting internalizing, externalizing, or mixture approach (Achenbach, 2017). For example, children/youth could be categorized based on scale scores (e.g., clinical, borderline, normal) and groups can be formed based on the individuals' symptomology (e.g., clinical for both internalizing and externalizing problems; clinical for one of internalizing/externalizing; clinical on neither). Results showed moderate to high overlap between the proactive and reactive aggressions, which was evident in various other studies (e.g., Smeets et al., 2017). Previous studies found reactive aggression is a normal phenomenon for children, but when it does not diminish it may be indicative of severe aggression in older adolescents (Smeets et al., 2017). A meaningful assessment of change in reactive aggression symptoms over time could have clinical utility in problem identification and monitoring.

Of note, a diagnosis of conduct disorder in the DSM-5 includes age-appropriate persistent pattern of violated behavior in which societal norms are violated (APA, 2013). These behaviours can range from a lower limit of 3 to an upper limit of 15 diverse forms of problematic behaviours which could be applicable to the individual (i.e., behaviours related to aggressive to people and animals, destruction of property, deceitful/theft, serious violations of rules). The interRAI ChYMH clinician is encouraged to gather information from other items to further assess for frequency and severity of problematic

externalizing behaviours (e.g., cruelty to animals, involvement in gang, etc.) if a child/youth scores highly on the measure.

Mean scores of the internalizing scale scores were significantly higher in children/youth with the mood/anxiety diagnosis, such that measures that are expected to be related to internalizing symptoms in the DSM are, in fact, adequately capturing the differences between groups. In addition, the effect sizes for the means difference were medium to large, which coincided with previous findings that indicated higher scores on the interRAI ChYMH anhedonia subscale were strongly associated with diagnosis of mood disorder (Stewart et al., 2015). Evidence of divergent validity was also evident in the known-groups validity analysis for the externalizing scale. Effect size were large when the externalizing scale was used to differentiate disruptive behaviour disorders and ADHD. Of note, the externalizing scale outperformed the narrow band scales (e.g., disruptive/aggressive behaviour scale, hyperactive/distraction scale) in differentiating children/youth with or without a disruptive behaviour disorder diagnosis in terms of effect size. Clinicians may use higher scores on the externalizing scale as an indicator to flag for a possible disruptive behaviour disorder diagnosis.

4.3 Limitations and Future Directions

Although this study used a large sample size of clinically referred youth across multiple mental health facilities across the province of Ontario, some limitations within the present study that should be evaluated when interpreting the results. First, the sample of clinically-referred youth were likely not representative of the child/youth population in general. In the absence of a normative sample, scores on the internalizing and externalizing subscales were taken from children/youth with a wide range of physical and mental health problems as standard clinical practice were used. However, the purpose of study was to investigate whether the two scales accurately captured clinically significant internalizing and externalizing conditions from a clinically-referred sample. Thus, the integrity of this study should not be compromised and instead, discriminant ability of both scales is hypothesized to be greater than reported in this study if the diagnostic group was compared to a representative sample, as opposed to a clinically referred sample.

Of concern is that the DSM diagnosis was not a specific type of standardized assessment (i.e., Structured Clinical Interview for DSM-IV) and it was not recorded in the dataset what type of interviews were conducted in each assessment (e.g., structured or unstructured, type of diagnostic assessment, etc.) when children/youths were given the diagnosis. While this may have affected the diagnoses the assessors gave to the child, the interRAI ChYMH assessor is recording a diagnosis that has been made previously from a formal assessment, and not inferring a diagnosis based only on less formal clinical judgment.

Given the transition from the DSM-IV-TR to DSM-5 in the last decade, a discussion on the limitations in using diagnostic criteria from the DSM-IV-TR is warranted. Changes in grouping of symptoms (e.g., ODD symptoms grouped as angry/irritable mood, argumentative/defiant behavior, and vindictiveness), additional specifiers (e.g., limited prosocial emotions specifier in CD), and severity ratings that changed between the DSM-IV-TR and DSM-5 should not affect the results of the present study, given that only presence or absence of the broad category of disorders was evaluated (APA, 2013). However, some changes in specific diagnosis criteria (e.g., the onset of ADHD symptoms can now occur as late as age 12 instead of age 6 years, obsessive-compulsive disorder is no longer an anxiety disorder) may have affected whether a child/youth would have received a diagnosis. Moreover, given changes in diagnostic categories, Axelson and colleagues (2012) indicated that disruptive mood dysregulation disorder (DMDD) was not clearly differentiated from disruptive behaviour disorders, given the unusual large overlap between DMDD and ODD in comorbid childhood mental health disorders. It is unclear whether using DSM-5 criteria, some children/youth in this sample would receive a DMDD diagnosis, which is a mood disorder diagnosis, rather than a disruptive behaviour disorder diagnosis. Future studies should evaluate the effectiveness of the internalizing and externalizing subscales in differentiating DSM-5 criteria childhood mental health disorders.

Future studies should also examine whether an IRT reweighting (i.e., scoring with items weighted by discrimination parameters) would characterize a more accurate representation of the latent construct, rather than scoring with unweighted item sums (Gray-Little, Williams, & Hancock, 1997). However, the obvious limitation would be the

loss of simplicity in scoring for clinicians in the field. Thus, it is recommended that researchers examine the correlation and discrepancy between the unweighted summed score and the internalizing and externalizing θ estimate from the unconstrained model to examine whether the discrepancy is large enough that IRT scoring would outweigh the convenience factor (Gray-Little et al., 1997). IRT methods are becoming increasingly common in psychological measurement; the rigorous measurement could promote better science and clinical practice in the development of new psychological measures and diagnostic assessments (Santor & Ramsay, 1998).

Finally, clinically elevated dimensional internalizing and externalizing scale scores can be identified if scores of children/youth within the clinic were compared to population norms of a same-aged child (Achenbach et al., 2016). The internalizing/externalizing scores obtained would provide clinical utility regarding provision of treatment of children/youth in group programs (e.g., internalizing, externalizing, or mixed problem patterns; Achenbach, 2016). The clinical utility of the measures should be further assessed in future studies.

4.4 Concluding Remarks

Overall, this study confirmed that the internalizing and externalizing subscales are useful for detecting symptoms related to broadband emotional and behavioural disturbances in children/youth. The scales on the interRAI ChYMH were developed as a brief and efficient estimation of the severity of internalizing and externalizing disturbances encountered in a clinical setting. The relative economy of using subscales to detect frequency and severity of symptoms is beneficial to decreasing the assessment burden and involves less clinician time than a fully structured diagnostic interview during intake. Although the interRAI is copyrighted, use of the scale is free of charge to researchers, clinicians, and government subject to the terms of a user license agreement with interRAI (www.interrai.org).

These findings support a growing body of evidence regarding the utility of the interRAI ChYMH for providing a comprehensive profile of children and adolescents' needs, risks, and strengths (Stewart et al., 2015; Stewart & Hamza, 2017, Lau et al., 2017). The interRAI ChYMH internalizing and externalizing subscales can be clinically useful indicators for providing information to clinicians, researchers, policy makers, and

intervention program developers. While traditional diagnostic interviews neglect sub-threshold symptoms, these two broadband measures identify features of disorders based on a dimensional approach to understanding mental health needs. Implications on triaging, prioritizing referrals, and service utilization in the community when utilizing these two psychometrically sound measures to match children's mental health needs will be an area of research needed in the future. With the benefits of early identification of internalizing and externalizing behaviours in guiding prevention and treatment planning, future research should continue to facilitate early screening in primary care to circumvent chronic emotional and emotional and behavioural problems later in life (Dumesnil & Verger, 2009; Kelly, Jorm, & Wright, 2007).

References

Achenbach, T. M., & Edelbrock, C. (1991). Child behavior checklist. Burlington (Vt), 7.

Achenbach, T. M. (1995). Diagnosis, assessment, and comorbidity in psychosocial treatment research. *Journal of Abnormal Child Psychology, 23*(1), 45–65.

Achenbach, T. M., Ivanova, M. Y., Rescorla, L. A., Turner, L. V., & Althoff, R. R. (2016). Internalizing/externalizing problems: Review and recommendations for clinical and research applications. *Journal of the American Academy of Child & Adolescent Psychiatry*, *55*(8), 647–656.

Achenbach, T. M. (2017). Future directions for clinical research, services, and training: evidence-based assessment across informants, cultures, and dimensional hierarchies. *Journal of Clinical Child & Adolescent Psychology*, *46*(1), 159–169.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Washington, DC: Author.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Asendorpf, J. B., & van Aken, M. A. (1999). Resilient, overcontrolled, and undercontroleed personality prototypes in childhood: Replicability, predictive power, and the trait-type issue. *Journal of Personality and Social Psychology, 77*(4), 815–832.

Armiento, J., Hamza, C. A., Stewart, S. L., & Leschied, A. (2016). Direct and indirect forms of childhood maltreatment and nonsuicidal self-injury among clinically-referred children and youth. *Journal of Affective Disorders, 200*, 212–217.

Axelson, D., Findling, R. L., Fristad, M. A., Kowatch, R. A., Youngstrom, E. A., Horwitz, S. M., ... & Gill, M. K. (2012). Examining the proposed disruptive mood dysregulation disorder diagnosis in children in the Longitudinal Assessment of Manic Symptoms study. *The Journal of Clinical Psychiatry*, *73*(10), 1342–1350.

Baglin, J. (2014). Improving your exploratory factor analysis for ordinal data: a demonstration using FACTOR. *Practical Assessment, Research & Evaluation*, *19*(5), 2.

Baiden, P., Stewart, S. L., & Fallon, B. (2017a). The role of adverse childhood experiences as determinants of non-suicidal self-injury among children and adolescents referred to community and inpatient mental health settings. *Child Abuse & Neglect, 69*, 163–176.

Baiden, P., Stewart, S. L., & Fallon, B. (2017b). The mediating effect of depressive symptoms on the relationship between bullying victimization and non-suicidal self-injury among adolescents: Findings from community and inpatient mental health settings in Ontario, Canada. *Psychiatry Research, 255*, 238–247.

Beauchaine, T. P., Hinshaw, S. P., & Pang, K. L. (2010). Comorbidity of attention-deficit/hyperactivity disorder and early-onset conduct disorder: Biological, environmental, and developmental mechanisms. *Clinical Psychology: Science and Practice, 17*(4), 327–336.

Beck, J. S. (2001). *Beck youth inventories*. Psychological Corporation Harcourt Brace.

Beck, J. S., Beck, A. T., & Jolly, J. B. (2001). *BECK Youth Inventories of Emotional & Social Impairment: Depression Inventory for Youth, Anxiety Inventory for Youth, Anger Inventory for Youth, Disruptive Behavior Inventory for Youth, Self-concept Inventory for Youth: Manual*. Psychological Corporation.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238.

Bernabei, R., Landi, F., Onder, G., Liperoti, R., & Gambassi, G. (2008). Second and third generation assessment instruments: the birth of standardization in geriatric care. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences, 6*3(3), 308–313.

Berenbaum, H. (2013). Classification and psychopathology research. *Journal of Abnormal Psychology, 122*(3), 894–901.

Bernabei, R., Gray, L., Hirdes, J., Pei, X., Henrard, J. C., Jonsson, P. V., ... & Carpenter, I. G. (2009). *International Gerontology in Hazzard's Geriatric Medicine and Gerontology 6th Edition:* Halter JB, Ouslander JG, Tinetti ME, Studenski S, High KP, Asthana S (Eds*).,* New York: McGraw Medical, pp 69–96.

Bernabei, R., Landi, F., Onder, G., Liperoti, R., & Gambassi, G. (2008). Second and third generation assessment instruments: the birth of standardization in geriatric care. *The*

Journals of Gerontology. Series A, Biological Sciences and Medical Sciences, 63(3), 308–313.

Blumberg, S. H., & Izard, C. E. (1985). Affective and cognitive characteristics of depression in 10-and 11-year-old children. *Journal of Personality and Social Psychology*, *49*(1), 194.

Bornstein, M. H., Hahn, C. S., & Suwalsky, J. T. (2013). Physically developed and exploratory young infants contribute to their own long-term academic achievement. *Psychological Science*, *24*(10), 1906–1917.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150.

Bufferd, S. J., Dougherty, L. R., Carlson, G. A., Rose, S., & Klein, D. N. (2012). Psychiatric disorders in preschoolers: continuity from ages 3 to 6. *American Journal of Psychiatry, 169*(11), 1157–1164.

Bose-Deakins, J. E., & Floyd, R. G. (2004). A review of the Beck Youth Inventories of emotional and social impairment. *Journal of School Psychology*, *42*(4), 333–340.

Boyle, M. H., Cunningham, C. E., Georgiades, K., Cullen, J., Racine, Y., & Pettingill, P. (2009). The Brief Child and Family Phone Interview (BCFPI): 2. Usefulness in screening for child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry*, *50*(4), 424–431.

Campbell, S. B., Shaw, D. S., & Gilliom, M. (2000). Early externalizing behavior problems: Toddlers and preschoolers at risk for later maladjustment. *Development and psychopathology*, *12*(3), 467–488.

Canadian Mental Health Association (2017). Mental Illness and Addictions: Facts and Statistics, Retrieved from www.camh.ca/News_events/Key_CAMH_facts_for.../addictionmentalhealthstatistics.html

Caspi, A., Henry, B., McGee, R. O., Moffitt, T. E., & Silva, P. A. (1995). Temperamental origins of child and adolescent behavior problems: From age three to age fifteen. *Child Development*, *66*(1), 55–68.

Capaldi, D. M., & Stoolmiller, M. (1999). Co-occurrence of conduct problems and depressive symptoms in early adolescent boys: III. Prediction to young-adult adjustment. *Development and Psychopathology, 11*(1), 59–84.

Children's Mental Health Ontario (2016). Ontario's children waiting up to 1.5 years for urgently needed mental healthcare. Retrieved from https://cmho.org/blog/article2/6519717-ontario-s-children-waiting-up-to-1-5-years-for-urgently-needed-mental-healthcare-3

Chan, C. L., Lai, C. K., & Chi, I. (2014). Initial validation of the Chinese interRAI Mental Health in people with psychiatric illness. *International Journal of Psychiatry in Clinical Practice, 18*(3), 182–189.

Child, D. (1990). *The essentials of factor analysis*. Cassell Educational.

Choi, J., Kim, S., Chen, J., & Dannels, S. (2011). A comparison of maximum likelihood and Bayesian estimation for polychoric correlation using Monte Carlo simulation. *Journal of Educational and Behavioral Statistics*, *36*(4), 523–549.

Cicchetti, D. V., & Sparrow, S. S. (1981). Developing criteria for the rating of specific items in a given inventory. *American Journal of Mental Deficiency*, *86*, 127–137.

Costello, A.B., & Jason, O. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation, 10*(7), 1–9.

Costello, E. J., Mustillo, S., Erkanli, A., Keeler, G., & Angold, A. (2003). Prevalence and development of psychiatric disorders in childhood and adolescence. *Archives of General Psychiatry, 60*(8), 837–844.

Cramer, D. (1998). *Fundamental statistics for social research: step-by-step calculations and computer techniques using SPSS for Windows*. Psychology Press.

Cramer, D., & Howitt, D. L. (2004). *The Sage dictionary of statistics: a practical resource for students in the social sciences*. Sage.

Crijnen, A. A., Achenbach, T. M., & Verhulst, F. C. (1997). Comparisons of problems reported by parents of children in 12 cultures: total problems, externalizing, and internalizing. *Journal of the American Academy of Child & Adolescent Psychiatry*, *36*(9), 1269–1277.

Cunningham, C. E., Boyle, M. H., Hong, S., Pettingill, P., & Bohaychuk, D. (2009). The Brief Child and Family Phone Interview (BCFPI): 1. Rationale, development, and description of a computerized children's mental health intake and outcome assessment tool. *Journal of Child Psychology and Psychiatry*, *50*(4), 416–423.

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 425–438.

Dozois, D. J., Dobson, K. S., & Ahnberg, J. L. (1998). A psychometric evaluation of the Beck Depression Inventory–II. *Psychological Assessment*, *10*(2), 83.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, *19*(2), 143–166.

Dumesnil, H., & Verger, P. (2009). Public awareness campaigns about depression and suicide: a review. *Psychiatric Services*, *60*(9), 1203–1213.

Eisenberg, N., Cumberland, A., Spinrad, T. L., Fabes, R. A., Shepard, S. A., Reiser, M., ... & Guthrie, I. K. (2001). The relations of regulation and emotionality to children's externalizing and internalizing problem behavior. *Child Development*, *72*(4), 1112–1134.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement, 20*, 201–212.

Embretson, S. E., & Hershberger, S. (1999). The new rules of measurement: What every psychologist and educator should know. Mahwah, NJ: Erlbaum.

Fact sheet no. 369: Depression. World Health Organization, 2017. Web site. http://www.who.int/mediacentre/factsheets/fs369/en/. Accessed June 01, 2017. Updated 2017.

Faraone, S. V., Biederman, J., Mennin, D., Russell, R., & Tsuang, M. T. (1998). Familial subtypes of attention deficit hyperactivity disorder: a 4-year follow-up study of children from antisocial-ADHD families. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *39*(7), 1045–1053.

Ferrando, P.J. & Lorenzo-Seva, U. (2013). Unrestricted item factor analysis and some relations with item response theory. Technical Report. Department of Psychology, Universitat Rovira i Virgili, Tarragona.

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema*, *29*(2), 236–240.

Fleiss, J. L. (1981). Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement*, *5*(1), 105–112.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466.

Foebel, A. D., Hirdes, J. P., Heckman, G. A., Kergoat, M. J., Patten, S., & Marrie, R. A. (2013). Diagnostic data for neurological conditions in interRAI assessments in home care, nursing home and mental health care settings: a validity study. *BMC Health Services Research, 13*(1), 457.

Forbes, M. K., Tackett, J. L., Markon, K. E., & Krueger, R. F. (2016). Beyond comorbidity: Toward a dimensional and hierarchical approach to understanding psychopathology across the life span. *Development and Psychopathology*, *28*(4pt1), 971–986.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of Personality and Social Psychology, 78*(2), 350–365.

Fries, B. E., Simon, S. E., Morris, J. N., Flodstrom, C., & Bookstein, F. L. (2001). Pain in US nursing homes: validating a pain scale for the minimum data set. *The Gerontologist, 41*(2), 173–179.

Garrido, L.E., Abad, J.J. and Ponsoda, V. (2013) A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods, 18*(4), 454–474.

Germain, M. L. (2006). Stages of Psychometric Measure Development: The Example of the Generalized Expertise Measure (GEM). *Online Submission*.

Government of Canada (2006). *The human face of mental health and mental illness in Canada*. Ottawa: Minister of Public Works and Government Services Canada. Retrieved from http://www.phac-aspc.gc.ca/publicat/human-humain06/pdf/human_face_e.pdf

Gray-Little, B., Williams, V. S., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 23*(5), 443–451.

Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system (SSRS)*. American Guidance Service.

Gresham, F., & Elliott, S. N. (2008). Social skills improvement system (SSIS) rating scales. *Bloomington, MN: Pearson Assessments*.

Gresham, F. M., Elliott, S. N., Vance, M. J., & Cook, C. R. (2011). Comparability of the Social Skills Rating System to the Social Skills Improvement System: Content and psychometric comparisons across elementary and secondary age levels. *School Psychology Quarterly*, *26*(1), 27.

Gresham, F. M., Elliott, S. N., Cook, C. R., Vance, M. J., & Kettler, R. (2010). Cross-informant agreement for ratings for social skill and problem behavior ratings: An investigation of the Social Skills Improvement System—Rating Scales. *Psychological Assessment*, *22*(1), 157–166.

Griffith, A. K. (2010). Committee on the prevention of mental disorders and substance abuse among children, youth, and young adults, National Research Council and the Institute of Medicine: preventing mental, emotional, and behavioral disorders among young people: progress and possibilities. *Journal of Child and Family Studies, 19*(5), 675–676.

Grant, J. S., & Davis, L. L. (1997). Selection and use of content experts for instrument development. *Research in Nursing & Health*, *20*(3), 269-274.

Gresham, F. M., Elliott, S. N., & Kettler, R. J. (2010). Base rates of social skills acquisition/performance deficits, strengths, and problem behaviors: An analysis of the Social Skills Improvement System—Rating Scales. *Psychological Assessment*, *22*(4), 809–815.

Hambleton, RK.; Swaminathan, H., & Rogers, WH. (1991). Fundamentals of item response theory. Newbury Park: Sage Publications.

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238.

Hébert, R., Durand, P. J., Dubuc, N., Tourigny, A., & PRISMA Group. (2003). Frail elderly patients. New model for integrated service delivery. *Canadian Family Physician, 49*(8), 992–997.

Heckman, G. A., Gray, L. C., & Hirdes, J. P. (2013). Addressing health care needs for frail seniors in Canada: the role of interRAI instruments. UWSpace. http://hdl.handle.net.proxy1.lib.uwo.ca/10012/11701

Hinshaw, S. P. (1987). On the distinction between attentional deficits/hyperactivity and conduct problems/aggression in child psychopathology. *Psychological Bulletin*, *101*(3), 443.

Hinkin, T. R., Tracey, J. B., & Enz, C. A. (1997). Scale construction: Developing reliable and valid measurement instruments. *Journal of Hospitality & Tourism Research*, *21*(1), 100–120.

Hirdes, J. P., Fries, B. E., Morris, J. N., Steel, K., Mor, V., Frijters, D., ... & Smith, T. (1999). Integrated health information systems based on the RAI/MDS series of instruments. In *Healthcare management forum* (Vol. 12, No. 4, pp. 30-40). Sage CA: Los Angeles, CA: SAGE Publications.

Hirdes, J. P., Smith, T. F., Rabinowitz, T., Yamauchi, K., Pérez, E., Telegdi, N. C., ... & Fries, B. E. (2002). The resident assessment instrument-mental health (RAI-MH): Inter-rater reliability and convergent validity. *The Journal of Behavioral Health Services and Research, 29*(4), 419–432.

Hirdes, J. P., Frijters, D. H., & Teare, G. F. (2003). The MDS-CHESS scale: a new measure to predict mortality in institutionalized older people. *Journal of the American Geriatrics Society, 51*(1), 96–100.

Hirdes, J. P., Fries, B. E., Morris, J. N., Ikegami, N., Zimmerman, D., Dalby, D. M., ... & Jones, R. (2004). Home care quality indicators (HCQIs) based on the MDS-HC. *The Gerontologist, 44*(5), 665–679.

Hirdes, J. P., Mitchell, L., Maxwell, C. J., & White, N. (2011). Beyond the 'iron lungs of gerontology': using evidence to shape the future of nursing homes in Canada. *Canadian Journal on Aging, 30*(3), 371–390.

Hirdes, J., Curtin-Telegdi, N., Mathias, K., Perlman, C. M., Saarela, T., Kolbeinsson, H., ... & Martin, L. (2011). InterRAI mental health clinical assessment protocols (CAPs) for use with community and hospital-based mental health assessment instruments. Version 9.1.

Hirdes, J. P., Ljunggren, G., Morris, J. N., Frijters, D. H., Soveri, H. F., Gray, L., ... & Gilgen, R. (2008). Reliability of the interRAI suite of assessment instruments: a 12-country study of an integrated health information system. *BMC Health Services Research, 8*(1), 277.

Hodges, K., & Wong, M. M. (1996). Psychometric characteristics of a multidimensional measure to assess impairment: The Child and Adolescent Functional Assessment Scale. *Journal of Child and Family Studies, 5*(4), 445–467.

Holgado–Tello, F.P., Chacón–Moscoso, S., Barbero–García, I. and Vila–Abad E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality and Quantity*, 44(1), 153–166.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424–453.

Huesmann, L. R., Eron, L. D., & Dubow, E. F. (2002). Childhood predictors of adult criminality: are all risk factors reflected in childhood aggressiveness?. *Criminal Behaviour and Mental Health, 12*(3), 185–208.

Jackson, D.N. (1970). A sequential system for personality scale development. In C.D. 96). New York: Academic Press.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*(1), 2.

Kagan, J., Snidman, N., Zentner, M., & Peterson, E. (1999). Infant temperament and anxious symptoms in school age children. *Development and Psychopathology, 11*(2), 209–224.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795.

Kelley, T. L. (1935). An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences, 21*(9), 554–559.

Kelly, C. M., Jorm, A. F., & Wright, A. (2007). Improving mental health literacy as a strategy to facilitate early intervention for mental disorders. *Medical Journal of Australia, 187*(7), S26.

Keltner, D., Moffit, T. E., & Stouthamer-Loeber, M. (1995). Facial expressions of emotion and psychopathology in adolescent boys. *Journal of Abnormal Psychology, 104*(4), 644–652.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62*(6), 593–602.

Kim, H., Jung, Y. I., Sung, M., Lee, J. Y., Yoon, J. Y., & Yoon, J. L. (2015). Reliability of the interRAI long term care facilities (LTCF) and interRAI home care (HC). *Geriatrics & Gerontology International, 15*(2), 220–228.

Kline, R. B. (2011). Principles and practices of structural equation modeling (3rd ed.). New York, NY: Guilford Press.

Knapp, M. (1997). Economic evaluations and interventions for children and adolescents with mental health problems. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 38*, 3–25.

Krueger, R. F., & DeYoung, C. G. (2016). The RDoC initiative and the structure of psychopathology. *Psychophysiology*, *53*(3), 351–354.

Krueger, R. F., & Eaton, N. R. (2012). Structural validity and the classification of mental disorders. *Philosophical Issues in Psychiatry II: Nosology*, 2, 199.

Krueger, R. F., Markon, K. E., Patrick, C. J., Benning, S. D., & Kramer, M. D. (2007). Linking antisocial behavior, substance use, and personality: An integrative quantitative model of the adult externalizing spectrum. *Journal of Abnormal Psychology, 116*, 645–666.

Lahey, B. B., Applegate, B., Waldman, I. D., Loft, J. D., Hankin, B. L., Rick, J. (2004). The structure of child and adolescent psychopathology: Generating new hypotheses. *Journal of Abnormal Psychology, 113*, 358–385.

Lahey, B. B., Waldman, I. D., & McBurnett, K. (1999). Annotation: The development of antisocial behavior: An integrative causal model. *The Journal of Child Psychology and Psychiatry and Allied Disciplines, 40*(5), 669–682.

Lau, C., Stewart, S.L., Saklofske, D. H., Tremblay, P.F., & Hirdes, J. Validation of the interRAI-ChYMH Aggressive Disruptive Behaviour Scale and Hyperactivity-Distractibility Scale. Manuscript under revision in Child Psychiatry and Human Development.

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949.

Liu, J. (2004). Childhood externalizing behavior: theory and implications. *Journal of Child and Adolescent Psychiatric Nursing*, *17*(3), 93–103.

Liu, J. Y., Chi, I., Chan, K. S., Lai, C. K., & Leung, A. Y. (2015). The reliability and validity of the pain items of the Hong Kong version interRAI community health assessment for community-dwelling elders in Hong Kong. *Journal of Clinical Nursing, 15*(24), 2352–2354.

Lesse S. (1983). Masked depression. *Current Psychiatric Therapies, 22,* 81–87.

Lorenzo-Seva, U. (1999). Promin: A method for oblique factor rotation. *Multivariate Behavioral Research*, *34*(3), 347–365.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods*, *38*(1), 88–91.

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*(2), 340–364.

Lorenzo-Seva, U. (2013). Why rotate my data using Promin? Technical Report. Department of Psychology, Universitat Rovira i Virgili, Tarragona.

Lorenzo-Seva, U., & Ferrando, P. J. (2013). FACTOR 9.2: A Comprehensive Program for Fitting Exploratory and Semiconfirmatory Factor Analysis and IRT Models. *Applied Psychological Measurement*, *37*(6), 497–498.

Luby, J. L., Heffelfinger, A. K., Mrakotsky, C., Brown, K. M., Hessler, M. J., Wallis, J. M., & Spitznagel, E. L. (2003). The clinical picture of depression in preschool children. *Journal of the American Academy of Child & Adolescent Psychiatry*, *42*(3), 340–348.

Luby, J. L., Si, X., Belden, A. C., Tandon, M., & Spitznagel, E. (2009). Preschool depression: Homotypic continuity and course over 24 months. *Archives of General Psychiatry, 66*(8), 897–905.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*(6), 382–386.

Martin, L., Hirdes, J. P., Fries, B. E., & Smith, T. F. (2007). Development and psychometric properties of an assessment for persons with intellectual disability—the interRAI ID. *Journal of Policy and Practice in Intellectual Disabilities, 4*(1), 23–29.

Mannuzza, S., Klein, R. G., Bessler, A., Malloy, P., & LaPadula, M. (1993). Adult outcome of hyperactive boys: Educational achievement, occupational rank, and psychiatric status. *Archives of General Psychiatry, 50*(7), 565–576.

Mesman, J., & Koot, H. M. (2000). Child-reported depression and anxiety in preadolescence: I. Associations with parent-and teacher-reported problems. *Journal of the American Academy of Child & Adolescent Psychiatry, 39*(11), 1371–1378.

Minichiello, V., Aroni, R., Timewell, E., & Alexander, L. (1995). In-depth interviewing 2nd ed.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*(2), 177–195.

Morris, J. N., Carpenter, I., Berg, K., & Jones, R. N. (2000). Outcome measures for use with home care clients. *Canadian Journal on Aging, 19*(S2), 87–105.

Nishtala, P. S., & Jamieson, H. A. (2017). New Zealand's interRAI: A Resource For Examining Health Outcomes in Geriatric Pharmacoepidemiology. *Journal of the American Geriatrics Society*, 65(4), 876–877.

Nunnally, J. C., & Bernstein, I. H. (1994). Validity. *Psychometric Theory*, 99–132.

Office of the Auditor General of Ontario (2016). Annual report 2016, volume 1. Toronto: Queen's Printer for Ontario. Retrieved from: http://www.camh.ca/en/hospital/about_camh/newsroom/for_reporters/Pages/addictionmentalhealthstatistics.aspx

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443-460.

Osborne, J.W. (2014). Best practices in exploratory factor analysis. Marston Gate: Amazon. co. uk, Ltd.

Patton, M. (2002). Utilization-focused evaluation. *Evaluation Models*, 425–438.

Perlman, C. M., & Hirdes, J. P. (2008). The aggressive behavior scale: a new scale to measure aggression based on the minimum data set. *Journal of the American Geriatrics Society*, *56*(12), 2298–2303.

Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*(5), 489–497.

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health, 30*(4), 459–467.

Poss, J., Murphy, K. M., Woodbury, M. G., Orsted, H., Stevenson, K., Williams, G., ... & Hirdes, J. P. (2010). Development of the interRAI Pressure Ulcer Risk Scale (PURS) for use in long-term care and home care settings. *BMC Geriatrics, 10*(1), 67.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21–33.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*(4), 401–412.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*(1), 25–36.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement*, *14*, 45–58.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*(6), 544–559

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*(1), 19–31.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*(4), 401–412.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25–36.

Reckase, M. D. (2006). 18 Multidimensional Item Response Theory. Handbook of statistics, 26, 607–642.

Ribbe, M., Jónsson, P. V., Bernabei, R., Onder, G., & Van Marum, R. J. (2012). InterRAI assessment instruments: Basics, applications and research. *European Geriatric Medicine, 3*, S27.

Robins, R. W., John, O. P., Caspi, A., Moffitt, T. E., & Stouthamer-Loeber, M. (1996). Resilient, overcontrolled, and undercontrolled boys: Three replicable personality types. *Journal of Personality and Social psychology*, *70*(1), 157.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement,* 17 (4, Pt. 2), 386–415.

Samejima, F. (1996). Graded response model. In *W. J. van der Linden & R. K. Hambleton* (Eds.), Handbook of modern item response theory (pp. 85–100). New York: Springer

Santor, D. A., & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment*, *10*(4), 345–359.

Salih, S. A., Paul, S., Klein, K., Lakhan, P., & Gray, L. (2012). Screening for delirium within the interRAI acute care assessment system. *The Journal of Nutrition, Health & Aging*, *16*(8), 1–6.

Schriesheim, C. A., Powers, K. J., Scandura, T. A., Gardiner, C. C., & Lankau, M. J. (1993). Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management*, *19*(2), 385–417.

Steer, R. A., Kumar, G., Beck, J. S., & Beck, A. T. (2001). Evidence for the construct validities of the Beck Youth Inventories with child psychiatric outpatients. *Psychological Reports*, *89*(3), 559–565.

Steer, R. A., Kumar, G., Beck, A. T., & Beck, J. S. (2005). Dimensionality of the Beck Youth Inventories with child psychiatric outpatients. *Journal of Psychopathology and Behavioral Assessment*, *27*(2), 123–131.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173–180.

Stewart, S., Hirdes, J., Curtin-Telegdi, N., Perlman, C. M., McKnight, M., MacLeod, K., ... & Berg, K. (2015). interRAI Child and Youth Mental Health (ChYMH) Assessment Form and User's Manual: For use with In-patient and Community-based Assessments.

Stewart, S. L., & Hamza, C. A. (2017). The Child and Youth Mental Health Assessment (ChYMH): An examination of the psychometric properties of an integrated assessment developed for clinically referred children and youth. *BMC Health Services Research, 17*(1), 82.

Smith, J.P., & Smith, G.C. (2010). Long-term economic costs of psychological problems during childhood. *Social Science & Medicine, 71* (1), 110–115.

Smeets, K. C., Oostermeijer, S., Lappenschaar, M., Cohn, M., van der Meer, J. M. J., Popma, A., ... & Buitelaar, J. K. (2017). Are proactive and reactive aggression

meaningful distinctions in adolescents? A variable-and person-based approach. *Journal of Abnormal Child Psychology*, *45*(1), 1–14.

Smetanin, P., Stiff, D., Briante, C., Adair, C., Ahmad, S., & Khan, M. (2011). *The life and economic impact of major mental illnesses in Canada: 2011 to 2041.* RiskAnalytica, on behalf of the Mental Health Commission of Canada. http://strategy.mentalhealthcommission.ca/the-facts/

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101.

Southam-Gerow, M. A., & Kendall, P. C. (2002). Emotion regulation and understanding: Implications for child psychopathology and therapy. *Clinical Psychology Review, 22*(2), 189–222.

Tabachnick, B. G., & Fidell, L. S. (1989). Using Multivariate Statistics, Harper Collins Publishers. *New York*.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209–220.

Travers, C., Byrne, G. J., Pachana, N. A., Klein, K., & Gray, L. (2013). Validation of the interRAI Cognitive Performance Scale against independent clinical diagnosis and the Mini-Mental State Examination in older hospitalized patients. The journal of nutrition, health & aging, 17(5), 435.

The InterRAI Organization: Who We Are. interRAI. 2017. Web site. http://www.interrai.org/organization.html. Accessed January 30, 2017.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Waltz, C. F., & Bausell, B. R. (1981). *Nursing research: design statistics and computer analysis*. Davis FA.

Weeks, M., Ploubidis, G. B., Cairney, J., Wild, T. C., Naicker, K., & Colman, I. (2016). Developmental pathways linking childhood and adolescent internalizing, externalizing, academic competence, and adolescent depression. *Journal of Adolescence, 51*, 30–40.

Weiss, G., Hechtman, L., Milroy, T., & Perlman, T. (1985). Psychiatric status of hyperactives as adults: a controlled prospective 15-year follow-up of 63 hyperactive children. *Journal of the American Academy of Child Psychiatry*, *24*(2), 211–220.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*(3), 291–298.

Wilkinson, P. (2009). Conceptualization about internalizing problems in children and adolescents. *Ciência & Saúde Coletiva*, *14*(2), 373–381.

Wolff, H. G., & Preising, K. (2005). Exploring item and higher order factor structure with the Schmid-Leiman solution: Syntax codes for SPSS and SAS. Behavior Research Methods, 37(1), 48–58.

Table 1. Internalizing Item Means (Standard Deviation) and Item Variance for the Full Sample (Aged 4 to 18 years), Children (Aged 4 to 11 years), and Adolescents (Aged 12 to 18 years).

Item	Item Mean (SD) for full sample	Item Variance for full sample	Item Mean (SD) for Children	Item Variance for Children	Item Mean (SD) for Adolescence	Item Variance for Adolescence
1. Episodes of panic	.78 (1.107)	1.226	.74 (1.144)	1.310	.81 (1.074)	1.153
2. Expressions of hopelessness	.75 (1.116)	1.246	.53(1.010)	1.020	.93 (1.167)	1.363
3. Nightmares	.61 (.975)	**.951**	.72 (1.029)	1.058	.52 (.918)	.842
4. Lack of interest in social interaction	.86 (1.337)	1.787	.74(1.276)	1.629	.97 (1.377)	1.696
5. Hyper-vigilance	.46 (1.025)	1.050	.46(1.033)	1.067	.46 (1.019)	1.896
6. Crying, tearfulness	1.29 (1.288)	1.659	1.52(1.368)	1.872	1.09 (1.182)	1.396
7. Sad, pained, or worried facial expressions	1.65 (1.435)	2.060	1.65(1.458)	2.127	1.64 (1.416)	2.006
8. Intrusive thoughts or flashbacks	.46(.929)	**.863**	.36 (.831)	.690	.54 (.998)	.995
9. Self-deprecation	1.23 (1.322)	1.746	1.24 (1.356)	1.838	1.22 (1.292)	1.670
10.Irritability	2.24 (1.475)	2.176	2.49 (1.454)	2.114	2.03 (1.461)	2.134
11.Anhedonia	.66 (1.149)	1.319	.50(1.033)	1.068	.80 (1.221)	1.492
12.Expressions of guilt or shame	.76 (1.125)	1.265	.70(1.113)	1.239	.82 (1.132)	1.282
13.Unrealistic fears	.98 (1.366)	1.866	1.07 (1.406)	1.976	.92 (1.328)	1.763
14.Made negative statements	1.31 (1.364)	1.860	1.31(1.389)	1.930	1.30 (1.342)	1.801
15.Repetitive anxious complaints/concerns	1.52 (1.574)	2.478	1.67(1.624)	2.636	1.40 (1.521)	2.312
16.Obsessive thoughts	1.04 (1.460)	2.131	1.01 (1.457)	2.153	1.07 (1.453)	2.111
17.Withdrawal from activities of interest	.52 (1.012)	1.025	.32 (.810)	.655	.68 (1.130)	1.276
18.Decreased Energy	1.03 (1.378)	1.899	.71(1.205)	1.452	1.31 (1.455)	2.116
19.Compulsive behaviour	.72 (1.327)	1.761	.79(1.385)	1.918	.65 (1.273)	1.621
20.Re-enactment through play of traumatic events	.09 (0.411)	**.169**	.10 (.448)	.201	.08 (.377)	.142
21.Lack of motivation	1.02 (1.423)	2.024	.70 (1.259)	1.584	1.29 (1.496)	2.237
22.Repetitive Health complaints	.86 (1.298)	1.684	.87 (1.293)	1.671	.85 (1.302)	1.696

Table 2. Externalizing Item Means (Standard Deviation) and Item Variance for the Full Sample (Aged 4 to 18), Children (Aged 4 to 11), and Adolescents (Aged 12 to 18).

Item Name and Description	Item Mean (SD) for full sample	Item Variance	Item Mean (SD) for Children	Item Variance for Children	Item Mean (SD) for Adolescence	Item Variance for Adolescence
1. Destructive behaviour towards property	.76 (1.041)	1.084	1.00 (1.173)	1.375	.56 (.864)	.746
2. Repetitive Lying	.84 (1.284)	1.648	.94 (1.357)	1.841	.75 (1.212)	1.468
3. Elopement attempts/threats	.82 (1.399)	1.957	.84 (1.408)	1.982	.80 (1.392)	1.937
4. Demonstrates limited understanding of consequences to behaviour	.86 (1.344)	1.806	1.13 (1.461)	2.135	.66 (1.217)	1.480
5. Preoccupation of violence – e.g., depictions of violence	.32 (1.079)	1.164	.35 (1.101)	1.213	.30 (1.059)	1.123
6. Easily distracted	2.44 (1.538)	2.364	2.72 (1.451)	2.107	2.21 (1.569)	2.463
7. Impulsivity	1.97 (1.629)	2.655	2.49 (1.550)	2.404	1.53(1.566)	2.451
8. Physical Abuse	.92 (1.166)	1.360	1.37 (1.306)	1.706	.55 (.875)	.765
9. Disorganization	2.04 (1.657)	2.745	2.19 (1.645)	2.705	1.92 (1.657)	2.746
10. Bullying peers	.67 (1.392)	1.938	.78 (1.510)	2.281	.57 (1.276)	1.628
11. Fire-setting or misuse of ignition	.20 (.676)	**.457**	.13 (.549)	.302	.27 (.762)	.581
12. Argumentativeness	1.87 (1.554)	2.416	2.20 (1.544)	2.384	1.58 (1.506)	2.267
13. Hyperactivity – excessive level of activity	1.64 (1.666)	2.774	2.23 (1.655)	2.740	1.15 (1.505)	2.264
14. Socially inappropriate or disruptive behaviour	.71 (1.188)	1.411	.95 (1.327)	1.761	.50 (1.012)	1.025
15. Intimidation of others or threatened violence	1.24 (1.787)	3.193	1.37 (1.902)	3.617	1.13 (1.676)	2.810

16. Outburst of anger –	1.63 (1.400)	1.960	2.06 (1.397)	1.950	1.28 (1.301)	1.692
17. Violent Ideation – e.g., reports of premeditated thoughts, statements	.47 (1.205)	1.452	.45 (1.191)	1.419	.48 (1.216)	1.479
18. Violence to others	.71 (1.419)	2.013	.89 (1.657)	2.744	.55 (1.160)	1.345
19. Cruelty to animals	.24 (.866)	**.749**	.33 (1.031)	1.062	.16 (.687)	.472
20. Verbal Abuse	1.40 (1.439)	2.070	1.63 (1.494)	2.233	1.21 (1.362)	1.855
21. Defiant behaviour– active persistent refusal to comply with reasonable requests by others	1.75 (1.541)	2.373	2.17 (1.520)	2.311	1.40 (1.468)	2.156
22. Stealing –e.g., theft from family, shoplifting	.68 (1.312)	1.720	.63 (1.306)	1.705	.72 (1.316)	1.731
23. Expressions supportive of criminal activity (e.g., “it’s only a crime if you get caught”)	.16 (.609)	**.370**	.11 (.489)	.239	.21 (.690)	.476

Table 3. I-CVI, Pc, and Modified Kappa (k^*) for Items Evaluated for Children aged 4 to 11 years in the Internalizing Scale during Expert Content Validation.

Item Name	I-CVI for children	Pc	k^*	Evaluation
1. Episodes of panic	0.867	0.00320	0.86624	Excellent
2. Expressions of hopelessness	0.933	0.00046	0.93330	Excellent
3. Nightmares	0.800	0.01389	0.79718	Excellent
4. Lack of interest in social interaction	0.667	0.09164	0.63304	Good
5. Hyper-vigilance	0.733	0.04166	0.72174	Good
6. Crying, tearfulness	0.867	0.00320	0.86624	Excellent
7. Sad, pained, or worried facial expressions	0.933	0.00046	0.93330	Excellent
8. Intrusive thoughts or flashbacks	0.667	0.09164	0.63304	Good
9. Self-deprecation	0.933	0.00046	0.93330	Excellent
10.Irritability	0.667	0.09164	0.63304	Good
11.Anhedonia	0.867	0.00320	0.86624	Excellent
12.Expressions of guilt or shame	0.933	0.00046	0.93330	Excellent
13.Unrealistic fears	0.867	0.00320	0.86624	Excellent
14.Made negative statements	0.800	0.01389	0.79718	Excellent
15.Repetitive anxious complaints/concerns	1	3.05176E-05	1	Excellent
16.Obsessive thoughts	0.667	0.09164	0.63304	Good
17.Withdrawal from activities of interest	0.933	0.00046	0.93330	Excellent
18.Decreased Energy	0.800	0.01389	0.79718	Excellent
19.Compulsive behaviour	0.733	0.04166	0.72174	Good
20.Re-enactment through play of traumatic events	0.600	0.15274	0.52789	Fair
21.Lack of motivation	1	3.05176E-05	1	Excellent
22.Repetitive Health complaints	0.800	0.01389	0.79718	Excellent

Table 4. I-CVI, Pc, and Modified Kappa (k^*) for Items Evaluated for Adolescents aged 12 to 18 years in the Internalizing Scale during Expert Content Validation.

Item Name	I-CVI for Adolescents	Pc	k^*	Evaluation
1. Episodes of panic	1	3.05E-05	1	Excellent
2. Expressions of hopelessness	1	3.05E-05	1	Excellent
3. Nightmares	0.73333	0.04166	0.72174	Good
4. Lack of interest in social interaction	0.86667	0.00320	0.86624	Excellent
5. Hyper-vigilance	0.80000	0.01389	0.79718	Excellent
6. Crying, tearfulness	0.86667	0.00320	0.86624	Excellent
7. Sad, pained, or worried facial expressions	0.80000	0.01389	0.79718	Excellent
8. Intrusive thoughts or flashbacks	0.66667	0.09164	0.63304	Good
9. Self-deprecation	1	3.05E-05	1	Excellent
10.Irritability	0.86667	0.00320	0.86624	Excellent
11.Anhedonia	1	3.05E-05	1	Excellent
12.Expressions of guilt or shame	1	3.05E-05	1	Excellent
13.Unrealistic fears	0.86667	0.00320	0.86624	Excellent
14.Made negative statements	1	3.05E-05	1	Excellent
15.Repetitive anxious complaints/concerns	0.93333	0.00046	0.93330	Excellent
16.Obsessive thoughts	0.8	0.01389	0.79718	Excellent
17.Withdrawal from activities of interest	1	3.05E-05	1	Excellent
18.Decreased Energy	1	3.05E-05	1	Excellent
19.Compulsive behaviour	0.86667	0.00320	0.86624	Excellent
20.Re-enactment through play of traumatic events	0.2	0.01389	0.18874	Poor
21.Lack of motivation	1	3.05E-05	1	Excellent
22.Repetitive Health complaints	1	3.05E-05	1	Excellent

Table 5. Demographics for Children/Youth Assessed using the interRAI ChYMH between 2015 to 2016 (*N*=2535).

		Number (% of sample)
Gender	Male	1461 (57.6%)
	Female	1074 (42.7%)
Patient type	Inpatient	160 (6.3%)
	Outpatient	2375 (93.7%)
Assessment Method	Person	1873 (73.9%)
	Phone	661 (26.1%)
	Video	1 (.0%)
Legal Guardianship	Both Parents	1391 (54.9%)
	Only Mother	753 (29.7%)
	Only Father	106 (4.2%)
	Neither parent but other relative(s) or non-relative(s)	133 (5.2%)
	Child Protection Agency (e.g., CAS)	134 (5.3%)
	Public Guardian	3 (.1%)
	Youth Cares for Self	15 (.6%)
DSM-IV Diagnosis	Reactive Attachment Disorder	46/1894 (2.4%)
	Attention deficit Hyperactive Disorder	1045/2019 (51.8%)
	Disruptive Behaviour Disorders (i.e., ODD, CD)	559/1959 (28.5%)
	Learning/Communication Disorder	508/1940 (26.2%)
	Autism Spectrum Disorder	239/1922 (12.4%)
	Substance-related disorders	64/1926 (3.4%)
	Schizophrenia and other psychotic disorders	19/1932 (0.98%)
	Mood Disorders	410/1885 (21.8%)
	Anxiety Disorders	917/1920 (47.8%)
	Eating Disorders	50/1923 (2.6%)
	Sleep Disorders	74/1912 (3.8%)
	Adjustment Disorders	72/1898 (3.8%)

Table 6. Robust Model-Fit Statistics for 12-item Internalizing Subscale with BC Bootstrap 95% Confidence interval.

	1-Factor Model	**3-Factor Model**	**Exploratory Bifactor Model**
Root Mean Square Error of Approximation (RMSEA)	0.137 (0.1267, 0.1437)	0.050 (0.0431, 0.0533)	0.054 (0.0349, 0.0584)
Non-Normed Fit Index (NNFI Tucker & Lewis)	0.873 (0.853, 0.895)	0.983 (0.980, 0.988)	0.981 (0.977, 0.991)
Comparative Fit Index (CFI)	0.896 (0.880, 0.914)	0.992 (0.990, 0.994)	0.993 (0.992, 0.997)
Schwarz's Bayesian Information Criterion (BIC)	2794.230 (2438.111, 3077.030)	615.282 (564.302, 646.561)	669.019 (568.520)
Goodness of Index (GFI)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Adjusted Goodness of Index (GFI)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Goodness of Fit without diagonal values	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Adjusted Goodness of Fit without diagonal values	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)

Table 7. Internalizing Scale Rotated Loading Matrix using Robust Diagonally Weighted Least Squares Estimator and Promin Rotation, with loadings greater than |0.300| bolded.

Items	**Anhedonia (F1)**	**Anxiety (F2)**	**Depression (F3)**
Repetitive anxious complaints/concerns	-0.081 [-0.155, -0.011]	**0.697 [0.634, 0.769]**	0.050 [-0.026, 0.123]
Hypervigilance	0.049 [-0.021, 0.104]	**0.627 [0.550, 0.682]**	0.039 [-0.016, 0.092]
Unrealistic fears	-0.088 [-0.127, -0.039]	**0.879 [0.829, 0.930]**	-0.041 [-0.089, 0.007]
Episodes of Panic	0.121 [0.070, 0.159]	**0.624 [0.563, 0.673]**	0.007 [-0.042, 0.065]
Lack of Motivation	**0.914 [0.873, 0.954]**	-0.057 [-0.112, -0.015]	-0.097 [-0.149, -0.053]
Anhedonia	**0.811 [0.762, 0.855]**	0.009 [-0.038, 0.055]	0.085 [0.031, 0.134]
Withdrawal from Activities of Interest	**0.825 [0.770, 0.873]**	0.012 [-0.047, 0.059]	-0.054 [-0.109, -0.004]
Decreased Energy	**0.626 [0.552, 0.673]**	0.018 [-0.035, -0.080]	0.093 [0.034, 0.153]
Made Negative Comments	0.015 [-0.030, 0.069]	-0.094 [-0.148, -0.054]	**0.900 [0.858, 0.945]**
Self-Deprecation	-0.133 [-0.183, -0.100]	0.049 [0.011, 0.099]	**0.932 [0.897, 0.979]**
Expressions of Guilt/Shame	-0.069 [-0.121, -0.008]	0.106 [0.052, 0.159]	**0.698 [0.646, 0.760]**
Expressions of Hopelessness	0.232 [0.182, 0.275]	-0.056 [-0.104, -0.003]	**0.656 [0.605, 0.703]**
	Factor 1 Anhedonia	**Factor 2 Anxiety**	**Factor 3 Depression**
Variance Explained by factor	44.50%	13.12%	11.74%
Cronbach's Alpha of Factor	0.772	0.704	0.820
Correlation with F1 (Anhedonia)	1.00	-	-
Correlation with F2 (Anxiety)	0.515 [0.464, 0.568]	1.00	-
Correlation with F3 (Depression)	0.586 [0.549, 0.629]	0.534 [0.485, 0.582]	1.00

Table 8. Exploratory Bifactor Solution of the 12-Item Internalizing Scale.

Items	Anxiety (F1)	Depression (F2)	Anhedonia (F3)	Internalizing General Factor
Repetitive anxious complaints/concerns	**0.556**	0.277	-0.027	**0.332**
Hypervigilance	**0.471**	0.144	-0.100	**0.490**
Unrealistic fears	**0.684**	0.203	-0.096	**0.408**
Episodes of Panic	**0.476**	0.150	-0.041	**0.513**
Lack of Motivation	-0.100	-0.071	**0.426**	**0.717**
Anhedonia	-0.032	0.187	**0.498**	**0.706**
Withdrawal from Activities of Interest	-0.029	0.060	**0.472**	**0.650**
Decreased Energy	-0.061	-0.042	0.122	**0.725**
Made Negative Comments	-0.108	**0.694**	0.037	**0.509**
Self-Deprecation	-0.010	**0.679**	-0.131	**0.548**
Expressions of Guilt/Shame	0.014	**0.389**	-0.274	**0.591**
Expressions of Hopelessness	-0.107	**0.417**	-0.007	**0.668**
Interfactor Correlations	**F1**	**F2**	**F3**	**General Factor**
F1	1.000	-	-	-
F2	-0.029	1.000	-	-
F3	0.190	-0.032	1.000	-
General Factor	0.000	0.000	0.000	1.000

Table 9. Reckase's Multidimensional Item Response Theory Parameterization (1985) with Item Discrimination, MDISC, and Category Threshold Values.

Items	**Item Discrimination Dimension (a_1)**	**Item Discrimination Dimension (a_2)**	**Item Discrimination Dimension (a_3)**	**MDISC**	**Category Threshold b_1**	**Category Threshold b_2**	**Category Threshold b_3**	**Category Threshold b_4**
Repetitive anxious complaints/concerns	-0.111	**0.958**	0.069	0.967	-0.146	0.347	0.817	1.370
Hypervigilance	0.066	**0.850**	0.053	0.854	1.058	1.542	1.954	2.365
Unrealistic fears	-0.153	**1.527**	-0.071	1.536	0.361	1.135	1.687	2.342
Episodes of Panic	0.169	**0.871**	0.010	0.887	0.240	1.195	1.897	2.521
Lack of Motivation	**1.661**	-0.103	-0.177	1.673	0.378	1.086	1.694	2.153
Anhedonia	**1.632**	0.017	0.171	1.641	1.021	1.912	2.715	3.340
Withdrawal from Activities of Interest	**1.378**	0.021	-0.090	1.381	1.005	1.926	2.512	2.974
Decreased Energy	**0.871**	0.025	0.130	0.881	0.210	0.698	1.293	1.859
Made Negative Comments	0.030	-0.185	**1.780**	1.790	-0.460	0.857	1.794	2.677
Self-Deprecation	-0.288	0.105	**2.013**	2.036	-0.445	1.020	2.017	3.021
Expressions of Guilt/Shame	-0.100	0.153	**1.007**	1.023	0.320	1.202	1.893	2.527
Hopelessness	0.374	-0.091	**1.056**	1.124	0.416	1.440	2.166	2.842

Table 10. Descriptive Statistics of the 12-Item Internalizing Scale and 12-Item Externalizing Scale.

	Internalizing Scale	**Externalizing Scale**
Mean	10.45	5.11
Std. Dev	8.74	3.23
Range	0 to 48	0 to 12
Skewness	1.057	.034
Std Error Skewness	.049	.042
Kurtosis	.820	-.909
Std. Error of Kurtosis	.097	.083
Score in 25th Percentile 25	4	3
Score in 50th Percentile	8	5
Score in 75th Percentile	15	8
Cronbach's Alpha	.88	.87

Table 11. Items Evaluated using Expert Content Validation for the Externalizing Scale rated for Children (aged 4 to 11 years)

Item Name	I-CVI for children	Pc	K*	Evaluation
1. Destructive behaviour towards property	0.84615	0.0095	0.84467	Excellent
2. Repetitive Lying – misrepresentations for personal gain	0.69231	0.0873	0.66288	Good
3. Elopement attempts/threats	0.46154	0.20947	0.31886	Poor
4. Demonstrates limited understanding of consequences to behaviour	0.69231	0.08728	0.66288	Good
5. Preoccupation of violence	0.69231	0.08728	0.66288	Good
6. Easily distracted	0.38462	0.15710	0.26992	Poor
7. Impulsivity	0.76923	0.03491	0.76088	Excellent
8. Verbal Abuse	1	0.00012	1	Excellent
9. Disorganization	0.23077	0.034912	0.20294	Poor
10. Bullying peers	1	0.00012	1	Excellent
11. Fire-setting or misuse of ignition	1	0.00012	1	Excellent
12. Argumentativeness – verbally combative, belligerent, quarrelsome	0.84615	0.00952	0.84467	Excellent
13. Hyperactivity – excessive level of activity	0.61538	0.15710	0.54370	Fair
14. Socially inappropriate or disruptive behaviour	0.69231	0.08728	0.66288	Good
15. Intimidation of others or threatened violence	1	0.00012	1	Excellent
16. Outburst of anger	0.92308	0.00159	0.92295	Excellent
17. Violent Ideation	0.76923	0.03491	0.76088	Excellent
18. Violence to others	0.92308	0.00159	0.92295	Excellent
19. Cruelty to animals	1	0.00012	1	Excellent
20. Physical Abuse	1	0.00012	1	Excellent
21. Defiant behaviour	1	0.00012	1	Excellent
22. Stealing	1	0.00012	1	Excellent
23. Expressions supportive of criminal activity	0.69231	0.08728	0.66288	Good

Table 12. Items Evaluated using Expert Content Validation for the Externalizing Scale rated for Adolescents (aged 12 to 18 years)

Item Name	I-CVI for children	Pc	K*	Evaluation
1. Destructive behaviour towards property	1	0.00012	1	Excellent
2. Repetitive Lying	0.92308	0.00159	0.92296	Excellent
3. Elopement attempts/threats	1	0.00012	1	Excellent
4. Demonstrates limited understanding of consequences to behaviour	0.92308	0.00159	0.92296	Excellent
5. Preoccupation of violence	0.84615	0.00952	0.84467	Excellent
6. Easily distracted	0.30769	0.08728	0.24149	Poor
7. Impulsivity	0.69231	0.08728	0.66288	Good
8. Verbal Abuse	0.92308	0.00159	0.92295	Excellent
9. Disorganization	0.23077	0.03491	0.20294	Poor
10. Bullying peers	1	0.00012	1	Excellent
11. Fire-setting or misuse of ignition	1	0.00012	1	Excellent
12. Argumentativeness	1	0.00012	1	Excellent
13. Hyperactivity	0.69231	0.08728	0.66288	Good
14. Socially inappropriate or disruptive behaviour	0.84615	0.00952	0.84467	Excellent
15. Intimidation of others or threatened violence	1	0.00012	1	Excellent
16. Outburst of anger	0.92308	0.00159	0.92295	Excellent
17. Violent Ideation	0.84615	0.00952	0.84467	Excellent
18. Violence to others	1	0.00012	1	Excellent
19. Cruelty to animals	1	0.00012	1	Excellent
20. Physical Abuse	1	0.00012	1	Excellent
21. Defiant behaviour	1	0.00012	1	Excellent
22. Stealing	1	0.00012	1	Excellent
23. Expressions supportive of criminal activity	0.92308	0.00159	0.92295	Excellent

Table 13. Demographics Table for Children/Youth Assessed using the interRAI ChYMH between 2012 to 2016 (*N*=3464).

		Number (% of sample)
Gender	Male	2090 (60.3%)
	Female	1374 (39.7%)
Patient type	Inpatient	331 (9.6%)
	Outpatient	3133 (90.4%)
Assessment Method	Person	2064(59.6%)
	Phone	1286 (37.1%)
	Video	1 (<0.01%)
	Unspecified	113 (3.3%)
Legal Guardianship	Both Parents	1920 (55.4%)
	Only Mother	1021 (29.5%)
	Only Father	139 (4.0%)
	Neither parent but other relative(s) or non-relative(s)	185 (5.3%)
	Child Protection Agency (e.g., CAS)	179 (5.2%)
	Public Guardian	5 (0.1%)
	Youth Cares for Self	15 (0.4%)
DSM-IV Diagnosis[1,2]	Reactive Attachment Disorder	80/2706 (3.0%)
	Attention deficit Hyperactive Disorder	1509/2849 (53.0%)
	Disruptive Behaviour Disorders	796/2787 (28.6%)
	Learning/Communication Disorder	716/2762 (25.9%)
	Autism Spectrum Disorder	361/2683 (13.5%)
	Substance-related disorders	83/2757 (3.0%)
	Schizophrenia and other psychotic disorders	25/2768 (0.9%)
	Mood Disorders	533/2659 (20.0%)
	Anxiety Disorders	1215/2700 (45.0%)
	Eating Disorders	62/2745 (2.3%)
	Sleep Disorders	102/2742 (3.7%)
	Adjustment Disorders	101/2726 (3.7%)

Table 14. Comparison of Model Fit Statistics of the 1-Factor, 2-Factor, and Exploratory Bifactor Model for the Externalizing Subscale.

	1-Factor Model	2-Factor Model	Exploratory Bifactor Model
Root Mean Square Error of Approximation (RMSEA)	0.121	0.052	0.046
Minimum Fit Chi-Square	467.640 ($df = 54$; $p < 0.001$)	221.441($df = 43$; $p < 0.001$)	110.365 ($df = 33$; $p < 0.001$)
Non-Normed Fit Index (NNFI Tucker & Lewis)	0.929	0.987	0.990
Comparative Fit Index (CFI)	0.942	0.992	0.995
Schwarz's Bayesian Information Criterion (BIC)	3003.457	741.027	666.634
Goodness of Index (GFI)	0.953	0.980	0.980
Adjusted Goodness of Index (AGFI)	0.942	0.970	0.959
Goodness of Fit without diagonal values	0.922	0.968	0.967
Adjusted Goodness of Fit without diagonal values	0.905	0.950	0.933

Table 15. Externalizing Scale Rotated Loading Matrix using Robust Diagonally Weighted Least Squares Estimator and Promin Rotation, with loadings greater than |0.300| bolded with Bias-Corrected and Accelerated (BCa).

Item	**Reactive Aggression Factor**	**Proactive Aggression Factor**
Stealing	0.113 [0.066, 0.158]	**0.394 [0.327, 0.454]**
Elopement attempts/threats	0.139 [0.095, 0.182]	**0.395 [0.345, 0.447]**
Bullying Peers	0.106 [0.068, 0.150]	**0.450 [0.387, 0.498]**
Impulsivity	**0.586 [0.554, 0.615]**	0.031 [-0.001, 0.070]
Verbal Abuse	**0.611 [0.577, 0.641]**	0.230 [0.197, 0.264]
Outburst of Anger	**0.826 [0.804. 0.849]**	-0.033 [-0.065, -0.009]
Defiant Behaviour	**0.894 [0.873, 0.914]**	-0.047 [-0.075, -0.025]
Argumentativeness	**0.895 [0.867, 0.912]**	-0.080 [-0.104, -0.053]
Preoccupation with Violence	-0.032 [-0.082, 0.004]	**0.456 [0.403, 0.515]**
Violence to others	-0.053 [-0085, 0.022]	**0.717 [0.676, 0.754]**
Intimidation of others or threatened violence	0.041 [0.007, 0.076]	**0.780 [0.743, 0.815]**
Violent Ideation	-0.126 [-0.154, -0.087]	**0.656 [0.601, 0.698]**
	F1 Reactive Aggression	**F2 Reactive Aggression**
Variance Explained By Factor	41.45%	12.80%
Cronbach's Alpha of Factor	.77	.88
Correlation with Reactive Aggression Factor	1.0	-
Correlation with Proactive Aggression Factor	.605	1.0

Table 16. Rotated Loading Matrix using a Diagonally-Weighted Least Squares Estimator of the Exploratory Bifactor Model of the Externalizing Subscale.

Items	**F1 (Reactive Aggression)**	**F2 (Proactive Aggression)**	**GF**
Stealing	-0.234	0.101	**0.534**
Elopement attempts/threats	-0.109	0.105	**0.528**
Bullying Peers	-0.026	0.067	**0.546**
Impulsivity	**0.237**	**0.470**	**0.411**
Verbal Abuse	**0.383**	**0.463**	**0.601**
Outburst of Anger	**0.482**	**0.655**	**0.472**
Defiant Behaviour	**0.389**	**0.721**	**0.527**
Argumentativeness	**0.384**	**0.717**	**0.493**
Preoccupation with Violence	-0.044	**0.717**	**0.454**
Violence to others	**0.244**	-0.053	**0.633**
Intimidation of others or threatened violence	**0.304**	-0.041	**0.750**
Violent Ideation	0.032	-0.151	**0.568**
	F1	**F2**	**GF**
F1	1.000	**-**	**-**
F2	-0.346	1.000	-
GF	0.000	0.000	1.000
ORION (Factor Determinacy Index)	0.384 (0.619)	0.695 (0.834)	0.826 (0.909)

Table 17. Multidimensional Two-Parameter Normal Ogive Item Response Theory Parameterization Model of the Externalizing Subscale using McDonald-Reckase Parameterization (McDonald, 1976; Reckase, 1985)

Item	**Item Discrimination (a1)**	**Item Discrimination (a2)**	**MDISC**	**Category Threshold (b)**	**MDIFF**
Stealing	0.397	0.437	0.590	-0.824	1.396
Elopement attempts/threats	0.303	0.536	0.616	-0.628	1.020
Bullying Peers	0.483	0.572	0.749	-1.045	1.395
Impulsivity	-0.156	1.069	1.081	0.771	-0.658
Verbal Abuse	0.478	1.266	1.353	0.565	-0.418
Outburst of Anger	-0.159	2.809	2.813	1.692	-0.601
Defiant Behaviour	-0.188	2.842	2.848	1.374	-0.482
Argumentativeness	-0.451	2.628	2.666	1.381	-0.518
Preoccupation with Violence	0.824	0.213	0.851	-1.824	2.144
Violence to others	1.998	-0.223	2.010	-1.435	0.714
Intimidation of others or threatened violence	2.882	-0.032	2.883	-0.913	0.317
Violent Ideation	1.845	-0.388	1.886	-1.859	0.986

Table 18. Pearson's rho Bayesian correlations for ChYMH Internalizing Scale and Criterion Measures of SSIS, CBCL, Beck, BCFPI, and CAFAS.

Criterion Scale	**Hypothesis**	**Pearson correlation ρ**	**Jeffreys's Bayes Factor BF_{+0}(d)**	**Evidence for H_1 (by Jeffreys Criterion)**
SSIS: Internalizing Behaviour	Positively correlated	0.605	30351.906	Decisive
SSIS: Self-Control	Negatively Correlated	-0.420	43.851	Very Strong
SSIS: Cooperation	Negatively Correlated	-0.289	2.988	Anecdotal
BCFPI: Managing Mood	Positively correlated	0.499	392.863	Decisive
BCFPI: Managing Mood and Self-Harm	Positively correlated	0.549	1396.289	Decisive
BCFPI: Internalizing Behaviours	Positively correlated	0.489	279.218	Decisive
BCFPI: Anxiety	Positively correlated	0.228	1.270	Anecdotal
BCFPI: Social Participation	Positively correlated	0.466	199.079	Decisive
CBCL: Internalizing	Positively correlated	0.624	15309.718	Decisive
CBCL: Social Withdrawal	Positively correlated	0.394	16.132	Strong
CBCL: Anxiety/Depression	Positively correlated	0.608	9574.265	Decisive
CBCL: Social Problems	Positively correlated	0.327	4.501	Substantial

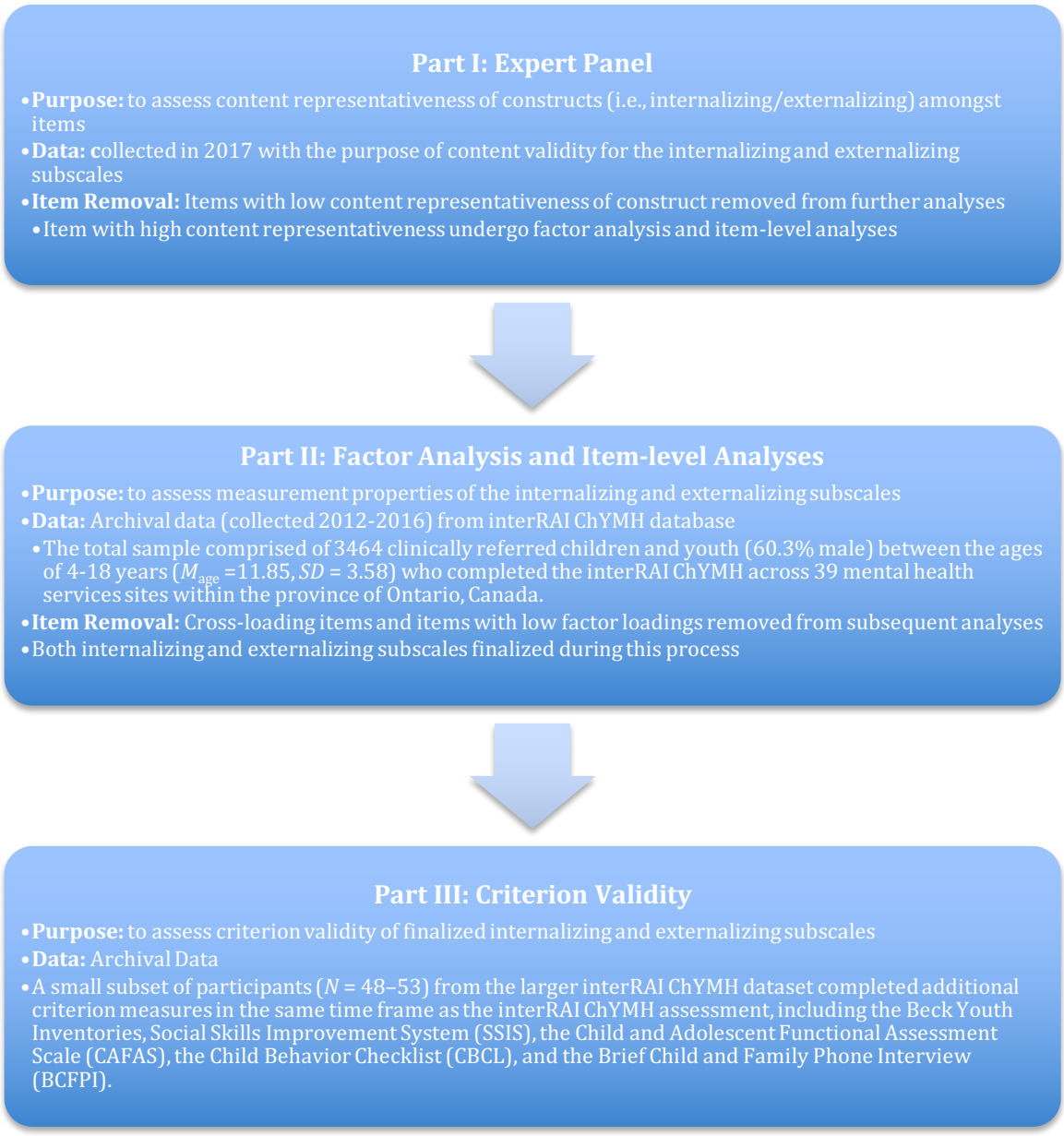
CBCL: Somatic Complaints	Positively correlated	0.415	6.237	Substantial
CAFAS: Mood/Emotions	Positively correlated	0.212	1.184	Anecdotal
Beck: Depression	Positively correlated	0.231	1.112	Anecdotal
Beck: Anxiety	Positively correlated	0.178	0.648	Anecdotal

Table 19. Pearson's rho Bayesian correlations between ChYMH Externalizing Scale and Criterion Measures of SSIS, CBCL, BCFPI, Beck, and CAFAS.

Criterion Scale	**Hypothesis**	**Pearson correlation ρ**	**Jeffreys's Bayes Factor BF_{+0}(d): one-sided extension**	**Evidence for H1 (by Jeffreys Criterion)**
SSIS: Externalizing Behaviours	Positively Correlate	0.648	216652	Decisive
SSIS: Hyperactivity	Positively correlate	0.413	34.22	Very Strong
SSIS: Cooperation	Negatively correlated	-0.549	2737.550	Decisive
SSIS: Responsibility	Negatively correlated	-0.525	947.189	Decisive
SSIS: Self-Control	Negatively correlated	-0.453	109.066	Decisive
SSIS: Bully Behaviour	Positively correlated	0.642	203140	Decisive
SSIS: Empathy	Negatively correlated	-0.345	7.267	Substantial
CAFAS: School Problems	Positively Correlated	0.308	6.201	Substantial
CAFAS: Behaviour toward others	Positively correlated	0.442	213.7	Decisive
CBCL: Aggressive Behaviours	Positively correlated	0.631	33905.047	Decisive
CBCL: Attention problems	Positively correlated	0.330	4.330	Substantial

CBCL: Externalizing	Positively correlated	0.645	65257.389	Decisive
CBCL: Social Problems	Negatively Correlated	-0.540	845.598	Decisive
CBCL: Rule Breaking	Positively correlated	0.409	22.38	Strong
BCFPI: Regulating Attention (RA)	Positively correlated	0.314	4.900	Substantial
BCFPI: Cooperation	Positively correlated	0.610	62570	Decisive
BCFPI: Externalizing Behaviours	Positively Correlated	0.632	199157	Decisive
BCFPI: Conduct	Positively correlated	0.621	107562	Decisive
BCFPI: Bullying	Positively correlated	0.335	7.323	Substantial
Beck: Anger	Positively correlated	0.486	120.5	Decisive
Beck: Disruptive Behaviour	Positively correlated	0.433	35.17	Very Strong

Figure 1. Flow Diagram of Study Design in Constructing the Internalizing and Externalizing Subscales.



APPENDIX A

Internalizing Scale Expert Panel Invitation Email

The expert panel members will be contacted via email and invited to participate in the expert panel. The following email will be sent:

Subject: Expert Panel for Internalizing Scale Development

Dear [EXPERT NAME],

Dr. Shannon Stewart and I would like to formally invite you to participate in the expert panel of the development of a measure on the interRAI ChYMH (Stewart et al., 2015) that can adequately capture internalizing conditions in children and adolescents. The interRAI ChYMH is a semi-structured assessment that consists of items assessing child/youth strengths, level of functioning, and areas of risk to assist care providers in treatment planning and resource allocation. The assessment tool is currently being used in over 60 mental health agencies across Ontario and is compatible with the adult suite (RAI-MH). We are in the process of testing a psychometric instrument on the interRAI Child and Youth Mental Health (ChYMH) to further refine scales that requires your expertise on internalizing mental health indicators. This should only take about 5 minutes. If you are interested in participating in the process, please click on the following link:
https://surveys.mcmaster.ca/limesurvey/index.php/392854?lang=en
Only summary data will be analyzed and no individual data will be used to identify you. We do ask for your name in the survey as we would like to ensure every participant is an expert we invited onto the panel.

Thank you very much. Please contact me at clau263@uwo.ca if you have any questions or concerns.

Kind regards,

Chloe
Survey:

Instructions on the Survey

Internalizing difficulties in children and adolescents refers to "conditions whose central feature is disordered mood or emotion" (Wilkinson, 2009). The terminology commonly used as ‘‘emotional’’ disorders versus ‘‘behavior’’ difficulties are synonymous with ‘‘internalizing’’ versus ‘‘externalizing’’ difficulties. Internalizing conditions are characterized by symptoms of depressed mood, anxiety, and anhedonia.

To what extent do you perceive each individual item to be representative as **an internalizing difficulty mental state indicator for a child (ages 4-11)**:

In a 4-point rating scale (1= not representative, 2=minimally representative, 3=moderately representative, and 4= strongly representative) (definition of items will be provided):

1. Episodes of panic
2. Expressions of hopelessness
3. Nightmares
4. Lack of interest in social interaction
5. Hyper-vigilance
6. Crying, tearfulness
7. Sad, pained, or worried facial expressions
8. Intrusive thoughts or flashbacks
9. Self-deprecation
10. Irritability
11. Anhedonia
12. Expressions of guilt or shame
13. Unrealistic fears
14. Made negative statements
15. Repetitive anxious complaints/concerns
16. Obsessive thoughts
17. Withdrawal from activities of interest
18. Decreased Energy
19. Compulsive behaviour
20. Re-enactment through play of traumatic events
21. Lack of motivation
22. Repetitive Health complaints

Are there any items that you would like to see that was missed?
[TEXT BOX]

To what extent do you perceive each individual item to be representative as **an internalizing difficulty mental state indicator for an adolescent (ages 12-18)**:

In a 4-point rating scale (1= not representative, 2=minimally representative, 3=moderately representative, and 4= strongly representative) (definition of items will be provided):

Adolescents aged (12-18)

1. Episodes of panic
2. Expressions of hopelessness

3. Nightmares
4. Lack of interest in social interaction
5. Hyper-vigilance
6. Crying, tearfulness
7. Sad, pained, or worried facial expressions
8. Intrusive thoughts or flashbacks
9. Self-deprecation
10. Irritability
11. Anhedonia
12. Expressions of guilt or shame
13. Unrealistic fears
14. Made negative statements
15. Repetitive anxious complaints/concerns
16. Obsessive thoughts
17. Withdrawal from activities of interest
18. Decreased Energy
19. Compulsive behaviour
20. Re-enactment through play of traumatic events
21. Lack of motivation
22. Repetitive Health Complaints

Are there any items that you would like to see that were missed? Please feel free to leave any general comments as well.

[TEXT BOX]

Appendix B: Description and Scoring of Items

Internalizing Scale Items:

0-4 items (14 items)

0-not present
1=present but not exhibited in last 3 days
2=exhibited on 1-2 of last 3 days
3=exhibited daily in last 3 days, 1-2 episodes
4= exhibited daily in last 3 days, 3 or more episodes or continuously

1. Episodes of panic
2. Expressions of hopelessness
3. Nightmares
4. Lack of interest in social interaction
5. Hyper-vigilance
6. Crying, tearfulness
7. Sad, pained, or worried facial expressions
8. Intrusive thoughts or flashbacks
9. Self-deprecation
10. Irritability
11. Anhedonia
12. Expressions of guilt or shame
13. Unrealistic fears
14. Made negative statements
15. Repetitive anxious complaints/concerns
16. Obsessive thoughts
17. Withdrawal from activities of interest
18. Decreased Energy
19. Compulsive behaviour
20. Re-enactment through play of traumatic events
21. Lack of motivation

Externalizing Scale Items:

- ➢ = Items in the Disruptive/Aggressive Behaviour Scale
- ▪ = Items in the Hyperactivity/Distractibility items
- ⇒ Other items related to Conduct Problems

0-4 items (14 items)

0-not present

1=present but not exhibited in last 3 days
2=exhibited on 1-2 of last 3 days
3=exhibited daily in last 3 days, 1-2 episodes
4= exhibited daily in last 3 days, 3 or more episodes or continuously

- iE3l; Verbal Abuse (e.g., others were threatened, cursed at)
- iE3m; Physical Abuse (e.g., hits schoolmate, punches sibling in the face)
- iE3n; Socially inappropriate or disruptive behaviour (e.g., screamed out during class, smeared or threw food or feces)
- iE3q; Destructive behaviour towards property (e.g., throwing or breaking objects, turning over beds or tables, vandalism)
- iE3r; Outburst of anger – intense flare-up of anger in reaction to a specific action or event (e.g., tantrums when told "no")
- iE1oo; Impulsivity (e.g., running into traffic, interrupts, taking risky actions without thinking; difficulty taking turns)
- iE1pp; Easily distracted (e.g., epsiodes of difficulty paying attention, gets sidetracked)
- iE1qq; Hyperactivity – excessive level of activity
- iE1rr; Disorganization –e.g., problems organizing personal belongings; difficulty adhering to schedule

⇒ iE3s; Defiant behaviour– active persistent refusal to comply with reasonable requests by others
⇒ iE3t; Argumentativeness – verbally combative, belligerent, quarrelsome
⇒ iE3v; Repetitive Lying – misrepresentations for personal gain
⇒ CY_C1vv; Demonstrates limited understanding of consequences to behaviour – e.g., consistently fails to realize that his or her actions will have a negative effect on others
⇒ iE1dddd; Expressions supportive of criminal activity (e.g., "it's only a crime if you get caught")

<u>0-5 Items (9 items)</u>

0=never
1=most recent instance was more than 1 year ago
2=31 days- 1 year ago
3= 8-30 days ago
4=4-7 days ago

5= in the last 3 days

- ⇒ iX18; Fire-setting or misuse of ignition
- ⇒ iX16b; Cruelty to animals – deliberate mistreatment of or physical injury to animals (exclude behaviours that are consistent with cultural norms)
- ⇒ iX16c; Preoccupation of violence – e.g., depictions of violence
- ⇒ iX2c; Violent Ideation –e.g., reports of premeditated thoughts, statements
- ⇒ iX2b; Intimidation of others or threatened violence – intentionally makes threatening gestures, verbalizations or stance with no physical contact (e.g., explicit threats of violence)
- ⇒ iX2a; Violence to others – acts with purposeful, malicious, or vicious intent, resulting in physical harm to another (e.g., stabbing, choking, beating)
- ⇒ iE15a; Stealing –e.g., theft from family, shoplifting
- ⇒ iE15e; Bullying peers – pattern of repeated oppression or victimization of others
- ⇒ iE15d; Elopement attempts/threats

APPENDIX C

Externalizing Scale Expert Panel Review

The expert panel members will be contacted via email and invited to participate in the expert panel. The following email will be sent:

SUBJECT: interRAI ChYMH Expert Panel for Externalizing Scale Development

Dear [EXPERT NAME],

Apologies for emailing you again about this, but we would really appreciate it if you could take a couple minutes to serve as an expert on this panel for a new scale development. Dr. Shannon Stewart and I would like to formally invite you to participate in the expert panel of the revision of a measure on the interRAI ChYMH (Stewart et al., 2015) that can adequately capture externalizing conditions in children and adolescents. The interRAI ChYMH is a semi-structured assessment that consists of items assessing child/youth strengths, level of functioning, and areas of risk to assist care providers in treatment planning and resource allocation. The assessment tool is currently being used in over 60 mental health agencies across Ontario and is compatible with the adult suite (RAI-MH). We would like to refine scales that require your expertise on externalizing mental health indicators. This should only take about 5 minutes. If you are interested in participating in the process, please click on the following link:

https://surveys.mcmaster.ca/limesurvey/index.php/195181?lang=en

Only summary data will be analyzed and no individual data will be used to identify you. We do ask for your name in the survey as we would like to ensure every participant is an expert we invited onto the panel.

Thank you very much. Please contact me at clau263@uwo.ca if you have any questions or concerns. Happy holidays!

Kind regards,

Chloe

Step 3: Instructions for Panel

"Externalizing" and "internalizing" difficulties in children and adolescents are synonymous with "behavioural" and "emotional" problems respectively. (Achenbach, 1978). Externalizing behaviour refers to a child/youth's problematic "outward behavior" as the child acts excessively negative towards the *external* environment (Campbell, Shaw, & Gilliom, 2000; Eisenberg et al., 2001). Other terms used to describe the externalizing behaviour construct includes aggression, antisocial, delinquency, hyperactivity, and "undercontrolled behaviour" (Hinshaw, 1987; Liu, 2004).

To what extent do you perceive each individual item to be representative as **an externalizing difficulty mental state indicator for a child (ages 4-11)**:

In a 4-point rating scale (1= not representative, 2=minimally representative, 3=moderately representative, and 4= strongly representative) (definition of items will be provided):

1. Destructive behaviour towards property (e.g., throwing or breaking objects, turning over beds or tables, vandalism)
2. Repetitive Lying – misrepresentations for personal gain
3. Elopement attempts/threats
4. Demonstrates limited understanding of consequences to behaviour – e.g., consistently fails to realize that his or her actions will have a negative effect on others
5. Preoccupation of violence – e.g., depictions of violence
6. Easily distracted (e.g., epsiodes of difficulty paying attention, gets sidetracked)
7. Impulsivity (e.g., running into traffic, interrupts, taking risky actions without thinking; difficulty taking turns)
8. Verbal Abuse (e.g., others were threatened, cursed at)
9. Disorganization –e.g., problems organizing personal belongings; difficulty adhering to schedule
10. Bullying peers – pattern of repeated oppression or victimization of others
11. Fire-setting or misuse of ignition
12. Argumentativeness – verbally combative, belligerent, quarrelsome
13. Hyperactivity – excessive level of activity
14. Socially inappropriate or disruptive behaviour (e.g., screamed out during class, smeared or threw food or feces)

15. Intimidation of others or threatened violence – intentionally makes threatening gestures, verbalizations or stance with no physical contact (e.g., explicit threats of violence)
16. Outburst of anger – intense flare-up of anger in reaction to a specific action or event (e.g., tantrums when told "no")
17. Violent Ideation –e.g., reports of premeditated thoughts, statements
18. Violence to others – acts with purposeful, malicious, or vicious intent, resulting in physical harm to another (e.g., stabbing, choking, beating)
19. Cruelty to animals – deliberate mistreatment of or physical injury to animals (exclude behaviours that are consistent with cultural norms)
20. Physical Abuse (e.g., hits schoolmate, punches sibling in the face)
21. Defiant behaviour– active persistent refusal to comply with reasonable requests by others
22. Stealing –e.g., theft from family, shoplifting
23. Expressions supportive of criminal activity (e.g., "it's only a crime if you get caught")

Are there any items that you would like to see that was missed?
[TEXT BOX]

To what extent do you perceive each individual item to be representative as **an externalizing difficulty mental state indicator for an adolescent (ages 12-18)**:

In a 4-point rating scale (1= not representative, 2=minimally representative, 3=moderately representative, and 4= strongly representative) (definition of items will be provided):

Adolescents aged (12-18)

1. Destructive behaviour towards property (e.g., throwing or breaking objects, turning over beds or tables, vandalism)
2. Repetitive Lying – misrepresentations for personal gain
3. Elopement attempts/threats
4. Demonstrates limited understanding of consequences to behaviour – e.g., consistently fails to realize that his or her actions will have a negative effect on others
5. Preoccupation of violence – e.g., depictions of violence
6. Easily distracted (e.g., epsiodes of difficulty paying attention, gets sidetracked)

7. Impulsivity (e.g., running into traffic, interrupts, taking risky actions without thinking; difficulty taking turns)
8. Verbal Abuse (e.g., others were threatened, cursed at)
9. Disorganization –e.g., problems organizing personal belongings; difficulty adhering to schedule
10. Bullying peers – pattern of repeated oppression or victimization of others
11. Fire-setting or misuse of ignition
12. Argumentativeness – verbally combative, belligerent, quarrelsome
13. Hyperactivity – excessive level of activity
14. Socially inappropriate or disruptive behaviour (e.g., screamed out during class, smeared or threw food or feces)
15. Intimidation of others or threatened violence – intentionally makes threatening gestures, verbalizations or stance with no physical contact (e.g., explicit threats of violence)
16. Outburst of anger – intense flare-up of anger in reaction to a specific action or event (e.g., tantrums when told "no")
17. Violent Ideation –e.g., reports of premeditated thoughts, statements
18. Violence to others – acts with purposeful, malicious, or vicious intent, resulting in physical harm to another (e.g., stabbing, choking, beating)
19. Cruelty to animals – deliberate mistreatment of or physical injury to animals (exclude behaviours that are consistent with cultural norms)
20. Physical Abuse (e.g., hits schoolmate, punches sibling in the face)
21. Defiant behaviour– active persistent refusal to comply with reasonable requests by others
22. Stealing –e.g., theft from family, shoplifting
23. Expressions supportive of criminal activity (e.g., "it's only a crime if you get caught")

Are there any items that you would like to see that were missed?

[TEXT BOX]

Do you have any comments?

[TEXT BOX]

Appendix D

Dear Dr. Hinson,

This is a confirmation email that we spoke again today in your office that my supervisors and I plan to have an expert panel portion for the externalizing scale (format is identical to the internalizing scale expert panel) and this is also considered quality assurance as part of questionnaire construction. Only summary data will be analyzed and no individual data will be used to identify any individual. I'll cc'd supervisors of this project Dr. Donald Saklofske and Dr. Shannon Stewart.

Thank you very much for your time today.

Kind regards,

Chloe

Chloe Lau, B.Sc.
M.Sc. Candidate
Clinical Psychology
Department of Psychology
Western University

From: Riley Hinson
Sent: Thursday, November 10, 2016 8:11 AM
To: Chloe Lau; Katelyn Harris
Subject: Re: Ethics Approval for Internalizing Scale Development Study

Thanks for your email. You have accurately captured our conversation, and I confirm that the first part may be considered secondary use of deidentified data and the second part would not require ethics approval as it is quality assurance for internal use of construction of the questionnaire.

Katelyn, I talked to one of Don Saklofse's students about the issues below. I will keep this email as confirmation of what we discussed, but based on her description I am comfortable identifying the first part as secondary use of deidentified data, and the second as not "research" but rather quality assurance as part of questionnaire construction.

On 11/8/2016 2:00 PM, Chloe Lau wrote:
Dear Dr. Hinson,

We had a discussion in person today about whether I need ethics approval for a project. To confirm this, we discussed that I would be using interRAI data collected across over 25 mental agencies and stored in the faculty of education to develop a new scale using that data. We will be analyzing deidentified data collected across these agencies to develop an internalizing scale (composes of symptoms of depression, anhedonia, and anxiety). We previously had ethics approval for storing and data analysis by the University ethics board (REB #106415).

For the expert panel portion, we plan to ask 5-15 senior psychology students, psychology residents, and practicing psychologists about their thoughts on the items of the scale. We would ask, "what items do you think would be important to identify a child who might have internalizing problems that you would want to look further into? Highlight 5-8 items out of a list of existing items on a list." We would ensure we are incorporating the important items onto the scale. You had mentioned that this would be considered quality assurance and we would not need further ethics approval for this project.

Thank you very much for your help today.

Kind regards,

Chloe

Chloe Lau, B.Sc.
M.Sc. Candidate
Clinical Psychology
Department of Psychology
Western University

Appendix E

Internalizing Scale Scoring Sheet

The internalizing scale consists of 12 items, with 4 items for each of the three factors (i.e., anhedonia, depression, anxiety). Scores range between 0 to 48, with higher scores indicating higher levels of internalizing symptoms (i.e., emotional distress/disturbance).

Score **Interpretation of the Score**

0 - Not present
1 - Present but not exhibited in last 3 days
2 - Exhibited on 1–2 of last 3 days
3 - Exhibited daily in last 3 days, 1–2 episodes
4 - Exhibited daily in last 3 days, 3 or more episodes or continuously

Item	iCode	Score (0-4)
Repetitive anxious complaints/concerns	iE1eee	
Hypervigilance	iE1eeee	
Unrealistic fears	iE1fff	
Episodes of Panic	iE1kkk	
	Total Score for Anxiety Factor:	
Lack of Motivation	iE1rrr	
Anhedonia	iE1sss	
Withdrawal from Activities of Interest	iE1ttt	
Decreased Energy	iE1uu	
	Total Score for Anhedonia factor:	
Made Negative Comments	iE1vv	
Self-Deprecation	iE1ww	
Expressions of Guilt/Shame	iE1xx	
Expressions of Hopelessness	iE1yy	
	Total score for depression factor:	
	Total Score (add up scores for ALL 3 factors):	

Externalizing Scale Scoring Criteria

The externalizing scale consists of 12 items. Scores range between 0 to 12, with higher scores indicating higher levels of externalizing symptoms (i.e., behavioural disturbance). 5 items are measured using the 0 to 4 scale, while 7 items are measured using a 0 to 5 ordinal scale. Before obtaining a total score for the externalizing scale, please recode the item scores accordingly:

0 to 1 Scale Recoded	**Original: 0 to 4 Scale**	**Original: 0- 5 Scale**
0 –Not Present	0 – Not present	0 – Never
1 – Present currently or in the past	1 – Present but not exhibited in last 3 days	1 – More than 1 year ago
1 – Present currently or in the past	2 – Exhibited on 1– 2 of last 3 days	2 – 31 days to 1 year ago
1 – Present currently or in the past	3 – Exhibited daily in last 3 days, 1-2 episodes	3 – 8 to 30 days ago
1 – Present currently or in the past	4 –Exhibited daily in last 3 days, 3 or more episodes or continuously	4 –4 to 7 days ago
1 – Present currently or in the past	–	5 –In the last 3 days

Externalizing Scale Scoring Sheet

Item Name	iCode	Original Score	Original Scale Score	Recoded (see legend above 0=not present; 1=present)
1. Stealing –e.g., theft from family, shoplifting	iE15a	0-5		
2. Elopement attempts/threats	iE15d	0-5		
3. Bullying peers	iE15e	0-5		
4. Impulsivity	iE1oo	0-4		
5. Verbal Abuse	iE3l	0-4		
6. Outburst of anger – intense flare-up of anger in reaction to a specific action or event	iE3r	0-4		
7. Defiant behaviour– active persistent refusal to comply with reasonable requests by others	iE3s	0-4		
8. Argumentativeness – verbally combative, belligerent, quarrelsome	iE3t	0-4		
9. Preoccupation of violence – e.g., depictions of violence	iX16c	0-5		
10. Violence to others	iX2a	0-5		
11. Intimidation of others or threatened violence	iX2b	0-5		
12. Violent Ideation	iX2c	0-5		
			TOTAL SCORE (add 0's and 1's recoded only):	

CURRICULUM VITAE

Name:	**Chloe Lau**
Post-secondary Education and Degrees:	University of Zurich, Department of Psychology Zurich, Switzerland 2017 Invited Doctoral Student
	Western University London, Ontario, Canada 2015-2017 Masters of Science in Clinical Psychology
	McMaster University Hamilton, Ontario, Canada 2011-2015 Bachelors of Science in Psychology
Honours and Awards	Joseph-Armand Bombardier Canada Graduate Scholarship Doctoral Award (2017-2020) Social Sciences and Humanities Research Council
	Michael Smith Foreign Supplement Study Supplements (2017) Social Sciences and Humanities Research Council
	Doctoral Ontario Graduate Scholarship (2017-2018) – *Declined* Province and Ontario and Ontario Universities
	ThinkSwiss Scholarship (2017) Embassy of Switzerland in the United States of America
	PSAC 610 Academic Achievement Scholarships (2017) PSAC 610 at Western University
	Best Research Poster in the Graduate Clinical/Education Section Department of Psychiatry & Behavioural Neurosciences, McMaster University
	Runner-Up Poster in the Clinical Psychology Section Canadian Psychological Association 2017
	Graduate Research Awards Fund (2017-2018) Faculty of Social Sciences, Western University
	Joseph-Armand Bombardier Canada Graduate Scholarship Master's Award (2016-2017) Social Sciences and Humanities Research Council
	Graduate Research Awards Fund (2016-2017) Faculty of Social Sciences, Western University
	University Senate Scholarships (2013; 2014) McMaster University
Selected Publications	**Lau, C**., Ford, J., Van Lieshout, R.J., Saperson, K., McConnell, M.M., & McCabe, R. *Enhancing Mentorship in Psychiatry: A Study*

Investigating Needs and Preferences in the Development of a Mentoring Program. Manuscript under review.

Lau, C., Stewart, S.L., Saklofske, D. H., Tremblay, P.F., & Hirdes, J. (2017) Validation of the interRAI-ChYMH Aggressive Disruptive Behaviour Scale and Hyperactivity-Distractibility Scale. Child Psychiatry and Human Development, 1–11.

Lau, C., & Forchuk, C. (in press). Hopelessness. In B. J. Carducci (Editor-in-Chief) & A. Di Fabio, D. H. Saklofske, & C. Stough (Vol. Eds.), *The Wiley-Blackwell encyclopedia of personality and individual differences: Vol. III. Personality processes and individual differences.* Hoboken, NJ: John Wiley & Sons.

Sarmiento, C., & **Lau, C**. (in press). DSM-5. In B. J. Carducci (Editor-in-Chief) & A. Di Fabio, D. H. Saklofske, & C. Stough (Vol. Eds.), *The Wiley-Blackwell encyclopedia of personality and individual differences: Vol. IV. Personality processes and individual differences.* Hoboken, NJ: John Wiley & Sons.

Tohver, G., & **Lau, C**. (in press). Genius, personality correlates of. In B. J. Carducci (Editor-in-Chief) & A. Di Fabio, D. H. Saklofske, & C. Stough (Vol. Eds.), *The Wiley-Blackwell encyclopedia of personality and individual differences: Vol. III. Personality processes and individual differences.* Hoboken, NJ: John Wiley & Sons.

Babcock, S. E., Wilson, C. A., & **Lau, C.** (2017). Test Review: School Motivation and Learning Strategies Inventory (SMALSI): College Form [Manual] by Stroud, KC, & Reynolds, CR. *Canadian Journal of School Psychology*, DOI 10.1177/0829573517697532

Lau, C., Ford, J., Van Lieshout, R. J., Saperson, K., McConnell, M., & McCabe, R. (2016). Developing mentoring competency: Does a one session training workshop have impact?. *Academic Psychiatry, 40*(3), 429–433.

Van Ameringen, M., Turna, J., Patterson, B., & **Lau, C.** (2015). Pharmacotherapy for adolescent social phobia. In K. Ranta, M. Marttunen, L-J. García-Lopez, and A. M. La Greca (Eds.), *Social anxiety and phobia in adolescents - Development, manifestation and intervention strategies* (pp.301–322). Powell, WY: Springer International Publishing.