Toward Validating a Dimensional Parent-Report Measure of Irritability in a National Sample: Initial Scale Development

A dissertation presented to

the faculty of

the College of Arts and Sciences of Ohio University

In partial fulfillment

of the requirements for the degree

Doctor of Philosophy

Sean T. Tams

December 2020

This dissertation titled

Toward Validating a Dimensional Parent-Report Measure of Irritability in a National Sample: Initial Scale Development

by

SEAN T. TAMS

has been approved for

the Department of Psychology

and the College of Arts and Sciences by

Brian T. Wymbs

Associate Professor of Psychology

Florenz Plassmann

Dean, College of Arts and Sciences

Abstract

TAMS, SEAN T., Ph.D., December 2020, Clinical Psychology

Toward Validating a Dimensional Parent-Report Measure of Irritability in a National Sample: Initial Scale Development

Director of Dissertation: Brian T. Wymbs

Research on the measurement of youth irritability lacks consensus, and there is need to address the underdeveloped literature on clinical assessment of the tonic (i.e., persistent irritable/angry mood) and phasic (i.e., temper outbursts) components of irritability. Currently, no measure has been developed or validated that assesses tonic/phasic irritability, which presents concerns with existing evidence related to the assessment and treatment of youth irritability. The current study developed and provided initial psychometric support for a parent-report, dimensional rating scale of irritability using empirically-supported steps to scale development. Mothers of children ages 6-12 (N = 397) were recruited through Amazon's Mechanical Turk (MTurk) and completed items measuring tonic/phasic irritability and related problems. Exploratory factor analysis (EFA) derived a 22-item, two-factor scale that demonstrated excellent reliability and showed convergent validity with other measures of child psychological problems. The ability for tonic and phasic irritability to differentially predict outcomes was not supported, and evidence for using the scale as a screening tool was mixed. These results provided, to my knowledge, support for the first tonic/phasic irritability scale and have important implications for research and practice related to youth irritability.

Dedication

This project is dedicated to the patients and families with whom I have worked, as well as the individuals (family, friends, colleagues, and mentors) who helped me complete this journey.

Acknowledgments

I would like to express the utmost appreciation to my dissertation director, Brian Wymbs, for his unwavering support and guidance toward the completion of this project. I am also grateful to my committee members (Nicholas Allan, Steven Evans, Julie Owens, and Thomas Vander Ven) for their steadfast commitment to improving the rigor and impact of this project.

Table of Contents

Page

List of Tables

Page

List of Figures

Page

Introduction

Contemporary perspectives on youth and adult psychopathology attempt to organize underlying neural mechanisms (e.g., brain activation/connectivity) and phenotypic expressions (e.g., observable behaviors) into categorical or dimensional constructs. The *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; *DSM-5*; American Psychiatric Association [APA], 2013) is the prevailing system—at least in the United States behavioral healthcare system—used to classify mental disorders as *categorical* constructs (e.g., attention-deficit/hyperactivity disorder [ADHD], major depressive disorder [MDD]). Alternatively, *dimensional* approaches to classification like the Research Domain Criteria (RDoC; Insel et al., 2010) examine transdiagnostic constructs (i.e., occurring across multiple mental disorders) that can be measured by multiple units of analysis (e.g., genetic, physiological, behavioral). Both categorical and dimensional approaches have been applied to the construct of youth irritability. Until recently, the scientific literature on youth irritability was plagued by a lack of consensus (Roy & Comer, 2020), with major advances related to phenomenology, assessment, and treatment occurring within the past 20 years. Nonetheless, as discussed in a recent special issue of the journal *Behavior Therapy* (March, 2020) on the state-of-the-science of youth irritability, clinical assessment of irritability remains a domain that requires further elucidation (Beauchaine & Tackett, 2020).

The most recent attempt to categorize youth irritability is represented in the *DSM-5* diagnosis of disruptive mood dysregulation disorder (DMDD), comprising persistent irritability/anger and severe temper outbursts (APA, 2013). Prior to (e.g., Axelson et al.,

2011; Parens, Johnston, & Carlson, 2010) and in the years since its publication (e.g., Evans et al., 2017; Goldstein et al., 2017), the scientific community has been wary of DMDD as an approach to capturing youth with high irritability. Key among reasons for this wariness is insufficient evidence supporting its diagnostic validity, including the concern that youth with DMDD would already meet criteria for one or more categorical diagnoses (i.e., high degrees of comorbidity with existing diagnoses that feature irritability; for review, see Evans et al., 2017). Rather, researchers have increasingly advocated for measurement of irritability as a dimensional construct.

Measuring irritability as a dimensional construct follows from work by several researchers (Brotman, Kircanski, Stringaris, Pine, & Leibenluft, 2017; Meyers, DeSerisy, & Roy, 2017) that identifies underlying neural mechanisms (i.e., frustrative nonreward, and response to threat, defined below) of youth irritability. Some research suggests that irritability can be measured as a unidimensional construct (e.g., Stringaris et al., 2012), although recent assertions indicate measures are needed to capture two dimensions of irritability (i.e., *tonic* irritability, which includes persistent irritable/angry mood; *phasic* irritability, which includes temper outbursts; Carlson & Klein, 2018; Stringaris, Vidal-Ribas, Brotman, & Leibenluft, 2018). However, the most up-to-date review of irritability assessment practices (Beauchaine & Tackett, 2020) concludes that "there are no psychometrically validated scales that differentiate between tonic and phasic irritability, even though a number of available measures assess irritability more broadly" (p. 353). Furthermore, psychometrically-sound adult irritability measures may have the ability to contribute additional coverage to the construct of youth irritability, particularly as

researchers call for measures that capture greater variability in the experience of irritability as a state (phasic) and trait (tonic; Toohey & DiGiuseppe, 2017).

It is clear that the conceptualization and measurement of irritability in childhood needs further refinement to align with contemporary theory in order to permit advances in its empirical study. In particular, empirical research is needed to explore the dimensional factor structure of youth irritability by (1) developing a measure that incorporates items assessing tonic and phasic components, and incorporates empirical knowledge from measures of adult and youth irritability, (2) testing whether tonic and phasic components are separate, but related factors (i.e., two-factor solution) or whether they are all interrelated items on a single factor (i.e., one-factor solution), and (3) establishing psychometric support (reliability, validity) for the measure. The current study developed and tested the psychometric properties of a dimensional measure of child irritability, including items assessing both tonic (i.e., irritable/angry mood) and phasic (i.e., temper outbursts) components, using theoretically-driven guidelines for scale development and item sampling from child and adult sources. Theoretical and empirical findings regarding youth irritability, including its clinical definition, course, and assessment, are reviewed below to illustrate their influence on the scale development procedures undertaken in this study.

Theoretical Definition of Irritability

Theoretical perspectives on irritability implicate emotion dysregulation, which involves having deficits in controlling emotional states to a degree that significantly impairs functioning (Bunford, Evans, & Wymbs, 2015; Cole, Michel, & Teti, 1994), as a

higher-order process that increases risk for persistent angry mood and temper outbursts (Carlson & Klein, 2018). Anger, or a negative affective state occurring as a result of being blocked from attaining a goal and/or appraising a threat or act of wrongdoing against oneself (Carver & Harmon-Jones, 2009; Fernandez & Johnson, 2016), is the primary affective component of irritability (Toohey & DiGiuseppe, 2017). The persistence of angry mood reflects having a diminished ability to regulate emotional arousal and express it in an adaptive way (i.e., increased proneness to anger; Stringaris et al., 2018). Temper outbursts are behavioral manifestations of irritability that reflect both anger and distress (Giesbrecht, Miller, & Müller, 2010; Potegal & Davidson, 2003), and the specific behaviors exhibited during temper outbursts are related to the level of anger experienced (e.g., foot stamping associated with low anger, verbal and/or physical aggression associated with high anger).

Translational models of irritability characterize its dimensions as related to the neural mechanisms of frustrative nonreward (i.e., emotional or behavioral responses that occur when an anticipated reward is not received) and responding to threat (i.e., maladaptive response to a threatening stimulus; Brotman et al., 2017). Terminology proposed by Stringaris and colleagues (2018) and reiterated by Beauchaine and Tackett (2020) serves to unify the affective and behavioral components of irritability by conceptualizing tonic and phasic dimensions. Together, these dimensions form the phenotypic measurement of frustrative nonreward and response to threat that were implicated in translational models of irritability (Brotman et al., 2017; Meyers et al., 2017). Despite growing evidence that supports the tonic/phasic conceptualization of

youth irritability, this contemporary terminology has yet to receive rigorous examination in literature pertaining to its clinical assessment (Beauchaine & Tackett, 2020). This underdeveloped area of the literature calls into question both theoretical definitions of irritability (i.e., absence of confirmatory evidence supporting multi-dimensional construct) as well as previous studies that have empirically assessed clinical characteristics and trajectories of irritability (i.e., findings may not align with construct definition).

Developmental Trajectories of Irritability

Trajectories of irritability have been studied across the lifespan, and like emotion dysregulation (Bunford et al., 2015; Cole et al., 1994), irritability peaks during preschool (ages 3-5), declines during middle childhood (ages 6-8), and decreases precipitously over the rest of childhood and adolescence (Brotman et al., 2017; Leibenluft & Stoddard, 2013). Several studies (including youth ages 3 to 15) have explored developmental trajectories of irritability with largely converging results, showing that most children (60-80%) fall within a low-irritability group that does not show changes over time (e.g., Orri et al., 2019; Riglin et al., 2019; Wiggins, Mitchell, Stringaris, & Leibenluft, 2014). Most concerning are classes of children that show increasing levels of irritability across childhood/early adolescence (2-13%) or those who show persistently high levels of irritability (2-11%), which are related with increased risk for mental health problems and functional impairment (Orri et al., 2019; Riglin et al., 2019).

Cross-sectional prevalence of irritability at specific points in development (rather than percent of those who fall within specific trajectories) have also been assessed.

Copeland, Brotman, and Costello (2015) conducted the first study to examine the prevalence (over a period of three months) of dimensional (tonic and phasic) irritability in a large, community-based sample (ages 9 to 16). Almost a quarter (23%) of their sample endorsed elevated tonic *and* phasic irritability at some point during the three-month period, with elevated phasic (51%) being more common than elevated tonic (28%). In a more recent study, Moore and colleagues (2019) examined tonic/phasic irritability prevalence in a large twin study that assessed children (ages 8 to 17) across two waves (18 months apart). Similar rates of children evinced (at least once across waves) elevated phasic (52%) and tonic (21%) irritability as observed in Copeland and colleagues' (2015) study. Rates of elevated tonic and phasic irritability were not qualified by gender or sex in either of these studies. Moving beyond community-based samples, a number of clinical samples (ages 5 to 19) assessing the categorical diagnosis of DMDD found prevalence rates of elevated tonic and phasic irritability to aggregate around 30% (Axelson et al., 2012; Freeman, Youngstrom, E. A., Youngstrom, J. K., & Findling, 2016; Margulies, Weintraub, Basile, Grover, & Carlson, 2012; Mitchell et al., 2016; Tiwari, Agarwal, Arya, Gupta, & Mahour, 2016). Altogether, community-based and clinical samples appear to indicate that somewhere between 20% to 30% of youth experience elevated tonic and phasic irritability at some point during childhood or adolescence.

However, caution is warranted when interpreting these findings. The community- and clinic-based studies above relied on items from structured interviews (e.g., Child and Adolescent Psychiatric Assessment; Angold & Costello, 2000) that were developed and

validated to assess diagnoses that appear in the *DSM-IV* (4th ed.; APA, 1994) and International Statistical Classification of Diseases and Related Health Problems (10th rev.; *ICD-10*; World Health Organization, 1992) as indicators of tonic and phasic irritability. Given that these tonic and phasic indices were not empirically-derived, but instead chosen from items that were individual symptoms of categorical diagnoses, the trajectories and cross sectional prevalence data reported above may not represent true tonic and phasic irritability (Beauchaine & Tackett, 2020). Specific instruments developed and validated to measure tonic and phasic irritability are needed to address this limitation.

Clinical Assessment of Irritability

Irritability is a heterogeneous construct with no universal, "gold-standard" method of assessment for individuals across development, including children and adolescents (Avenevoli, Blader, & Leibenluft, 2015) or adults (Toohey & DiGiuseppe, 2017).[1] Research examining adult irritability is more well-established both in terms of construct definition and clinical assessment, so findings derived from adult instruments are relevant given more accumulation of evidence (although, as mentioned above, this has not translated into greater consensus). In the adult literature, Toohey and DiGiuseppe (2017) advocate for measuring irritability using self-report rating scales that include items that

[1] Stringaris and colleagues (2018) discussed some of the most commonly used methods for assessing irritability, which include self- and parent-report questionnaires. Since that time, additional, novel measures (including semi-structured interviews) of youth irritability were developed and included in a recent special issue (Roy & Comer, 2020). Questionnaires assessing child and adult irritability were located in a literature review conducted in 2018-2019; these questionnaires are reviewed below and appear in Appendix A. The recent instruments (published 2020) are briefly reviewed in Appendix B but were not available to incorporate into the literature review and methods considered during the proposal of the current study.

capture the feeling of irritability, agitation, and a lowered threshold for arousal. They identify the Brief Irritability Test (BITe; Holtzman, O'Connor, Barata, & Stewart, 2015) as the best existing measure to accomplish these criteria, though they also note shortcomings that include its reliance on tonic items, limiting its ability to distinguish state/trait irritability (similar to the phasic/tonic distinction advocated in the child/adolescent literature). Additionally, the Irritability, Depression, and Anxiety Scale (IDA; Snaith, Constantopoulos, Jardine, & McGuffin, 1978) and the Irritability Questionnaire (IRQ; Craig, Hietanen, Markova, & Berrios, 2008) each contribute some items that capture tonic or phasic irritability, although none of them capture both domains within the same questionnaire. Across these measures, there does not appear to be adequate coverage of tonic and phasic irritability either due to insufficient items (IRQ) or coverage of only tonic or phasic items (BITe and IDA, respectively).

In the child/adolescent literature, Stringaris and colleagues (2018) identify rating scales (parent- and self-report) and semi-structured/structured interviews as the primary tools for measuring irritability, though rating scales offer more concise and expedient results. More specifically, *parent*-report rating scales offer benefits that include greater correspondence with externalizing problems (strongly associated with elevated irritability) and more reliable ratings for young children (when research shows that irritability is highest, and when problematic trajectories begin to appear). Of the parent-report measures that were developed to assess youth irritability, the Affective Reactivity Index (ARI; Stringaris et al., 2012) appears to have the most support for its psychometric properties (internal consistency, test-retest reliability) as a brief (7-item) parent-report

questionnaire (sample age: 5-18); however, its items only cover the tonic component of irritability, meaning it is not an appropriate instrument to comprehensively assess the modern concept of two-dimensional irritability. The Multidimensional Assessment of Preschool Disruptive Behavior Temper Loss scale (MAP-DB; Wakschlag et al., 2012, 2014) provides the most comprehensive coverage of the phasic component of irritability and has demonstrated strong psychometric support (internal consistency) as a parent-report measure. However, it was developed for use within preschool samples (ages 3-5) and requires empirical validation for use within a school-age sample; it also suffers from the same limitation as the ARI in that it only comprehensively captures one dimension of irritability. Finally, The CBCL irritability scale (Tseng et al., 2017) has adequate psychometric support (internal consistency, test-retest reliability) and covers both tonic and phasic irritability (sample age: 9-14), but the short length (3 items) may not capture the full dimensionality of irritability, particularly within tonic and phasic dimensions.

Questionnaires that measure child anger (*subordinate* construct of irritability) contribute additional coverage of tonic and phasic irritability beyond irritability-specific measures. The PROMIS Pediatric Anger Scale (PAS; Irwin et al., 2012) and State-Trait Anger Expression Inventory for children and adolescents (STAXI-CA; del Barrio, Aluja, & Spielberger, 2004) are self-report measures that each contain items that capture both tonic (PAS = 3 items, STAXI-CA = 5 items) and phasic (PAS = 2 items, STAXI-CA = 7 items) irritability. Due to coverage of each domain, these measures already have superior content compared to extant child irritability scales. The PAS has demonstrated good reliability in past research (sample age: 8-17), though published psychometric data only

exists for its self-report version. The STAXI-CA psychometric data was adequate in a Spanish-speaking sample (sample age: 7-17), but this data does not exist within English-speaking samples, where raters may respond differently to items. Taken together, these anger questionnaires require further empirical validation to apply them to irritability assessment for parents' report on their school-age children, though some of their individual items may be useful as indicators for irritability-focused measures, particularly as anger is subsumed under irritability.

Finally, there are additional questionnaires measuring the constructs of emotion (dys)regulation (*superordinate* construct of irritability), emotion reactivity (*related* construct), and temperament (*related* construct) that may contribute to assessments of youth irritability. The Difficulties in Emotion Regulation Scale—Parent Report (DERS-P; Bunford et al., 2018) contains items that measure tonic and phasic irritability and has shown psychometric support. However, the DERS-P is limited in that its items primarily correspond to tonic irritability (six tonic vs. two phasic across subscales), which means it may not be an appropriate measure of the modern, multidimensional conceptualization of irritability. The Emotion Regulation Index for Children and Adolescents (ERICA; MacDermott, Gullone, Allen, King, & Tonge, 2010), which includes items assessing tonic and phasic irritability, has not been validated as a parent-report measure. The Emotion Reactivity Scale (ERS; Nock, Wedig, Holmberg, & Hooley, 2008) has good psychometric support and offers adequate coverage of the tonic component of irritability. However, it does not contain items that measure phasic irritability, and it has only been validated as a self-report measure. Lastly, the parent-report Temperament in Middle

Childhood Questionnaire (TMCQ; Simonds, 2006) captures tonic irritability through its Anger/Frustration subscale, which has been validated within a variety of studies. However, the TMCQ does not contain items that correspond to phasic irritability.

Altogether, there are parent-report measures of school-age child irritability and superordinate (emotion dysregulation), subordinate (anger), and related constructs (emotional reactivity, temperament). However, none of the available measures of youth irritability—regardless of format (rating scale, interview), informant (child, parent, self), or population (children, adolescents, adults)—include items covering tonic *and* phasic irritability (Beauchaine & Tackett, 2020). Despite researchers making repeated calls for the development of tonic and phasic irritability assessment tools, none currently exist. The current study sought to address this significant gap by developing a parent-report measure of irritability among school-age children that includes items corresponding to both tonic and phasic components of irritability.

Current Study

The aims of the current study were to (1) Develop a novel parent-report, dimensional rating scale of youth irritability (including tonic and phasic dimensions), and (2) Evaluate the reliability, factor structure, and convergent validity of the novel scale (as well as tonic and phasic subscales) with other measures of child psychological problems (ODD, internalizing problems, externalizing problems, anxiety, and depression). With regards to the second aim, given prior findings that irritability is associated with youth psychopathology (for review and meta-analysis, see Vidal-Ribas, Brotman, Valdivieso, Leibenluft, & Stringaris, 2016), it was hypothesized that scores on the irritability rating

scale would be moderately, positively associated with internalizing problems, anxiety, and depression, whereas irritability scores would be strongly, positively associated with ODD and externalizing problems. Separate hypotheses were not specified for tonic and phasic irritability given the lack of previous studies' associations between child psychological problems and tonic vs. phasic irritability. Findings generated can begin to provide empirical support for tonic/phasic dimensions of irritability, as well as illuminate distinct relations with other comorbid problems that youth experience. Such information will propel advances in assessment practices and may inform psychosocial interventions that target youth with impairing levels of irritability.

Methods

Participants and Procedure

All research procedures were approved by the Institutional Review Board at Ohio University. Participants were recruited through Amazon's Mechanical Turk (MTurk) platform, a prominent crowdsourcing site that matches participants (otherwise known as "workers") with available surveys or experiments. MTurk has distinct advantages over more traditional methods of data collection, such as access to large, nontraditional samples (e.g., Dworkin, Hessel, Gliske, & Rudi, 2016), and MTurk has been used in a growing number of research studies with children and families that demonstrate the ability to collect high-quality data at a low cost (e.g., Dworkin et al., 2016; Schleider & Weisz, 2015). Recruitment occurred between August and September 2019 in two separate cohorts (n = 200) to allow for comparisons during planned analyses (see Data Analytic Plan). The length of time between completion of cohort 1 and start of cohort 2 was approximately 6 weeks (44 days). Inclusion criteria for the current study restricted participation to those who (1) were 18 years of age or older, (2) were residents of the United States, (3) identified as female, (4) identified as a parent, (5) had at least one child ages 6-12 in their home, and (6) had greater than 95% approval rate for their work on MTurk. For the second cohort, a seventh condition was added so that anyone who had participated in the first cohort was unable to participate in the second cohort. Study measures were completed via a survey link that redirected participants to a secure, web-based data collection platform (i.e., Research Electronic Data Capture [REDCap]; Harris et al., 2009). The average completion time was 20 minutes (range = 13 - 26 minutes) and

workers were compensated $3.00 for their participation. Recruitment materials indicated the study was investigating mothers, children, and behavior.

Figure 1 depicts participant recruitment (screening, eligibility, and completion) in cohorts 1 and 2. Cohort 1 screened 328 participants, 211 of whom were eligible to participate, and 200 (95% of eligible) contributed complete data to the current study. Cohort 2 screened 891 participants, 213 of whom were eligible to participate, and 197 (92% of eligible) contributed complete data to the current study. The completion rate in the overall eligible sample (N = 424) was 94% (n = 397). Most participants (98%) who contributed data to the study analyses answered all five attention check questions correctly, and the remaining passed at least four out of five attention check questions (see Appendix C).[2]

[2] There does not appear to be consensus on the recommended number of attention check questions to employ for online surveys, but review of previous literature indicated studies use anywhere from five items (e.g., Flessner, Murphy, Brennan, & D'Auria, 2017) to 10 items (e.g., Parent, Forehand, Pomerantz, Peisch, & Seehuus, 2017) and allow for at least one incorrect response. Thus, for those who did not pass all five attention check questions (n = 7), individual responses were examined to see if they appeared to be invalid (i.e., provided the incorrect response). Three participants endorsed their gender as "male" on the parent gender attention check item (correct response: "female"); however, due to employing an MTurk qualification that only allowed participants who identified as female to view the task, these responses were deemed to most likely be a mistaken selection, especially in light of them answering remaining attention check questions correctly. Three participants did not provide a response for (i.e., skipped) one attention check question; however, these participants provided correct responses to all other attention check questions. Finally, one participant provided an incorrect response to an attention check question (*Please select the "Almost Never" response option*); however, using the criterion cited above (Flessner et al., 2017; Parent et al., 2017), one incorrect response out of five was considered allowable. Therefore, these respondents were deemed to contribute valid responses and were retained for study analyses.

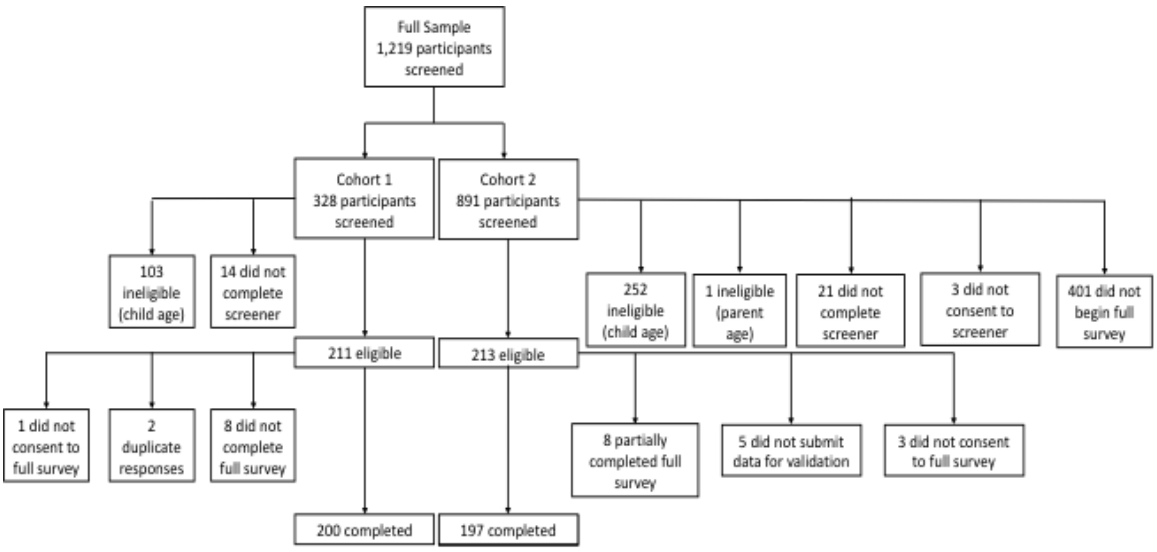

Figure 1. Flow of participant recruitment from screening, eligibility, and study completion.

Demographic characteristics are presented in Table 1 for the full sample and each cohort. The overall sample comprised 397 mothers (mean age in years = 38.21, *SD* = 7.36) of children ages 6-12 (mean age in years = 9.09, *SD* = 2.26). Per the study aims, data collection attempted to recruit equal numbers of mothers of boys and girls, children ages 6-9 and 10-12, and children with and without behavior problems. The full sample (*N* = 397) achieved this goal in regard to child gender, age, and behavior when collapsing across cohorts (each group approximately 50% in both cohorts; see Table 1). The majority of mothers were biological parents (94%), White/European Origin (76%), married (71%), and had attained a bachelor's degree (33%). Most families earned $25,000 to $49,999 (32%) or $50,000 to $74,999 (27%) household yearly income.

Table 1

Demographic Characteristics of Mothers and Children in the Full Sample

Variable	Full Sample	Cohort 1	Cohort 2	χ^2 p value
Parental status				.276
Biological mother	93.5%	93.5%	93.4%	
Stepmother	3.0%	2.5%	3.6%	
Adoptive mother	2.2%	3.5%	1.0%	
Grandmother	1.0%	0.5%	1.5%	
Foster mother	0.3%	-	0.5%	
Race/ethnicity				.672
White/European Origin	76%	74.5%	77.2%	
Biracial/multiracial	9%	10.0%	7.6%	
African American/Black/African Origin	8%	8.0%	8.6%	
Latinx/Hispanic	4%	5.0%	3.6%	
Asian American/Asian Origin/Pacific Islander	3%	2.0%	3.0%	
American Indian/Alaskan Native	<1%	0.5%	-	
Marital status				.688
Married	71.0%	72.5%	69.5%	
Single	15.6%	13.5%	17.8%	
Divorced	9.6%	10.5%	8.6%	
Separated	2.5%	2.0%	3.0%	
Widowed	1.3%	1.5%	1.0%	
Educational attainment				.021
Less than high school degree	1.3%	-	2.5%	
High school graduate or equivalent (GED)	10.3%	11.5%	9.1%	
Some college	23.7%	21.5%	25.9%	

Table 1: continued

Variable	Full Sample	Cohort 1	Cohort 2	χ^2 p value
Associate degree	20.1%	25.5%	14.7%	
Bachelor's degree	33.0%	31.5%	34.5%	
Graduate degree (masters or doctoral)	11.6%	10.0%	13.2%	
Household yearly income				.028
up to $10,000	1.3%	0.5%	2.0%	
$10,001-14,999	1.3%	0.5%	2.0%	
$15,000-24,999	9.3%	6.0%	12.7%	
$25,000-49,999	26.7%	28.0%	25.4%	
$50,000-74,999	31.7%	38.5%	24.9%	
$75,000-99,999	15.1%	14.0%	16.2%	
$100,000-149,999	10.8%	9.5%	12.2%	
$150,000-199,999	3.0%	2.0%	4.1%	
$200,000 or more	0.8%	1.0%	0.5%	
Child gender				.578
Female	51.4%	50.0%	52.8%	
Male	48.6%	50.0%	47.2%	
Child age				<.001
6 to 9 years old	51.6%	61.0%	42.1%	
10 to 12 years old	48.4%	39.0%	57.9%	
Child behavior				<.001
No behavior problems	48.1%	36.0%	60.4%	
Behavior problems	51.9%	64.0%	39.6%	

Note. Full sample N = 397. Cohort 1 n = 200. Cohort 2 n = 197.

In comparisons between cohort 1 and cohort 2, mothers did not differ significantly on parental status, race/ethnicity, or marital status (χ^2 *ns*). However, participants differed on educational attainment and household yearly income (χ^2 $p < .05$), such that cohort 1 reported higher rates of attaining an associate's degree or higher and earning yearly income greater than $50,000, as compared to cohort 2.[3] Further, child participants in cohorts 1 and 2 did not differ significantly on child gender (χ^2 *ns*). However, cohorts differed on child age, $\chi^2(1) = 14.148, p < .001$, such that cohort 1 had more 6- to 9-year-olds, and cohort 2 had more 10- to 12-year-olds. In addition, cohorts differed on behavior problems, $\chi^2(1) = 23.680, p < .001$, such that cohort 1 had more children with behavior problems, and cohort 2 had more children with no behavior problems. Given these cohort differences on child age and behavior, as well as parent education and income, and the recommendation to replicate results of scale development in different samples (e.g., Boateng et al., 2018; DeVellis, 2012), all item-level analyses, exploratory factor analyses (EFA), and reliability analyses were initially performed separately by cohort. Confirmatory factor analyses (CFA) were then performed prior to structural equation modeling (SEM) to specify the best-fitting model

[3] Socioeconomic status (SES) has been associated with child externalizing behaviors (i.e., aggression, disruptive behavior, or delinquency) in prior research (e.g., Kraatz Keiley, Bates, Dodge, & Pettit, 2000), although studies that examine youth irritability have shown mixed evidence for this association: some studies find that characteristics of SES such as poverty status (Copeland et al., 2013; Copeland, Shanahan, Egger, Angold, & Costello, 2014) or educational attainment (Wiggins et al., 2014) predict higher levels of youth irritability, whereas other studies find that parent income (Althoff et al., 2016) or parent education (Carlson, Danzig, Dougherty, Bufferd, & Klein, 2016; Dougherty et al., 2014) do not differentially predict irritability. Given these conflicting findings, educational attainment and household yearly income were not statistically accounted for in planned analyses; however, future research that uses empirically-validated measures of irritability may wish to clarify the role of SES as a predictor variable.

for the full sample. Cohort demographic differences were accounted for by testing measurement invariance (MI) in the full sample CFA models (see Data Analytic Plan).

Initial Scale Development (Item Selection and Retention)

As the chief component of Study Aim 1, guidelines for empirical scale development (Boateng et al., 2018; DeVellis, 2012) were followed to create the irritability rating scale evaluated in the current study. Relevant research was reviewed to identify the underlying construct, or latent variable, to be measured by the scale (irritability) and contemporary approaches to its measurement (general factor vs. tonic and phasic factors). Then, an initial item pool (three to four times larger than the final scale; DeVellis, 2012) was generated from review of psychometrically-sound instruments measuring a variety of relevant constructs in children and adults, including anger, emotion reactivity, emotion (dys)regulation, irritability, temper loss, and temperament. The initial item pool (N = 76) and instruments of origin appear in Appendix A; Appendices D and E include items organized by tonic and phasic domains, respectively. Thirty items (denoted in Appendices D and E) were not included in the irritability scale because their authors indicated that replicating individual items without receiving written permission and/or without paying per use to administer the instrument would violate the copyright terms. These items came from two measures of anger (PAS, STAXI-CA), two measures of irritability (ARI, CBCL), and one measure of temperament (TMCQ); it is possible that their inclusion (in particular those from the ARI and CBCL) would have affected results given the addition of eight tonic items that had already been examined in

samples of irritable youth. Nonetheless, 46 candidate items (22 tonic, 24 phasic) were retained in the initial item pool.

Next, four individuals who were a combination of parents, clinicians, or academicians with experience and/or expertise in child psychopathology, research methods, and scale development reviewed the initial item pool and provided feedback to (1) help assess the relevance of items to the construct the scale intends to measure and (2) offer input on the clarity and conciseness of individual items (DeVellis, 2012). The irritability scale expert panel survey was administered via REDCap (Harris et al., 2009) to collect information about its written instructions to respondents, the response scale, and individual items. Each item was rated on one categorical domain ("*Does this item tap the construct of irritability?*" [yes/no]), and if participants endorsed "yes," they rated the item on two continuous domains ("*How clear is this item? In responding, consider if a parent would understand it and how to answer it*", "*Do you believe, given a sufficient sample, that there are parents who will use the entire scale of responses for this item?*" [0-100]). Items were retained if the following conditions were satisfied (two items notwithstanding; see further discussion below): (1) if they were rated by consensus (≥ 75% agreement) as measuring the construct of irritability (for discussion on quantifying consensus, see Lawshe, 1975; Lynn, 1986), (2) if their average rating for clarity and/or variability was ≥ 50 (on a scale from 0 to 100; e.g., Dawson, 2013), and (3) if they were not deemed problematic per participants' qualitative feedback. Some examples of qualitative feedback that resulted in revision or omission of items included awkward phrasing of item, difficult interpretation of item due to wordiness, and vocabulary beyond

average adult reading level. Nineteen items (7 tonic, 12 phasic) were excluded due to failing to achieve consensus agreement. Two items, one tonic and one phasic, surpassed consensus and qualitative review, but did not pass clarity or variability 50% thresholds (tonic item: mean scores of 36 and 31; phasic item: 46 and 45, respectively). The decision was made to retain these items. Although this departed from one out of three of the item retention conditions specified above, DeVellis (2012) notes that final decisions for item retention are at the discretion of the scale developer, so retaining these items was felt to be permissible. All other items (n = 25) met the criteria above for consensus construct validity, clarity/variability ratings, and qualitative feedback. The final set of items (N = 27; including 15 tonic and 12 phasic irritability items) comprised the irritability rating scale administered to parent participants (Appendix F).

Measures

Predictors

Brief Screener (Appendix G). Mothers completed a three-item screener to determine that they were eligible to participate. The screening items asked mothers to report on their age, the number of children they had between ages 0 to 18 years old, and whether they or their child's teacher thought the child had behavior problems. Parent-reported behavior problems was used to classify children into groups with and without clinically significant behavior problems to predict patterns of psychological problems in discrimination analyses (Study Aim 2).

Parent Demographic Questionnaire (Appendix H). Mothers completed demographic information about themselves, including their age, gender, race/ethnicity,

relationship to child, marital status, number of children, educational attainment, and annual household income. Mothers also completed demographic information about their child, including the child's age, gender, race/ethnicity, grade level, and diagnostic history (i.e., any psychiatric/mental health diagnoses). Child age and gender were included as predictors to test measurement invariance in multi-group CFA models (Study Aim 2).

Impairment Rating Scale (IRS; Appendix I). Participants completed the IRS (Fabiano et al., 2006) to briefly assesses child impairment across multiple domains (i.e., relationship with peers, relationship with siblings, relationship with parents, academic progress, self-esteem, influence on family functioning, and overall impairment). Raters are instructed to put an "X" on a line for each domain that represents the child's impairment, from "No Problem, definitely does not need treatment or special services" to "Extreme Problem, definitely needs treatment or special services." Each line is broken into seven segments and scored from 0 to 6, with scores of 3 or more considered clinically significant. The IRS has shown test-retest reliability scores ranging from .54 to .89 over a period of six months to one year, and there is evidence that it has concurrent, convergent, and discriminant validity with other measures of functional impairment (Fabiano et al., 2006). The IRS (overall impairment item) was used to examine evidence of functional impairment to classify children into groups with and without clinically significant behavior problems to predict patterns of psychological problems in discrimination analyses (Study Aim 2).

Outcomes

The following measures were examined in relation to the developed irritability measure as part of validation analyses (Study Aim 2).

Disruptive Behavior Disorders Rating Scale (DBD; Appendix J). Participants completed an amended version of the DBD (Pelham, Gnagy, Greenslade, & Milich, 1992) to measure *DSM-IV* (APA, 1994) ODD symptoms. The DBD ODD was used to evaluate convergent validity with the irritability rating scale. Items are rated on a 4-point scale (0 = not at all, 1 = just a little, 2= pretty much, 3 = very much), and items that are rated as occurring "pretty much" or "very much" by parents are considered clinically significant. The DBD has shown good psychometric properties with excellent reliability (α = .87-.95) for the ODD composite score (Pelham et al., 1992; Pillow, Pelham, Hoza, Molina, & Stultz, 1998). Two items from the DBD ODD composite measure tonic irritability ("Is often angry and resentful" and "Is often touchy or easily annoyed by others") and one item measures phasic irritability ("Often loses temper"). To reduce inflated associations between the irritability rating scale and DBD ODD scores, the three items were excluded from the composite used in analyses. The DBD ODD composite score (sum of remaining five items) demonstrated acceptable reliability in the current study (Cronbach's α = .86).

Strengths and Difficulties Questionnaire (SDQ). Participants completed the 25-item SDQ (Goodman, 1997) to screen for emotional problems, conduct problems, hyperactivity-inattention, peer problems, and prosocial behavior. The SDQ was used to evaluate convergent validity with the irritability rating scale. Items are rated on a 3-point Likert-type scale (0 = not true, 1 = somewhat true, 2 = certainly true). An externalizing

composite score (conduct problems + hyperactivity-inattention) and an internalizing score (emotional problems + peer problems) were calculated. Goodman, Lamping, and Ploubidis (2010) found acceptable internal consistency (α = .73-.78) for the externalizing and internalizing composite scores within a large, population-based sample of youth ages 5 to 16. One item from the SDQ externalizing scale ("Often loses temper") measures phasic irritability and was not included in the SDQ externalizing scale used in analyses to reduce inflated associations. The SDQ externalizing (Cronbach's α = .83) and internalizing (Cronbach's α = .82) composites showed acceptable reliability in the current study.

Revised Children's Anxiety and Depression Scale 25—Parent Version (RCADS-P-25). Participants completed the 25-item RCADS-P-25 (Ebesutani, Korathu-Larson, Nakamura, Higa-McMillan, & Chorpita, 2017) to assess *DSM-IV* anxiety and depression symptoms. The RCADS-P-25 was used to evaluate convergent validity with the irritability rating scale. Items are rated on a 4-point Likert-type scale (0 = never, 1 = sometimes, 2 = often, 3 = always). The RCADS-P-25 yields an Anxiety composite score (15 items) and Depression composite score (10 items). The RCADS-P-25 has shown acceptable internal consistency (α = .80-.90), test-retest reliability (*r*s = .77-.90), and convergent and discriminant validity with another measure of child psychological problems (Ebesutani et al., 2017). The RCADS-P-25 Anxiety (Cronbach's α = .86) and Depression (Cronbach's α = .87) composites both showed acceptable reliability in the current study.

Data Analytic Plan

Study Aim 1

Develop A Parent-Report Irritability Rating Scale. *Item Evaluation*. Items included in the irritability rating scale were evaluated at the item level. First, Pearson's correlations were calculated among all items. An ideal scale contains items that are highly, positively intercorrelated, suggesting that items are measuring a shared latent construct (DeVellis, 2012). Negative correlations among items were examined to determine if reverse scoring was appropriate; no items met this criterion. Next, item-scale correlations were calculated using the corrected item-scale correlation, which correlates an item with all the other scale items excluding itself. High item-scale correlations indicate a desirable scale (DeVellis, 2012), whereas items with weaker item-scale correlations (i.e., less than .40) should be considered for dropping to enhance the scale's ability to reliably measure the latent construct of interest. Finally, variances and means were computed for all items. Good items should demonstrate high variance, indicating the ability to differentiate among different levels of the latent construct being measured (DeVellis, 2012). Good items should also have a mean close to the center of the range of possible values, as opposed to a mean near the extreme low or high values (DeVellis, 2012).

Exploratory Factor Analysis (EFA). After completing item-level analyses, EFA was performed to determine whether the retained items measured a unidimensional latent construct (e.g., general irritability factor) or a multidimensional latent construct (e.g., tonic and phasic irritability factors). EFA analyses were performed in Mplus Version 8.4

(Muthén & Muthén, 2019) specifying items on the irritability scale as categorical indicators using a robust weighted least squares mean and variance adjusted (WLSMV) estimator. Mplus uses a GEOMIN oblique rotation as the default method of rotation in EFA analyses, which assumes that factors are correlated and is recommended for most cases when conducting factor analysis (Fabrigar, Wegener, MacCallum, & Strahan, 1999). Models from one to three factors were compared using the χ^2 difference test (a nonsignificant test indicating that the more parsimonious model, the one with fewer factors, should be retained). In addition, the χ^2 value, the comparative fit index (CFI), the root mean square error of approximation (RMSEA) with accompanying 90% confidence intervals (CI), and the SRMR (standardized root mean square residual) were used to assess global model fit (Kline, 2016). Indices of acceptable model fit include nonsignificant χ^2, CFI $\geq$ .95, RMSEA $\leq$ .05 (close fit), and SRMR $\leq$ .08 (Boateng et al., 2018; Browne & Cudeck, 1993; Hu & Bentler, 1999; Kline, 2016). Models were considered fair fit (i.e., less than acceptable) if .05 $\leq$ RMSEA $\leq$ .08 and marginal fit (i.e., less than fair) if .08 $\leq$ RMSEA $\leq$ .10 (Browne & Cudeck, 1993). Items with factor loadings $\geq$ .32 that did not load $\geq$ .32 on more than one factor (i.e., cross-loading) were retained (e.g., Judah, Saulnier, Hager, & Allan, 2020; Tabachnick & Fidell, 2007), and factors were considered interpretable if they contained at least three items (Costello & Osborne, 2005).

Reliability Analysis. The resultant group of items was evaluated for their internal consistency by calculating Cronbach's alpha (α) for the entire scale (i.e., all items) as well as any subscales identified through factor analysis. Alpha values greater than .70

were considered acceptable, with values greater than .80 desirable (DeVellis, 2012). In addition, coefficient omega was calculated as an alternative to coefficient alpha, which has been identified as a biased estimate of internal consistency (for review, see Dunn, Baguley, & Brunsden, 2014). Omega is a measure of internal consistency that can be calculated along with bootstrapped confidence intervals (CI) to provide a distribution of possible values (e.g., 95% CI). Calculation was performed in the *R* statistical software (R Core Team, 2020) using the MBESS package (Kelley, 2019). Both alpha and omega values were compared as evidence of internal consistency.

Study Aim 2

Explore Revised Irritability Rating Scale Validity. *Measurement Invariance*. Once the optimal EFA solution was determined, multi-group confirmatory factor analysis (MG-CFA; e.g., Pendergast, von der Embse, Kilgus, & Eklund, 2017) was performed to explore measurement invariance (MI) across groups according to child age (6-9 years, 10-12 years) and child gender (females, males). Per recommendations for testing MI (Kline, 2016; Pendergast et al., 2017), configural, metric (weak), and scalar/threshold (strong) invariance was tested to determine if the factor structure, factor loadings, and factor loadings and thresholds (respectively) differed across groups. Global model fit was assessed using the indices (χ^2, CFI, RMSEA, SRMR) reviewed above. Change in χ^2 (*ns*), change in CFI$< .01$, and change in RMSEA $< .015$ were used to provide evidence of measurement invariance when comparing models (Chen, 2007). Models were estimated using WLSMV and theta parameterization for categorical data.

Validation Analysis. Structural equation modeling (for overview, see Kline, 2016) was used to model relationships between irritability ratings and dimensional measures of psychological problems, which included the DBD ODD composite, internalizing and externalizing composites of the SDQ, and anxiety and depression composites of the RCADS-P-25 (see Measures). The SEM model permitted items on the exogenous variable (irritability scale) to load onto different factors (tonic and phasic, per results of EFA), and structural regression coefficients were estimated to identify relationships with the endogenous variables (DBD, SDQ, and RCADS-P-25) to explore convergent validity. Any groups (e.g., girls vs. boys, 6- to 9-year-olds vs. 10- to 12-year-olds) found to have noninvariant scores on the irritability rating scale were included as covariates (i.e., predictors) in the model. Global model fit was assessed using the indices (χ^2, CFI, RMSEA, SRMR) reviewed above.

Discrimination Analyses. To enhance the practical application of the irritability rating scale, discrimination analyses were performed to identify scale summary score cut points to categorize individuals based on the severity of their irritability. First, the distribution of scores on the irritability rating scale was examined to explore whether groups of children with significantly higher scores existed. These groups were operationalized as those with scores < 1 *SD* (low risk; *n* = 286-299), ≥ 1 *SD* (moderate risk; *n* = 40-43), and ≥ 2 *SD* (high risk; *n* = 7-9) above the mean score. Although the high-risk group contained a smaller number of individuals relative to the low and moderate groups, it met the minimum criterion of the sample in each cell being greater than the number of dependent variables included (five; Hair, Anderson, Tatham, & Black,

1995). Standard scores are often used in measures of child behavior (e.g., Child Behavior Checklist; Achenbach & Rescorla, 2007) to operationalize groups (e.g., those with borderline clinical concerns vs. clinically significant concerns), and as no prior studies have specified classification ranges for tonic/phasic irritability, the ranges above were selected as a more conservative criterion. Chi square tests were conducted to determine if gender or age differences existed among groups. Finally, one-way multivariate analysis of variance (MANOVA) tests were performed to examine the ability for membership in a moderate- or high-risk group (vs. low-risk) to predict different scores on measures of dimensional psychological problems (DBD ODD composite, internalizing and externalizing composites of the SDQ, and anxiety and depression composites of the RCADS-P-25).

In addition, conditional probabilities were calculated to determine the ability of specific items on the irritability rating scale to differentiate between clinically significant and nonsignificant groups (e.g., Fabiano et al., 2006; Milich, Widiger, & Landau, 1987; for review, see Trevethan, 2017). Children were classified into groups with and without clinically significant behavior problems based on two indicators: (1) endorsement of the parent-reported behavior problems screening item ("yes" response to brief screener item 3 [Appendix G]) and (2) score of 3 or higher on the overall impairment item from the IRS, which is an indicator of clinically significant impairment (Fabiano et al., 2006). Scores on the irritability rating scale were explored for their ability to predict children with behavior problems ("yes" behavior problems + "yes" impairment) and children without behavior problems ("no" behavior problems + "no" impairment). Items on the

irritability scale endorsed as 4 ("often") or 5 ("always") were counted as a symptom being present. Conditional probabilities calculated for each item included: Base Rate (*N* with symptom/Total *N*); Sensitivity (*N* with behavior problems who show symptom/*N* with behavior problems); Specificity (*N* without behavior problems who do not show symptom/*N* without behavior problems); Positive Predictive Power (PPP; *N* with behavior problems who show symptom/*N* with symptom); and Negative Predictive Power (NPP; *N* without behavior problems who do not show symptom/*N* without symptom).

Results

Study Aim 1: Develop a Parent-report Irritability Rating Scale

Exploratory Factor Analysis of Irritability Rating Scale

Cohort 1. Item-level statistics were computed for the irritability rating scale, including Pearson's correlations, item-scale correlations, and means/variances (see Appendix K). All items (n = 27) were significantly, positively correlated ($r = .32 - .91$, all ps $< .001$). Item-scale correlations were high ($r = .56 - .85$) and item mean values (grand mean = 2.64, range = 2.07 - 3.25) were close to the scale midpoint (score of 3). No items contained problematic skew (values > 3.0) or kurtosis (values > 10.0). All items were deemed appropriate to include in EFA analyses.

Data were fit to a one-, two-, or three-factor model. The three-factor model fit the data better than the two-factor model ($\Delta\chi^2 = 194.60, df = 25, p < .001$). However, only three items (item 1 [$\lambda = .94$], item 2 [$\lambda = 1.02$], and item 11 [$\lambda = .35$]) loaded above .32 on the first factor, and despite meeting the minimum threshold for an interpretable factor (e.g., Costello & Osborne, 2005), these items did not appear to

constitute a distinct construct.[4] Therefore, the two-factor model was examined. For this model, only two items (item 1 [$\lambda = .76$] and item 2 [$\lambda = .79$]) loaded on the first factor, and these items exhibited cross-loadings above .32 with the second factor. The decision was made to drop items 1 and 2 from the EFA model due to their failure to constitute an interpretable factor, apparent influence on the overall scale, and high inter-item correlation ($r = .91$).

EFA analyses were rerun without items 1-2 using the same specifications (one- vs. two- vs. three-factor model). The three-factor model fit the data better than the two-factor model ($\Delta\chi^2 = 174.50, df = 23, p < .001$). However, the only items on factor three that loaded above .32 (item 24 [$\lambda = .38$], item 25 [$\lambda = .41$], and item 26 [$\lambda = .47$]) also exhibited cross-loadings above .32 with factor two. Therefore, the two-factor model was examined. For this model, the χ^2 was significant ($769.01, df = 251, p < .001$) and the RMSEA was .10 (90% CI [.09, .11]), inconsistent with good fit. However, the CFI was .98 and the SRMR was .05, both of which indicate acceptable model fit (Kline, 2016). All items (n = 12) on factor one loaded greater than .46, and all items (n = 7) on factor two loaded greater than .55. The items not retained (n = 6) all demonstrated cross-loadings greater than .32 on both factors.

Cohort 2. Item-level statistics (Pearson's correlations, item-scale correlations, and means/variances) were computed for the irritability rating scale (see Appendix L). All items (n = 27) were significantly, positively correlated ($r = .27 - .90$, all ps <

[4] Item 1: "When my child is upset, s/he becomes angry with him/herself for feeling that way." Item 2: "When my child is upset, s/he becomes irritated with him/herself for feeling that way." Item 11: "Other people get on my child's nerves."

.001). Item-scale correlations were high ($r = .49 - .87$) and item mean values (grand mean = 2.45, range = 1.84 – 3.16) were close to the scale midpoint (score of 3). No items contained problematic skew or kurtosis. All items were deemed appropriate to include in EFA.

As in cohort 1, data were fit to a one-, two-, or three-factor model. The three-factor model fit the data better than the two-factor model ($\Delta\chi^2 = 125.37, df = 25, p <$ .001). However, only two items (item 1 [$\lambda = 1.03$] and item 2 [$\lambda = .84$]) loaded above .32 on the first factor, which does not constitute an interpretable factor. Therefore, the two-factor model was examined. For this model, the same two items (item 1 [$\lambda = .98$] and item 2 [$\lambda = .95$]) loaded on the first factor, and all other items loaded on factor two. Similar to cohort 1, the decision was made to drop items 1 and 2 from the EFA model due to their failure to constitute an interpretable factor, apparent influence on the overall scale, and high inter-item correlation ($r = .90$).

EFA analyses were rerun without items 1-2 using the same specifications (one- vs. two- vs. three-factor model). The three-factor model fit the data better than the two-factor model ($\Delta\chi^2 = 103.22, df = 23, p <$ .001). However, no items loaded above .32 on either the second or third factor. Therefore, the two-factor model was examined. For this model, the χ^2 was significant ($452.43, df = 251, p <$ .001), inconsistent with good fit. However, the RMSEA was .06 (90% CI [.05, .07]), the CFI was .99, and the SRMR was .04, all of which indicate acceptable fit. All items (n = 10) on factor one loaded greater than .56, and all items (n = 13) on factor two loaded greater than .39. The items not retained (n = 2) demonstrated cross-loadings greater than .32 on both factors.

Factor Structure Clarification. Results of EFA analyses supported a 19-item, two-factor solution in cohort 1 and a 23-item, two-factor solution in cohort 2. Theoretically, the 23-item scale appeared to replicate the hypothesized tonic and phasic dimensions of irritability more appropriately: the first 10-item subscale was exclusively composed of tonic items identified during scale development, and the second 13-item subscale was mostly composed of phasic items (two items from the tonic pool notwithstanding). Conversely, the 19-item scale appeared a worse theoretical fit to tonic and phasic irritability: the first 12-item subscale was composed of both tonic (*n* = 8) and phasic (*n* = 4) items, and the second 7-item subscale was composed of both tonic (*n* = 2) and phasic (*n* = 5) items.

To examine which solution was more empirically appropriate (i.e., better fit to the data), CFA models were fit to the data using the 19-item solution and the 23-item solution for both cohorts. Model fit statistics were inspected to evaluate differences in model fit. In cohort 1, both the 19-item solution, $\chi^2(151) = 453.55, p < .001$, CFI = .98, RMSEA = .10 (90% CI [.09, .11]), SRMR = .04, and the 23-item solution, $\chi^2(229) = 659.30, p < .001$, CFI = .98, RMSEA = .10 (90% CI [.09, .11], SRMR = .04, indicated acceptable model fit. Model fit statistics (CFI, RMSEA, and SRMR) were identical when rounding to the hundredth (.01). In cohort 2, both the 19-item solution, $\chi^2(151) = 305.90, p < .001$, CFI = .99, RMSEA = .07 (90% CI [.06, .08]), SRMR = .03, and the 23-item solution, $\chi^2(229) = 411.16, p < .001$, CFI = .99, RMSEA = .06 (90% CI [.05, .07]), SRMR = .03, indicated acceptable model fit. RMSEA was slightly

improved for the 23-item solution, but the other fit indices (CFI and SRMR) were identical.

Notably, change in CFI ($\Delta CFI < .01$) and change in RMSEA ($\Delta RMSEA < .015$) have been recommended when evaluating change in model fit in multi-group analysis (e.g., Pendergast et al., 2017). Although these CFA analyses were not comparing groups using the same model specification (19-item vs. 23-item), which does not permit chi-square difference testing (i.e., requires nested models), ΔCFI and ΔRMSEA demonstrated favorable comparisons between the 19-item and 23-item solutions (cohort 1: ΔCFI = .001, ΔRMSEA = .003; cohort 2: ΔCFI = .002, ΔRMSEA = .008). Given the converging results of theoretical and empirical indices, the decision was made to select the 23-item scale for planned analyses. Consistent with theorized dimensions of irritability, Factor 1 was titled “Tonic” and Factor 2 was titled “Phasic.” Item loadings for the 23-item solution in cohort 2 are provided in Table 2.

Table 2

Irritability Rating Scale and Factor Loadings for the Two-Factor Exploratory Factor Analysis

Items	Tonic λ	Phasic λ
3. My child gets angry at people very easily.	.81*	.08
4. My child is easily agitated.	.99*	-.13
5. When something bad happens, my child's mood changes very quickly. People tell me s/he has a very short fuse.	.63*	.28*
6. When things don't go my child's way s/he gets upset easily.	.63*	.26
7. My child gets angry when adults tell him/her what s/he can and cannot do.	.69*	.11
8. My child has days at a time where s/he is touchy and gets angry easily.	.62*	.22
9. My child is grumpy.	.78*	-.00
10. My child feels like s/he might snap.	.61*	.32*
11. Other people get on my child's nerves.	.85*	-.07
12. Things bother my child more than they normally do.	.75*	.13
13. My child feels frustrated.	.57*	.24*
14. When my child is irritated, s/he needs to vent his/her feelings immediately.	.16	.40*
15. My child gets extremely angry.	.24*	.73*
16. When my child is upset, s/he loses control over his/her behaviors.	-.11	.99*
17. My child has angry outbursts.	.15	.80*
18. My child loses his/her temper and shouts or snaps at others.	.09	.85*
19. My child slams doors or bangs about when people upset him/her.	.17	.67*
20. My child feels like a bomb, ready to explode.	.29*	.66*
21. My child has temper tantrums or melt-downs.	-.00	.91*
22. My child loses their temper or has a tantrum with other adults (e.g., teacher, babysitter, family member).	.01	.86*
23. My child loses their temper or has a tantrum when frustrated, angry, or upset.	-.04	.96*
24. My child loses their temper or has a tantrum when tired, hungry, or sick.	.02	.71*

Table 2: continued

Items	Tonic λ	Phasic λ
25. My child loses their temper or has a tantrum during daily routines, such as bedtime, mealtime, or getting dressed.	-.02	.84*
26. My child loses their temper or has a tantrum out of the blue or for no reason.	.32*	.56*
27. My child has a hot or explosive temper.	.24*	.72*

Note. Factor loadings are provided from the solution in cohort 2 (n = 197). Significant factor loadings are bolded. Factors with loadings $\geq$ .32 on both factors were considered cross loadings.

* $p < .05$

Reliability Analyses

All items (n = 23) retained from EFA analyses were evaluated for their internal consistency by calculating Cronbach's alpha (α) for the full scale and both subscales identified through factor analysis. Item-scale analyses revealed that, for both cohorts 1 and 2, the overall scale and subscale alpha coefficients would increase if item 14 was deleted. Of note, this item showed the only factor loading $<$ .40 in cohort 2 (λ = .397). The item ("When my child is irritated, s/he needs to vent his/her feelings immediately") loaded onto the phasic factor despite being identified as tonic during scale development, and given that factor loadings $\geq$ .40 provide a more stringent criterion for item retention during factor analysis (Boateng et al., 2018), the decision was made to drop it from the scale. No other items were identified through item-scale analyses as improving the overall scale, or subscale, reliability by deletion in both cohorts.

The full scale score (sum of 22 items; range = 22 - 110) demonstrated excellent reliability in cohort 1 and cohort 2 (both αs = .97). The first subscale score (sum of 10 items; range = 10 - 50), which comprised items solely focusing on tonic components of irritability, showed excellent reliability in cohort 1 ($\alpha = .95$) and cohort 2 ($\alpha = .94$). The second subscale score (sum of 12 items; range = 12 - 60), which comprised items focusing on phasic components of irritability, also showed excellent reliability in cohort 1 ($\alpha = .96$) and cohort 2 ($\alpha = .97$). Scores on the Tonic subscale were strongly, positively correlated with scores on the Phasic subscale in cohort 1 and cohort 2 (both rs = .88). As expected, reliability coefficients in the full sample (N = 397) for the tonic scale

($\alpha = .94$), phasic scale ($\alpha = .96$), and full scale ($\alpha = .97$) were excellent, and scores on the tonic and phasic subscales were strongly related (r = .88).

Coefficient omega was calculated separately for each subscale (tonic and phasic) given Dunn and colleagues' (2014) recommendation to treat scales as unidimensional (i.e., it would not be appropriate to calculate omega for a bidimensional index [total scale score]). In cohort 1, the tonic (ω = .95, 95% CI [.94, .96]) and phasic (ω = .96, 95% CI [.95, .97]) subscales both demonstrated excellent internal consistency. In cohort 2, the tonic (ω = .94, 95% CI [.93, .95]) and phasic (ω = .97, 95% CI [.96, .97]) subscales also demonstrated excellent internal consistency. The full sample also showed excellent internal consistency for the tonic (ω = .95, 95% CI [.94, .95]) and phasic (ω = .96, 95% CI [.96, .97]) subscales. Of note, these internal consistency values were nearly identical to those calculated with Cronbach's alpha, suggesting the high internal consistency of the subscales.

Study Aim 2: Explore Revised Irritability Rating Scale Validity

Confirmatory Factor Analysis Exploring Measurement Invariance

CFA models for the 22-item irritability rating scale were conducted to examine measurement invariance. The two-factor CFA in cohort 1 fit the data adequately, $\chi^2(208) = 640.72, p < .001$, CFI = .98, RMSEA = .10 (90% CI [.09, .11]), SRMR = .04, and showed better fit to the data than a one-factor solution, $\Delta\chi^2 = 57.10, df = 1, p < .001$. The two-factor CFA in cohort 2 also fit the data adequately, $\chi^2(208) = 385.16, p < .001$, CFI = .99, RMSEA = .07 (90% CI [.06, .08]), SRMR = .03, and showed better fit to the data than a one-factor solution, $\Delta\chi^2 = 51.60, df = 1, p < .001$.

The two-factor CFA in the full sample showed analogous results with adequate fit to the data, $\chi^2(208) = 869.54, p < .001$, CFI = .98, RMSEA = .09 (90% CI [.08, .10]), SRMR = .03, and showed improved fit over the one-factor solution, $\Delta\chi^2 = 110.49, df = 1, p < .001$.

Given requirements for large samples with at least 150 participants in each group (e.g., Chen, 2007), MG-CFA analyses were only conducted using the full sample (*N* = 397). Table 3 displays results from CFA findings. Beginning with child age, models assessing configural, metric, and scalar/threshold invariance were tested. In the first step, configural invariance was established (CFI = .984, RMSEA = .089), showing that the same general factor structure holds for children ages 6 to 9 and 10 to 12. In the second step, metric invariance was tested wherein factor loadings were constrained to be equal across ages. This model fit the data adequately (CFI = .984, RMSEA = .087) and indeed showed improved fit to the configural invariance model ($\Delta\chi^2$ *ns*, ΔCFI < .01, ΔRMSEA < .015). Finally, scalar/threshold invariance was tested whereby factor loadings and thresholds were constrained to be equal across groups. This model fit the data adequately (CFI = .985, RMSEA = .078) and again showed improved fit over the metric invariance model ($\Delta\chi^2$ *ns*, ΔCFI < .01, ΔRMSEA < .015). Together, these analyses support configural, metric, and scalar invariance for children ages 6 to 9 and 10 to 12. A two-factor model of irritability appears to capture this construct in children ages 6-12, and among children within this age range, they demonstrate similar levels of tonic/phasic irritability and respond similarly to individual items. As such, child age was not included as a covariate in the SEM analyses.

Table 3

Fit Statistics for Measurement Invariance of Irritability Scores across Groups

Model	Retained?	χ^2	*df*	Model comparison	$\Delta\chi^2$	Δdf	RMSEA (90% CI)	CFI
Age								
1. Configural	Y	1076.60**	416	-	-	-	.089 [.083, .096]	.984
2. Metric	Y	1096.40**	436	2 vs. 1	26.08	20	.087 [.081, .094]	.984
3. Scalar/threshold	Y	1098.13**	500	3 vs. 2	63.62	64	.078 [.071, .084]	.985
Gender								
1. Configural	Y	1023.49**	416	-	-	-	.086 [.079, .092]	.986
2. Metric	Y	1040.99**	436	2 vs. 1	22.79	20	.084 [.077, .090]	.986
3. Scalar/threshold	N	1089.56**	500	3 vs. 2	108.34*	64	.077 [.071, .083]	.986

Note. CI = confidence interval. All results were computed by Mplus with theta parameterization.

* $p < .01$; ** $p < .001$.

Next, models assessing configural, metric, and scalar/threshold invariance across child gender were tested (see Table 3). In the first step, configural invariance was established (CFI = .986, RMSEA = .086), showing that the same general factor structure holds for males and females. In the second step, metric invariance was tested wherein factor loadings were constrained to be equal across genders. This model fit the data adequately (CFI = .986, RMSEA = .084) and indeed showed improved fit to the configural invariance model ($\Delta\chi^2$ *ns*, ΔCFI < .01, ΔRMSEA < .015). Finally, scalar/threshold invariance was tested whereby factor loadings and thresholds were constrained to be equal across groups. This model fit the data adequately (CFI = .986, RMSEA = .077), although the chi-square difference test was significant, $\Delta\chi^2$ = 108.34, *df* = 64, $p < .001$. Therefore, scalar/threshold invariance was not supported, suggesting that factor thresholds vary across females and males (i.e., although females and males demonstrate comparable levels of tonic/phasic irritability, they appear to respond differently to specific items measuring these constructs).

When scalar/threshold invariance is not supported, modification indices can be used as an exploratory approach to improving model fit by relaxing (freely estimating) parameters that are being held to equality across groups. There are important caveats to this approach: it is recommended to have solid theoretical justification for model re-specifications, and any specification searches are likely to be more stable in larger sample sizes ($N \geq 500$) with less complex models (Silvia & MacCallum, 1988). Each of these guidelines are difficult to meet in the current study: theoretical justification for multidimensional measurement of irritability is novel (hence the purpose of the current

study), and the overall sample size is less than recommended. As an alternative to exploring model modification indices, child gender was included as a covariate in SEM analyses to measure its potential influence on results.

Structural Equation Modeling

Descriptive statistics and correlations of the variables included in the SEM model are provided in Table 4. One item on the SDQ and three items on the RCADS-P-25 demonstrated problematic skew (values $>$ 3.0) and/or kurtosis (values $>$ 10.0; Kline, 2016). These items were normalized using a logarithmic transformation and showed acceptable values. All other variables showed acceptable skew/kurtosis.

Table 4

Means and Intercorrelations between Variables in Structural Equation Model

	Gender	Irr-T	Irr-P	ODD	SDQ Int	SDQ Ext	RCADS-A	RCADS-D
Gender	--							
Irr-T	.04	--						
Irr-P	.04	.88*	--					
ODD	.09	.71*	.74*	--				
SDQ Int	-.01	.59*	.52*	.49*	--			
SDQ Ext	.09	.67*	.67*	.69*	.55*	--		
RCADS-A	-.08	.53*	.50*	.44*	.70*	.43*	--	
RCADS-D	-.01	.62*	.60*	.62*	.72*	.57*	.77*	--
Mean	.49	27.92	28.44	4.75	5.79	6.43	7.82	5.18
SD	.50	8.72	10.86	3.74	4.40	4.01	6.23	4.99

Note. Gender (0 = female, 1 = male); Irr-T = Irritability Rating Scale-Tonic subscale; Irr-P = Irritability Rating Scale-Phasic subscale; ODD = Disruptive Behavior Disorder Rating Scale ODD scale; SDQ Int = Strengths and Difficulties Questionnaire internalizing scale; SDQ Ext = Strengths and Difficulties Questionnaire externalizing scale; RCADS-A = Revised Children's Anxiety and Depression Scale anxiety scale; RCADS-D = Revised Children's Anxiety and Depression Scale depression scale; *SD* = standard deviation.

* $p < .001$

Figure 2 depicts the SEM model examining the relations between the two-factor irritability rating scale and the dimensional measures of psychological problems (DBD-ODD, SDQ-Int, SDQ-Ext, RCADS-P-25-Anx, and RCADS-P-25-Dep). Child gender was entered as a covariate (tonic and phasic factors regressed on gender; DBD, SDQ, RCADS-P-25 regressed on gender) to account for scalar/threshold non-invariance. The model provided good fit to the data, χ^2 = 1108.53, *df* = 328, $p < .001$, CFI = .978, RMSEA = .077 (90% CI [.072, .082]), SRMR = .035.

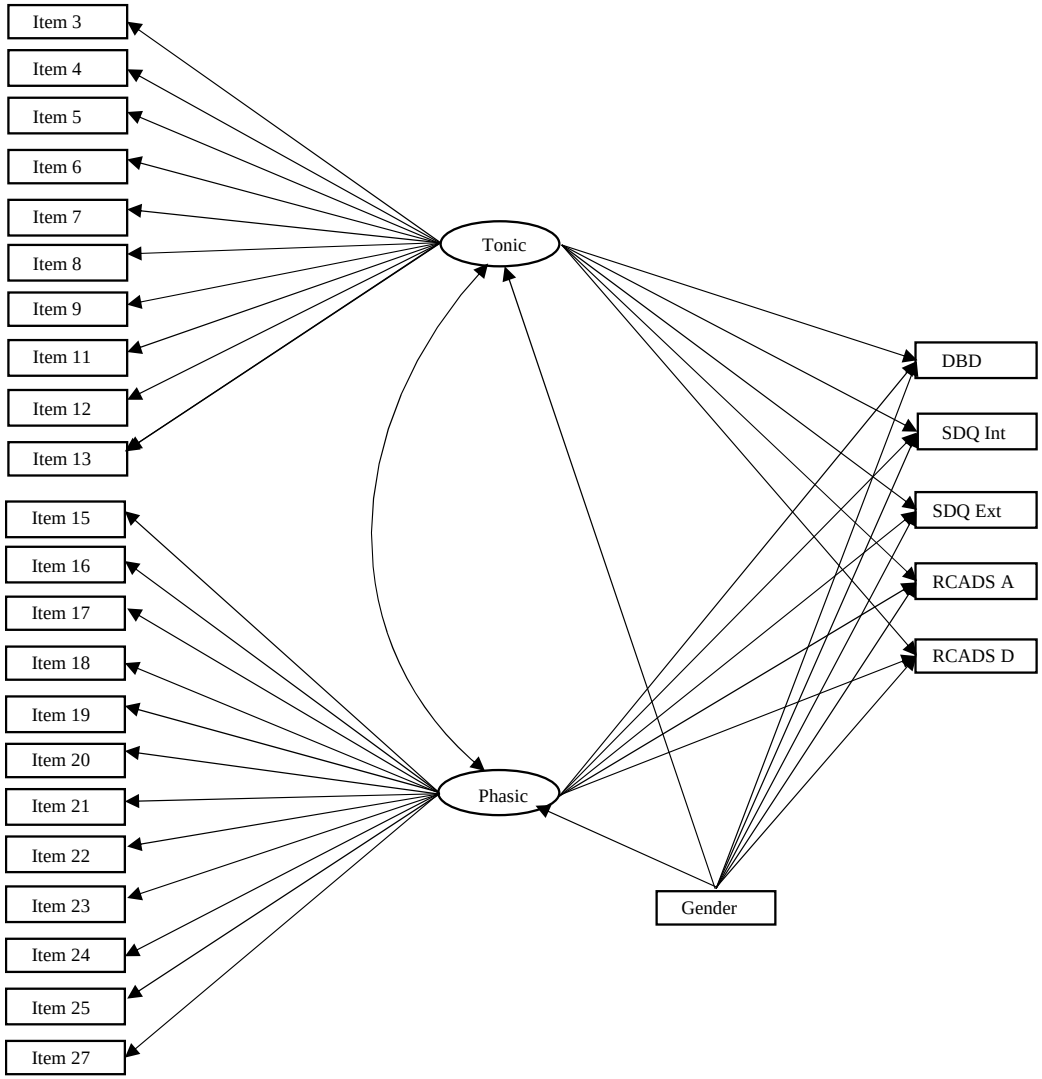

Figure 2. Structural equation model examining relations between Irritability Rating Scale factors and DBD ODD, SDQ Internalizing, SDQ Externalizing, RCADS Anxiety, and RCADS Depression.

For evidence of convergent validity, SDQ externalizing problems were significantly, positively related to the tonic (β = .35, p < .001) and phasic (β = .36, p < .001) irritability factors (applying a Bonferroni-corrected p value of .003; .05/17). The phasic factor was significantly, positively related to ODD symptoms (β = .51, p < .001) whereas the tonic irritability factor (applying the p value correction) was not significantly related to ODD symptoms (β = .25, p = .01). The phasic factor showed nonsignificant relations with SDQ internalizing problems (β = -.13, p = .26), anxiety (β = .05, p = .64), and depression (β = .12, p = .17). Finally, the tonic factor was significantly, positively related to SDQ internalizing problems (β = .71, p < .001), anxiety (β = .49, p < .001), and depression (β = .51, p < .001). Gender did not demonstrate significant relations with any of the variables in the model (applying the p value correction), including tonic irritability (β = .04, p = .50), phasic irritability (β = .04, p = .44), ODD (β = .06, p = .09), SDQ internalizing problems (β = -.03, p = .53), SDQ externalizing problems (β = .07, p = .08), anxiety (β = -.10, p = .03), and depression (β = -.03, p = .47).

Discrimination Analyses

Chi-square tests confirmed that parents of children in the low- (< 1 *SD* above mean), moderate- (≥ 1 *SD* above mean), and high-risk irritability groups (≥ 2 *SD* above mean) did not vary on age or gender (all *p*s > .05) across the three scales (total irritability, tonic irritability, phasic irritability). Bonferroni-corrected p value adjustments were applied to account for the multiple ANOVAs being conducted for the dependent variables (ODD, internalizing, externalizing, anxiety, depression; .05/5 = .01). Scores on dimensional psychological problems were significantly different based on a participant's

risk for total irritability problems, $F(10, 660) = 13.47$, $p < .001$; Wilk's $\Lambda = .690$, partial $\eta^2 = .170$. Bonferroni-corrected *p* value adjustments were then applied to account for multiple post-hoc comparisons (.01/3 = .003). Scores are presented by group in Table 5. For scores on ODD symptoms, SDQ internalizing problems, SDQ externalizing problems, anxiety, and depression, there were significant pairwise differences between the low and moderate-risk groups (all *p*s < .001), such that participants in the moderate-risk group had greater psychological problems. The same pattern of results was observed between the low and high-risk groups (all *p*s < .003), with participants in the high-risk group showing more psychological problems, with the difference being that SDQ internalizing problems was not significantly different for the low and high-risk groups (*p* > .003). The moderate- and high-risk groups were not significantly different on any of these measures, all *p*s > .05.

Table 5

Psychological Problems Compared by Irritability Rating Scale Score Risk Status

	Total Score		
Domain	Low Risk (*n* = 286)	Moderate Risk (*n* = 43)	High Risk (*n* = 8)
DBD ODD	3.90	8.91[a]	10.63 [a]
SDQ Internalizing	5.05	9.21[a]	8.88
SDQ Externalizing	5.63	9.90[a]	12.20[a]
RCADS Anxiety	6.81	13.00[a]	13.89[a]
RCADS Depression	4.19	9.40[a]	11.75[a]
	Tonic Score		
	Low Risk (*n* = 299)	Moderate Risk (*n* = 40)	High Risk (*n* = 7)
DBD ODD	4.02	9.05[a]	9.29[a]
SDQ Internalizing	5.00	9.45[a]	9.57
SDQ Externalizing	5.69	10.09[a]	10.61[a]
RCADS Anxiety	7.00	12.42[a]	13.88
RCADS Depression	4.26	9.80[a]	11.14[a]
	Phasic Score		
	Low Risk (*n* = 292)	Moderate Risk (*n* = 42)	High Risk (*n* = 9)
DBD ODD	3.89	9.38[a]	9.89[a]
SDQ Internalizing	4.99	9.48[a]	8.67
SDQ Externalizing	5.65	10.01[a]	11.54[a]
RCADS Anxiety	6.77	12.96[a]	14.09[a]
RCADS Depression	4.08	10.12[a]	10.89[a]

Note. DBD ODD = Disruptive Behavior Disorder Rating Scale ODD scale; SDQ = Strengths and Difficulties Questionnaire; RCADS = Revised Children's Anxiety and Depression Scale.

[a]Comparison with low risk group significant at $p < .003$.

Scores on dimensional psychological problems were also significantly different based on a participant's risk for tonic irritability problems, $F(10, 678) = 10.73$, $p < .001$; Wilk's $\Lambda = .745$, partial $\eta^2 = .137$. Bonferroni-corrected *p* value adjustments were applied to account for multiple post-hoc comparisons (.01/3 = .003). Scores are presented by group in Table 5. For scores on ODD symptoms, SDQ internalizing problems, SDQ externalizing problems, anxiety, and depression, there were significant pairwise differences between the low and moderate-risk groups (all *p*s < .001), such that participants in the moderate-risk group had more psychological problems. A similar pattern of results was observed between the low and high-risk groups (all *p*s < .003), with the difference being that SDQ internalizing problems and anxiety were not significantly different for the low and high-risk groups (all *p*s > .003). The moderate- and high-risk groups were not significantly different on any of these measures, all *p*s > .05.

Finally, scores on dimensional psychological problems were significantly different based on a participant's risk for phasic irritability problems, $F(10, 672) = 15.09$, $p < .001$; Wilk's $\Lambda = .667$, partial $\eta^2 = .183$. Bonferroni-corrected *p* value adjustments were applied to account for multiple post-hoc comparisons (.01/3 = .003). Scores are presented by group in Table 5. For scores on ODD symptoms, SDQ internalizing problems, SDQ externalizing problems, anxiety, and depression, there were significant pairwise differences between the low and moderate-risk groups (all *p*s < .001), such that participants in the moderate-risk group had greater psychological problems. The same pattern of results was observed between the low and high-risk groups (all *p*s < .003), with the difference being that SDQ internalizing problems was not significantly different for

the low and high-risk groups ($p > .003$). The moderate- and high-risk groups were not significantly different on any of these measures, all ps > .05.

Conditional Probabilities

Table 6 presents base rates of each item and the sensitivity, specificity, Positive Predictive Power (PPP) and Negative Predictive Power (NPP). The mean base rate among items was .19 (range = .08 - .38) in the sample, meaning roughly 1 in 5 children exhibited symptoms of irritability, which falls within previous estimates in community and clinical samples (Copeland et al., 2015; Moore et al., 2019). Sensitivity (i.e., proportion of people with a condition who are correctly identified by a screening test as indeed having that condition; "true positives"; Trevethan, 2017) was lower than desired (mean = .37; range = .17 - .68), meaning the majority of items on the scale may not be adequate at identifying the presence of significant irritability in youth who experience problematic irritability and impairment. This finding would direct caution when using it as a one-time assessment (i.e., screening instrument) for irritability, for example. Specificity (i.e., proportion of people without a condition who are correctly identified by a screening test as indeed not having the condition; "true negatives"; Trevethan, 2017) values were high (mean = .94; range = .85 - .98), meaning the majority of items on the scale are effective at identifying the absence of significant irritability in youth who do not experience problematic irritability and impairment. The PPP (i.e., probability that people with a positive screening test result indeed do have the condition of interest; "avoiding false positives"; Trevethan, 2017) was high (mean = .85; range = .78 - .91), indicating that the scale would be effective at not misidentifying those without these problems as

having problems. Finally, the NPP (i.e., probability that people with a negative screening test result indeed do not have the condition of interest; “avoiding false negatives”; Trevethan, 2017) was moderate (mean = .67; range = .61 - .78), indicating that the scale may be effective at classifying children with irritability problems who have problems. At the item level, item 4 (“My child is easily agitated”), item 6 (“When things don’t go my child’s way s/he gets upset easily”), and item 13 (“My child feels frustrated”) demonstrated the most desirable benchmarks with moderate sensitivity and high PPP, and high specificity and moderate NPP. High scores on these items may be most effective at identifying youth with problematic irritability among those whose mothers endorse significant behavior problems and impairment, as opposed to those whose mothers do not endorse significant behavior problems and impairment. These items were all from the tonic subscale, suggesting that phasic items may be overall less effective at discriminating between clinical and nonclinical groups.

Table 6

Conditional Probabilities for Irritability Rating Scale Items (N = 22)

Item	Base Rate	Sensitivity	Specificity	PPP	NPP
Item 3	.19	.38	.95	.86	.67
Item 4	.26	.51	.93	.85	.72
Item 5	.26	.50	.91	.82	.71
Item 6	.38	.68	.85	.78	.78
Item 7	.26	.47	.90	.79	.69
Item 8	.21	.38	.92	.80	.67
Item 9	.17	.32	.95	.84	.65
Item 11	.18	.35	.95	.85	.66
Item 12	.18	.38	.97	.91	.67
Item 13	.27	.56	.95	.90	.74
Item 15	.12	.23	.95	.82	.62
Item 16	.16	.33	.96	.88	.65
Item 17	.16	.31	.94	.82	.64
Item 18	.17	.35	.96	.90	.66
Item 19	.17	.32	.94	.83	.65
Item 20	.10	.20	.97	.90	.62
Item 21	.18	.36	.95	.89	.67
Item 22	.11	.24	.97	.91	.63
Item 23	.20	.38	.94	.85	.67
Item 24	.21	.38	.91	.78	.66
Item 25	.08	.17	.98	.85	.61
Item 27	.12	.25	.97	.89	.63
Average	.19	.37	.94	.85	.67

Note. Items endorsed as 4 (“often”) or 5 (“always”) were counted as a symptom being present. Base Rate = *N* with symptom/Total *N*; Sensitivity = *N* with behavior problems who show symptom/*N* with behavior problems; Specificity = *N* without behavior problems who do not show symptom/*N* without behavior problems; PPP = Positive Predictive Power, *N* with behavior problems who show symptom/*N* with symptom; NPP = Negative Predictive Power, *N* without behavior problems who do not show symptom/*N* without symptom.

Discussion

The current study developed and examined a parent-report, dimensional measure of youth irritability based on theoretical and empirical evidence of two related, but distinguishable aspects of irritability (tonic and phasic). To my knowledge, this is the first measure that was psychometrically developed to measure tonic and phasic irritability in children or adults (Beauchaine & Tackett, 2020). Irritability has been inadequately defined, both as a categorical and dimensional construct, prompting repeated calls to improve its definition and measurement in youth psychopathology. The tonic/phasic conceptualization has been advanced recently (Copeland et al., 2015) without identifiable progress when it comes to novel, original assessment tools. As contemporary knowledge about the underlying mechanisms (e.g., neurobiological systems) of irritability is advanced, instruments are needed that can keep pace with measuring irritability as a transdiagnostic risk factor for diverse forms of psychological problems.

Results of EFA and reliability analyses supported a 22-item scale with distinct, but strongly related tonic (10 items) and phasic (12 items) subscales. Notably, the tonic and phasic subscales comprised items that, a priori, were considered representative of tonic and phasic constructs, respectively. Moreover, total scale and subscale scores had excellent reliability. This extends findings from previous irritability measures that were designed to measure broad irritability (e.g., Stringaris et al., 2012), but importantly, it illustrates that separate irritability dimensions can be empirically-derived with corresponding, highly-reliable indices within a parent-report rating scale format.

Furthermore, these empirical findings align with the tonic/phasic theoretical model that has been advanced as the contemporary approach to measuring youth irritability.

Factor analyses supported a tonic/phasic structure in children across ages 6 to 12. Given that irritability (broadly) occurs at different rates during childhood (Brotman et al., 2017; Leibenluft & Stoddard, 2013), it is an important contribution to appropriately capture tonic and phasic irritability across a wide age range. With this finding, future research could be undertaken to assess whether this measure, when administered over multiple time points, may be useful to distinguish youth on one of the problematic trajectories of irritability (increasing, persistently-high) shown to predict deleterious outcomes in adolescence (Orri et al., 2019; Riglin et al., 2019). Since support for measurement invariance means it is appropriate to apply this measure in a longitudinal context, at least for children aged 6 to 12 years, findings would be able to shed light on which children—and at what age in this range—experience membership in specific groups, and whether group membership is stable or dynamic. Such findings would be informative when identifying critical windows during which school-age youth with elevated tonic and/or phasic irritability are candidates for monitoring and/or intervention. In addition, future studies can examine whether the factor structure holds when applying to younger (ages 3-5) or older (ages 13 and up) children who experience higher and lower rates of problematic irritability, respectively.

Although the tonic/phasic factor structure was also supported in parent reports of irritability in their male and female children, it appears that parents of these children show different patterns of responding to individual items. Previous research (Copeland et

al., 2015; Moore et al., 2019) has shown that prevalence rates of tonic/phasic irritability tend to be similar in boys and girls. However, the same studies underscored the possibility that children who demonstrate similar levels of the underlying construct may differ in their expression of each domain (particularly when comparing by gender). For example, the phasic irritability item "When my child is upset, s/he loses control over his/her behaviors" asks raters to reflect on behavioral responses that represent a significant departure from the child's typical behavior that shows poor control. In male children, these behaviors may include physical aggression and destruction of property, whereas in female children, these behaviors may include screaming/crying and verbal aggression (for meta-analysis, see Archer, 2004; see also Lansford et al., 2012). Thus, parents might endorse different levels for items used as indicators of the same construct depending on the gender-specific expression of the behavior by their child. Evidence from previous studies in related types of child psychopathology (ODD) demonstrate consistent differences among male and female children in the manner in which externalizing behaviors are expressed (for meta-analysis, see Demmer, Hooley, Sheen, McGillivray, & Lum, 2017), though overall rates of ODD are higher in males than females. It is unclear if similar such differences may exist when capturing tonic and phasic irritability, so it as an empirical question whether overall levels of tonic/phasic irritability are the same, but symptom manifestation is different, between males and females. Future studies can test this by sampling equal numbers of boys and girls and testing associations between the irritability scale and specific externalizing behaviors (e.g., noncompliance, aggression) to explore gender differences.

Findings from SEM analyses were consistent with hypotheses that the irritability scale would be related to a variety of childhood psychological problems. This study extended the literature by examining whether tonic and phasic irritability differentially predicted psychological problems. Results of these analyses, upon close examination, do not appear to bear this out. For example, children with higher scores on phasic irritability scored higher on measures of ODD symptoms, whereas children with higher scores on tonic irritability did not show this relationship (after accounting for multiple statistical tests among the dependent variables in the SEM model). However, bivariate analyses (Pearson's correlations) were nearly identical between ODD symptoms and tonic (r = .71) and phasic (r = .74) irritability. This suggests that the results of SEM analyses could be attributed to suppression effects that occur in multiple regression analyses when one predictor (but not the other) shows a significant relationship with an outcome variable due to both predictors being highly correlated with each other (for discussion, see Smith, Ager, & Williams, 1992). Other significant findings from the SEM model when examining tonic/phasic irritability and associations with childhood psychological problems also demonstrate this presumed suppression effect. Thus, the absence of a measure with the ability to differentially predict outcomes based on tonic vs. phasic irritability remains an issue to be addressed in the empirical literature. As such, prospective studies exploring youth irritability are still needed to examine whether tonic/phasic irritability can evince discriminant validity. If such evidence is found, then longitudinal studies may also help clarify the role that different trajectories of tonic and phasic irritability play in predicting child psychological problems, which could assist

with screening for specific disorders (e.g., anxiety/depression vs. ADHD/ODD) to provide more accurate identification and appropriate treatment referrals when children present with irritability problems.

Discrimination analyses suggested parent-rated moderate and high irritability (using the total scale and subscale scores) can be used to discriminate among children who score higher on dimensional measures of psychological problems than youth with low irritability. As above, any differences between tonic and phasic irritability subscale scores may have been more pronounced due to the high correlation between these scores, though the discussion below focuses on converging findings with total scale scores. For all types of psychological problems (ODD, internalizing, externalizing, anxiety, depression), children in the moderate risk group showed higher rates than children in the low risk group. Most types of psychological problems were also significantly higher for those in the high risk group, although this was not observed for internalizing problems and anxiety. This indicates that, for youth whose scores are even 1 *SD* above the scale mean, they exhibit higher levels of psychological problems and are likely to be at higher risk to experience significant impairment in functioning. Surprisingly, there were no observed differences between the moderate (≥1 *SD*) and high (≥2 *SD*) groups, though low membership in the latter group likely impacted power to detect some significant differences. These results should therefore be interpreted with caution, as it is unclear whether a sample with a larger number of cases in the high risk group would detect significant differences. Results from the current study recommend setting a cutoff of 1 *SD* above the scale mean (moderate irritability) may be sufficient at identifying children

who are likely to have more significant psychological problems. This cutoff should be replicated within a larger sample to clarify its appropriateness, and future studies may wish to examine whether lower cutoffs (e.g., .5 *SD* above mean = moderate, 1.5 *SD* above mean = high) yield similar results. However, the lack of significant differences between the moderate and high groups indicates that using one cutoff (rather than two) may be sufficient to differentiate concerning and unremarkable levels of irritability, unless future research suggests otherwise.

Finally, findings from conditional probability analyses provided mixed support for the utility of the dimensional irritability rating scale in a clinical context. Base rates of individual items indicated that approximately 1 in 5 children exhibited symptoms of elevated irritability, which is similar to prevalence rates found in prior studies (Copeland et al., 2015; Moore et al., 2019). Replicating the prevalence of clinically elevated irritability is an important contribution given the novelty of the scale. These prior studies were based on flawed measurement tools (i.e., not developed to measure tonic vs. phasic irritability), so prevalence rates found using the current scale can also be viewed as additional evidence of the rate at which problematic youth irritability occurs. Specificity and PPP were desirable, though sensitivity and NPP were more modest. Sensitivity (i.e., identifying true positives, avoiding false negatives) merits additional consideration when examining a transdiagnostic construct like youth irritability. Given that irritability is not specific to any one childhood psychological disorder, but rather a feature of many, a lower sensitivity value could reflect that youth with elevated irritability will not be identified as frequently when using a reference standard (in this study, behavior problems

plus impairment) that is also nonspecific with regard to diagnosis to meet a positive result on a screening test. A high rate of true positives is not expected to begin with, so a high sensitivity value would be surprising. It will be important to empirically test this assertion with this instrument, as well as future instruments, to determine if there are circumstances in which high sensitivity on an irritability instrument would be desirable (e.g., when trialing a medication with serious side effects in order to treat severe irritability).

These conditional probability findings direct caution when using the scale as a one-time assessment (i.e., screening instrument) for irritability given that it may be effective at classifying children with clinically-significant irritability problems (especially those with higher scores), though some children (especially those with lower scores) could be missed and go on to meet cutoffs for significant problems at a later time point. Clinically, this would hamper attempts at preventing escalation of significant irritability and associated impairment, as well as potentially interfere with youth connecting with services that might address associated psychological problems (e.g., ODD, anxiety, depression). At the descriptive level, there was evidence that three tonic scale items may provide more optimal screening properties (moderate sensitivity and high PPP, high specificity and moderate NPP), though desired values for these indices depend on the purpose of the scale as a screening instrument (Trevethan, 2017). Given the status of the literature and that this is the first known attempt to develop a tonic/phasic irritability rating scale, it is important to consider ways that the sensitivity/PPP and specificity/NPP can be enhanced with the goal of utilizing this measure beyond research contexts (i.e., in clinical practice). Future studies can explore whether modifications to the irritability scale

(e.g., raising or lowering cutpoints; Trevethan, 2017) improve these characteristics to enhance confidence in its use for screening purposes.

The current study had a number of strengths, including a theory-driven approach to scale development, recruiting a community sample of mothers with balanced child age and gender, and using sophisticated analyses to explore initial psychometric properties in the first tonic/phasic irritability rating scale. However, several limitations are important to consider. First, the study recruited a community sample of mothers who were fairly homogenous in terms of demographic characteristics. Racial/ethnic disparities (in particular for African American and Hispanic youth) in externalizing disorders are well-documented (e.g., Fadus et al., 2019), including in prevalence rates that some contend are inflated by biased assessment practices. On the other hand, there is robust evidence to suggest that racial/ethnic minority youth may experience higher rates of internalizing disorders relative to their White peers (for review, see Anderson & Mayes, 2010). To prevent the potential for assessment tools to be inappropriately applied across racial/ethnic groups (either due to being over-representative or under-representative), samples with larger groups of racial/ethnic minority youth should probe whether ratings are invariant (e.g., among White, African American, and Hispanic youth).

Second, the design of this study, though appropriate for an initial investigation of a novel measure of youth irritability, limits the impact of the findings. Mothers provided ratings for the youth irritability measure and the outcomes used to assess predictive validity. The concordance between informants on measures of child behavior is variable, particularly between those who interact with children in different settings (e.g.,

Achenbach, McConaughy, & Howell, 1987). Thus, analyses that compare the two-factor structure of the irritability scale among different sets of raters (e.g., fathers, teachers) may allow for evidence in support of its equivalence across informants. In addition, future studies should not only include additional informants for irritability ratings, but also for outcomes of interest (e.g., father or teacher report of psychological problems, impairment). Examining associations with objective outcomes (e.g., school suspensions) would allow for opportunities to increase confidence in the clinical relevance of this measure. Finally, all data were collected at a single point in time, which limits the ability to make temporal predictions or examine stability of ratings. Studies that use prospective, longitudinal designs would extend the psychometric support of this scale by administering it, as well as connections with covariates (including psychological problems), across time.

Third, contributions to the literature (Roy & Comer, 2020) were made after the conclusion of data collection, which may have resulted in relevant items not being considered for the item pool that would have contributed incremental strength to a tonic/phasic irritability rating scale. For example, the Clinician Affective Reactivity Index (CL-ARI; Haller et al., 2020) uses 12 items that are rated by a clinician to assess temper outbursts, mood between outbursts, and impairment; although these items were developed to be administered in a semi-structured interview to parent and child, they could offer alternative wording for items to evaluate tonic/phasic irritability.[5] Regardless,

[5] It is important to note that these authors have not made their measure publicly available, so it is unclear whether this point is true (due to the inability to examine items for correspondence with tonic/phasic irritability).

an updated literature search (conducted August, 2020) did not reveal any additional irritability instruments that purport to measure tonic/phasic domains, underscoring the novelty of the scale developed in this study.

Fourth, there are additional analytical approaches, such as item response theory (IRT; Foster, Min, & Zickar, 2017), that could contribute important information about item discrimination that were not performed as part of the current study. IRT analyses that estimate item characteristic curves (ICC) can be used to determine which items have optimal ability, based on the probability of individuals with a specific level of the latent attribute (e.g., irritability) endorsing a response category, to better discriminate among those with or without the underlying attribute (i.e., clinically-significant irritability). This would allow examination of whether the criteria used to classify problematic irritability (scores of 4 or 5 at the item level) is appropriate, and it would also clarify whether specific items benefit from a lower (e.g., score of 3, 4, or 5) or higher (e.g., score of 5 only) category threshold for clinical significance. Future research with the current measure, or with other novel measures of tonic/phasic irritability, may wish to use IRT to provide additional evidence of psychometric properties.

Fifth, data for the current study was collected via respondents on the internet, which is susceptible to threats to validity such as invalid responding. Although evidence-based steps to minimize these threats were taken (i.e., setting high threshold for MTurk worker completion rate, embedding attention check questions in survey), future research that collects responses through alternative designs would increase confidence in the findings from the current study. This might include having parents complete ratings in a

research laboratory setting, complete ratings as part of a clinical visit (e.g., integrated care setting), or embed the irritability scale as an outcome measure within a larger study (e.g., population-based study). In addition, exploring irritability ratings in clinical samples is necessary to extend its utility; studies that recruit clinically-referred youth may also wish to examine whether youth with specific diagnoses (e.g., anxiety disorders, ODD) demonstrate the pattern of differences between tonic/phasic irritability that were observed in the current study. Finally, the analyses that compared youth in different risk categories (low, moderate, and high) were likely underpowered to detect significant differences among those in the high risk group. Studies that specifically recruit large samples of youth across a broader spectrum of irritability, even perhaps oversampling those at highest risk, would address the inconclusive findings observed between the moderate and high risk groups in this study.

In closing, the development of a novel tonic/phasic irritability rating scale has important implications for research and practice. Measures of youth irritability that can be used to clarify the prevalence and longitudinal trajectories of tonic and phasic irritability across childhood may optimize classification of children who are at-risk for specific types of psychological problems, potentially aiding with identification of those most in need of preventative services. If the current scale can be revised to enhance its application as a screening instrument, or if it is identified to function best for more severe populations, then it could fill the need to have psychometrically-sound measures of youth irritability as a dimensional construct. Additionally, measures of tonic and phasic irritability in children may also be useful to enhancing psychosocial interventions for

children with clinically significant internalizing and externalizing problems. For example, preliminary studies implicate severe irritability as an important predictor of the response of children to behavioral parent training and cognitive behavioral treatment delivered in a modular approach (e.g., Evans et al., 2020). Measures of tonic and phasic irritability would be instrumental in assessing the likelihood of these constructs being of concern for clinicians working with these children, and thus in need of specific clinical focus during therapy. Overall, the development of a novel dimensional irritability rating scale will assist with correcting the pervasive gap in the literature that has limited advances in the measurement of youth irritability, which in turn may spur better alignment among theoretical perspectives, empirical examinations, and, ultimately, clinical tools used for the assessment and treatment of irritability.

References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*, 213–232. doi:10.1037/0033-2909.101.2.213

Achenbach, T. M., & Rescorla, L. A. (2007). *Manual for the ASEBA school-age forms & profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Althoff, R. R., Crehan, E. T., He, J., Burstein, M., Hudziak, J. J., & Merikangas, K. R. (2016). Disruptive mood dysregulation disorder at ages 13-18: Results from the National Comorbidity Survey—Adolescent Supplement. *Journal of Child and Adolescent Psychopharmacology*, *26*, 107-113. doi:10.1089/cap.2015.0038

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

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

Anderson, E. R., & Mayes, L. C. (2010). Race/ethnicity and internalizing disorders in youth: A review. *Clinical Psychology Review*, *30*, 338–348. doi:10.1016/j.cpr.2009.12.008

Angold, A., & Costello, E. J. (2000). The Child and Adolescent Psychiatric Assessment (CAPA). *Journal of the American Academy of Child & Adolescent Psychiatry, 39*, 39-48. doi:10.1097/00004583-200001000-00015

Archer, J. (2004). Sex differences in aggression in real-world settings: A meta-analytic review. *Review of General Psychology, 8,* 291–322. doi:10.1037/1089-2680.8.4.291

Avenevoli, S., Blader, J. C., & Leibenluft, E. (2015). Irritability in youth: An update. *Journal of the American Academy of Child & Adolescent Psychiatry, 54,* 881-883. doi:10.1016/j.jaac.2015.08.012

Axelson, D. A., Birmaher, B., Findling, R. L., Fristad, M. A., Kowatch, R. A., Youngstrom, E. A.,…Diler, R. S. (2011). Concerns regarding the inclusion of temper dysregulation disorder with dysphoria in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. *Journal of Clinical Psychiatry, 72,* 1257-1262. doi:10.4088/JCP.10com06220

Beauchaine, T. P., & Tackett, J. L. (2020). Irritability as a transdiagnostic vulnerability trait: Current issues and future directions. *Behavior Therapy, 51*, 350–364. doi:10.1016/j.beth.2019.10.009

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6,* 149. doi:10.3389/fpubh.2018.00149

Brotman, M. A., Kircanski, K., Stringaris, A., Pine, D. S., & Leibenluft, E. (2017). Irritability in youths: A translational model. *American Journal of Psychiatry, 174,* 520-532. doi:10.1176/appi.ajp.2016.16070839

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Bunford, N., Dawson, A. E., Evans, S. W., Ray, A. R., Langberg, J. M., Owens, J. S.,…Allan, D. M. (2018). The Difficulties in Emotion Regulation Scale–Parent Report: A psychometric investigation examining adolescents with and without ADHD. *Assessment.* Advance online publication. doi:10.1177/1073191118792307

Bunford, N., Evans, S. W., & Wymbs, F. (2015). ADHD and emotion dysregulation among children and adolescents. *Clinical Child and Family Psychology Review, 18,* 185–217. doi:10.1007/s10567-015-0187-5

Carlson, G. A., Danzig, A. P., Dougherty, L. R., Bufferd, S. J., & Klein, D. N. (2016). Loss of temper and irritability: The relationship to tantrums in a community and clinical sample. *Journal of Child and Adolescent Psychopharmacology, 26,* 114-122. doi:10.1089/cap.2015.0072

Carlson, G. A., & Klein, D. N. (2018). Commentary: Frying pan to fire? Commentary on Stringaris et al. (2018). *Journal of Child Psychology and Psychiatry, 59,* 740-743. doi:10.1111/jcpp.12873

Carver, C. S., & Harmon-Jones, E. (2009). Anger is an approach-related affect: Evidence and implications. *Psychological Bulletin, 135,* 183-204. doi:10.1037/a0013965

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464–504. doi:10.1080/10705510701301834.

Cole, P. M., Michel, M. K., & Teti, L. O. (1994). The development of emotion regulation and dysregulation: A clinical perspective. *Monographs of the Society for Research in Child Development, 59*, 73-102. doi:10.1111/j.1540-5834.1994.tb01278.x

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology, 56*, 754-761. doi:10.1037/0022-006X.56.5.754

Copeland, W. E., Angold, A., Costello, E. J., & Egger, H. (2013). Prevalence, comorbidity, and correlates of DSM-5 proposed disruptive mood dysregulation disorder. *American Journal of Psychiatry, 170*, 173-179. doi:10.1176/appi.ajp.2012.12010132

Copeland, W. E., Brotman, M. A., & Costello, E. J. (2015). Normative irritability in youth: Developmental findings from the Great Smoky Mountains Study. *Journal of the American Academy of Child and Adolescent Psychiatry, 54*, 635–642. doi:10.1016/j.jaac.2015.05.008

Copeland, W. E., Shanahan, L., Egger, H., Angold, A., & Costello, E. J. (2014). Adult diagnostic and functional outcomes of DSM-5 disruptive mood dysregulation disorder. *American Journal of Psychiatry, 171*, 668-674. doi:10.1176/appi.ajp.2014.13091213

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation*, *10*(7). Retrieved from: http://pareonline.net/getvn.asp?v=10&n=7

Craig, K. J., Hietanen, H., Markova, I. S., & Berrios, G. E. (2008). The Irritability Questionnaire: A new scale for the measurement of irritability. *Psychiatry Research*, *159*, 367-375. doi:10.1016/j.psychres.2007.03.002

Dawson, A. E. (2013). *Conflict in the parent-teacher relationship: Initial scale development* (Unpublished master's thesis). Ohio University, Athens, OH.

del Barrio, V., Aluja, A., & Spielberger, C. (2004). Anger assessment with the STAXI-CA: psychometric properties of a new instrument for children and adolescents. *Personality and Individual Differences*, *37*, 227-244. doi:10.1016/j.paid.2003.08.014

Demmer, D. H., Hooley, M., Sheen, J., McGillivray, J. A., & Lum, J. A. G. (2017). Sex differences in the prevalence of oppositional defiant disorder during middle childhood: A meta-analysis. *Journal of Abnormal Child Psychology*, *45*, 313–325. doi:10.1007/s10802-016-0170-8

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd. ed). Los Angeles, CA: SAGE Publications, Inc.

Dougherty, L. R., Smith, V. C., Bufferd, S. J., Carlson, G. A., Stringaris, A., Leibenluft, E., & Klein, D. N. (2014). DSM-5 disruptive mood dysregulation disorder:

Correlates and predictors in young children. *Psychological Medicine, 44*, 2339-2350. doi:10.1017/S0033291713003115

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*, 399–412. doi:10.1111/bjop.12046

Dworkin, J., Hessel, H., Gliske, K., & Rudi, J. H. (2016). A comparison of three online recruitment strategies for engaging parents. *Family Relations: An Interdisciplinary Journal of Applied Family Studies*, *65*, 550-561. doi:10.1111/fare.12206

Ebesutani, C., Korathu-Larson, P., Nakamura, B. J., Higa-McMillan, C., & Chorpita, B. (2017). The Revised Child Anxiety and Depression Scale 25–Parent Version: Scale development and validation in a school-based and clinical sample. *Assessment, 24*, 712-728. doi:10.1177/1073191115627012

Evans, S. C., Burke, J. D., Roberts, M. C., Fite, P. J., Lochman, J. E., de la Peña, F. R., & Reed, G. M. (2017). Irritability in child and adolescent psychopathology: An integrative review for ICD-11. *Clinical Psychology Review*, *53*, 29–45. doi:10.1016/j.cpr.2017.01.004

Evans, S. C., Weisz, J. R., Carvalho, A. C., Garibaldi, P. M., Bearman, S. K., Chorpita, B. F., & Research Network on Youth Mental Health (2020). Effects of standard and modular psychotherapies in the treatment of youth with severe irritability. *Journal of Consulting and Clinical Psychology, 88*, 255–268. doi:10.1037/ccp0000456

Fadus, M. C., Ginsburg, K. R., Sobowale, K., Halliday-Boykins, C. A., Bryant, B. E., Gray, K. M., & Squeglia, L. M. (2019). Unconscious bias and the diagnosis of disruptive behavior disorders and ADHD in African American and Hispanic youth. *Academic Psychiatry, 44*, 95–102. doi:10.1007/s40596-019-01127-6

Fabiano, G. A., Pelham, W. E., Jr., Waschbusch, D. A., Gnagy, E. M., Lahey, B. B., Chronis, A.…Burrows-Maclean, L. (2006). A practical measure of impairment: Psychometric properties of the Impairment Rating Scale in samples of children with Attention Deficit Hyperactivity Disorder and two school-based samples. *Journal of Clinical Child and Adolescent Psychology, 35*(3), 369–385. doi:10.1207/s15374424jccp3503_3

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272-299. doi:10.1037/1082-989X.4.3.272

Fernandez, E., & Johnson, S. L. (2016). Anger in psychological disorders: Prevalence, presentation, etiology and prognostic implications. *Clinical Psychology Review, 46*, 124-135. doi:10.1016/j.cpr.2016.04.012

Flessner, C. A., Murphy, Y. E., Brennan, E., & D'Auria, A. (2017). The Parenting Anxious Kids Ratings Scale-Parent Report (PAKRS-PR): Initial scale development and psychometric properties. *Child Psychiatry and Human Development, 48*, 651-667. doi:10.1007/s10578-016-0688-6

Foster, G. C., Min, H., & Zickar, M. J. (2017). Review of Item Response Theory practices in organizational research. *Organizational Research Methods, 20*, 465-486. doi:10.1177/1094428116689708

Freeman, A. J., Youngstrom, E. A., Youngstrom, J. K., & Findling, R. L. (2016). Disruptive mood dysregulation disorder in a community mental health clinic: Prevalence, comorbidity and correlates. *Journal of Child and Adolescent Psychopharmacology, 26,* 123-130. doi:10.1089/cap.2015.0061

Giesbrecht, G. F., Miller, M. R., & Müller, U. (2010). The anger-distress model of temper tantrums: Associations with emotional reactivity and emotional competence. *Infant and Child Development, 19*, 478-497. doi:10.1002/icd.677

Goldstein, B. I., Birmaher, B., Carlson, G. A., DelBello, M. P., Findling, R. L., Fristad, M.,…Youngstrom, E. A. (2017). The International Society for Bipolar Disorders Task Force report on pediatric bipolar disorder: Knowledge to date and directions for future research. *Bipolar Disorders, 19,* 524-543. doi:10.1111/bdi.12556

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Child Psychology & Psychiatry, 38*, 581-586. doi:10.1111/j.1469-7610.1997.tb01545.x

Goodman, A., Lamping, D. L., & Ploubidis, G. B. (2010). When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the Strengths and Difficulties Questionnaire (SDQ): Data from British parents, teachers and children. *Journal of Abnormal Child Psychology, 38*, 1179-1191. doi:10.1007/s10802-010-9434-x

Hair, J. F., Anderson, R. E., Tatham, R. L. & Black, W. C. (1995). *Multivariate data analysis* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Haller, S. P., Kircanski, K., Stringaris, A., Clayton, M., Bui, H., Agorsor, C.,…Brotman, M. A. (2020). The Clinician Affective Reactivity Index: Validity and reliability of a clinician-rated assessment of irritability. *Behavior Therapy, 51*, 283–293. doi:10.1016/j.beth.2019.10.005

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*, 377-381. doi:10.1016/j.jbi.2008.08.010

Holtzman, S., O'Connor, B. P., Barata, P. C., & Stewart, D. E. (2014). The Brief Irritability Test (BITe). *Assessment, 22*, 101-115. doi:10.1177/1073191114533814

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

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K.,…Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *The American Journal of Psychiatry*, *167*, 748-751. doi:10.1176/appi.ajp.2010.09091379

Irwin, D., Stucky, B., Langer, M., Thissen, D., DeWitt, E., Lai, J.-S.,…DeWalt, D. (2012). PROMIS Pediatric Anger Scale: An item response theory analysis. *Quality of Life Research*, *21*, 697-706. doi:10.1007/s11136-011-9969-5

Judah, M. R., Saulnier, K. G., Hager, N. M., & Allan, N. P. (2020). A bifactor model of the Straightforward Attentional Control Scale. *Journal of Psychopathology & Behavioral Assessment, 42*, 127–136. doi:10.1007/s10862-019-09737-y

Kelley, K. (2019). MBESS: The MBESS R Package (Version 4.6.0) [Computer software and manual]. Retrieved from https://CRAN.R-project.org/package=MBESS

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

Kraatz Keiley, M., Bates, J. E., Dodge, K. A., & Pettit, G. S. (2000). A cross-domain growth analysis: Externalizing and internalizing behaviors during 8 years of childhood. *Journal of Abnormal Child Psychology, 28*, 161–179. doi:10.1023/A:1005122814723

Lansford, J. E., Skinner, A. T., Sorbring, E., Giunta, L. D., Deater, D. K., Dodge, K. A.,…Chang, L. (2012). Boys' and girls' relational and physical aggression in nine countries. *Aggressive Behavior, 38*, 298–308. doi:10.1002/ab.21433

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575. doi:10.1111/j.1744-6570.1975.tb01393.x

Leibenluft, E., Cohen, P., Gorrindo, T., Brook, J. S., & Pine, D. S. (2006). Chronic versus episodic irritability in youth: A community-based, longitudinal study of clinical

and diagnostic associations. *Journal of Child and Adolescent Psychopharmacology, 16*, 456-466. doi:10.1089/cap.2006.16.456

Leibenluft, E., & Stoddard, J. (2013). The developmental psychopathology of irritability. *Development and Psychopathology*, *25*, 1473-1487. doi:10.1017/s0954579413000722

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*, 382–385. doi:10.1097/00006199-198611000-00017

MacDermott, S. T., Gullone, E., Allen, J. S., King, N. J., & Tonge, B. (2010). The Emotion Regulation Index for Children and Adolescents (ERICA): A psychometric investigation. *Journal of Psychopathology and Behavioral Assessment*, *32*, 301–314. doi:10.1007/s10862-009-9154-0

Margulies, D. M., Weintraub, S., Basile, J., Grover, P. J., & Carlson, G. A. (2012). Will disruptive mood dysregulation disorder reduce false diagnosis of bipolar disorder in children?. *Bipolar Disorders, 14*, 488-496. doi:10.1111/j.1399-5618.2012.01029.x

Meyers, E., DeSerisy, M., & Roy, A. K. (2017). Disruptive Mood Dysregulation Disorder (DMDD): An RDoC perspective. *Journal of Affective Disorders*, *216*, 117-122. doi:10.1016/j.jad.2016.08.007

Milich, R., Widiger, T. A., & Landau, S. (1987). Differential diagnosis of attention deficit and conduct disorders using conditional probabilities. *Journal of Consulting and Clinical Psychology, 55*, 762-767. doi:10.1037/0022-006x.55.5.762

Moore, A. A., Lapato, D. M., Brotman, M. A., Leibenluft, E., Aggen, S. H., Hettema, J. M.,…Roberson-Nay, R. (2019). Heritability, stability, and prevalence of tonic and phasic irritability as indicators of disruptive mood dysregulation disorder. *Journal of Child Psychology and Psychiatry, 60*, 1032–1041. doi:10.1111/jcpp.13062

Muthén, L. K., & Muthén, B.O. (1998-2019). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Nock, M. K., Wedig, M. M., Holmberg, E. B., & Hooley, J. M. (2008). The emotion reactivity scale: development, evaluation, and relation to self-injurious thoughts and behaviors. *Behavior Therapy*, *39*, 107–116. doi:10.1016/j.beth.2007.05.005

Orri, M., Galera, C., Turecki, G., Boivin, M., Tremblay, R. E., Geoffroy, M. C., & Côté, S. M. (2019). Pathways of association between childhood irritability and adolescent suicidality. *Journal of the American Academy of Child and Adolescent Psychiatry*, *58*, 99–107. doi:10.1016/j.jaac.2018.06.034

Parens, E., Johnston, J., & Carlson, G. A. (2010). Pediatric mental health care dysfunction disorder?. *New England Journal of Medicine*, *362*, 1853-1855. doi:10.1056/NEJMp1003175

Parent, J., Forehand, R., Pomerantz, H., Peisch, V., & Seehuus, M. (2017). Father participation in child psychopathology research. *Journal of Abnormal Child Psychology*, *45*, 1259–1270. doi:10.1007/s10802-016-0254-5

Pelham, W. E., Gnagy, E. M., Greenslade, K. E., & Milich, R. (1992). Teacher ratings of DSM-III—R symptoms for the disruptive behavior disorders. *Journal of the*

American Academy of Child & Adolescent Psychiatry, *31*, 210-218. doi:10.1097/00004583-199203000-00006

Pendergast, L. L., von der Embse, N., Kilgus, S. P., & Eklund, K. R. (2017). Measurement equivalence: A non-technical primer on categorical multi-group confirmatory factor analysis in school psychology. *Journal of School Psychology, 60*, 65-82. doi:10.1016/j.jsp.2016.11.002

Pillow, D. R., Pelham, W. J., Hoza, B., Molina, B. G., & Stultz, C. H. (1998). Confirmatory factor analyses examining attention deficit hyperactivity disorder symptoms and other childhood disruptive behaviors. *Journal of Abnormal Child Psychology*, *26*, 293-309. doi:10.1023/A:1022658618368

Potegal, M., & Davidson, R. J. (2003). Temper tantrums in young children: 1. Behavioral composition. *Developmental and Behavioral Pediatrics, 24*, 140-147. doi:10.1097/00004703-200306000-00002

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

Riglin, L., Eyre, O., Thapar, A. K., Stringaris, A., Leibenluft, E., Pine, D. S.,…Thapar, A. (2019). Identifying novel types of irritability using a developmental genetic approach. *The American Journal of Psychiatry, 176*, 635–642. doi:10.1176/appi.ajp.2019.18101134

Roy, A. K., & Comer, J. S. (2020). Advances in the conceptualization, assessment, and treatment of pediatric irritability. *Behavior Therapy, 51*, 207–210. doi:10.1016/j.beth.2019.10.010

Schleider, J. L., & Weisz, J. R. (2015). Using Mechanical Turk to study family processes and youth mental health: A test of feasibility. *Journal of Child and Family Studies*, *24*, 3235–3246. doi:10.1007/s10826-015-0126-6

Silvia, E. S. M., & MacCallum, R. C. (1988). Some factors affecting the success of specification searches in covariance structure modeling. *Multivariate Behavioral Research, 23*, 297-326. doi:10.1207/s15327906mbr2303_2

Simonds, J. (2006). *The role of reward sensitivity and response: Execution in childhood extraversion* (Unpublished doctoral dissertation). University of Oregon, Eugene, OR.

Smith, R. L., Ager, J. W., & Williams, D. L. (1992). Suppressor variables in multiple regression/correlation. *Educational & Psychological Measurement, 52*, 17-29. doi:10.1177/001316449205200102

Snaith, R. P., Constantopoulos, A. A., Jardine, M. Y., & McGuffin, P. (1978). A clinical scale for the self-assessment of irritability. *The British Journal of Psychiatry: The Journal of Mental Science, 132*, 164–171. doi:10.1192/bjp.132.2.164

Stringaris, A., Goodman, R., Ferdinando, S., Razdan, V., Muhrer, E., Leibenluft, E., & Brotman, M. A. (2012). The Affective Reactivity Index: A concise irritability scale for clinical and research settings. *Journal of Child Psychology and Psychiatry, 53*, 1109-1117. doi:10.1111/j.1469-7610.2012.02561.x

Stringaris, A., Vidal-Ribas, P., Brotman, M. A., & Leibenluft, E. (2018). Practitioner review: Definition, recognition, and treatment challenges of irritability in young

people. *Journal of Child Psychology and Psychiatry, 59,* 721-739. doi:10.1111/jcpp.12823

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson/Allyn & Bacon.

Tinsley, H. E., & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology, 34,* 414-424. doi:10.1037/0022-0167.34.4.414

Tiwari, R., Agarwal, V., Arya, A., Gupta, P. K., & Mahour, P. (2016). An exploratory clinical study of disruptive mood dysregulation disorder in children and adolescents from India. *Asian Journal of Psychiatry, 21,* 37-40. doi:10.1016/j.ajp.2016.02.005

Toohey, M. J., & DiGiuseppe, R. (2017). Defining and measuring irritability: Construct clarification and differentiation. *Clinical Psychology Review, 53,* 93–108. doi:10.1016/j.cpr.2017.01.009

Trevethan, R. (2017). Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. *Frontiers in Public Health, 5,* 307. doi:10.3389/fpubh.2017.00307

Tseng, W.-L., Moroney, E., Machlin, L., Roberson-Nay, R., Hettema, J. M., Carney, D.,…Brotman, M. A. (2017). Test-retest reliability and validity of a frustration paradigm and irritability measures. *Journal of Affective Disorders, 212,* 38-45. doi:10.1016/j.jad.2017.01.024

Vidal-Ribas, P., Brotman, M. A., Valdivieso, I., Leibenluft, E., & Stringaris, A. (2016). The status of irritability in psychiatry: A conceptual and quantitative review. *Journal of the American Academy of Child and Adolescent Psychiatry, 55*, 556–570. doi:10.1016/j.jaac.2016.04.014

Wakschlag, L. S., Briggs-Gowan, M. J., Choi, S. W., Nichols, S. R., Kestler, J., Burns, J. L.,…Henry, D. (2014). Advancing a multidimensional, developmental spectrum approach to preschool disruptive behavior. *Journal of the American Academy of Child & Adolescent Psychiatry, 53*, 82-96. doi:10.1016/j.jaac.2013.10.011

Wakschlag, L. S., Choi, S. W., Carter, A. S., Hullsiek, H., Burns, J., McCarthy, K.,…Briggs-Gowan, M. J. (2012). Defining the developmental parameters of temper loss in early childhood: Implications for developmental psychopathology. *Journal of Child Psychology and Psychiatry, 53*, 1099-1108. doi:10.1111/j.1469-7610.2012.02595.x

Wakschlag, L. S., Krogh-Jespersen, S., Estabrook, R., Hlutkowsky, C. O., Anderson, E. L., Burns, J.,…Perlman, S. B. (2020). The Early Childhood Irritability-Related Impairment Interview (E-CRI): A novel method for assessing young children's developmentally impairing irritability. *Behavior Therapy, 51*, 294–309. doi:10.1016/j.beth.2019.07.008

Wiggins, J. L., Mitchell, C., Stringaris, A., & Leibenluft, E. (2014). Developmental trajectories of irritability and bidirectional associations with maternal depression. *Journal of the American Academy of Child and Adolescent Psychiatry, 53*, 1191-1205. doi:10.1016/j.jaac.2014.08.005

World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders*: *Clinical descriptions and diagnostic guidelines* (10th rev.). Geneva: Author.

Appendix A

Measures and Items (*N* = 76) Sampled for Irritability Scale

Measure	Domain	Items[a]	Included[b]
PROMIS Pediatric Anger Scale (PAS; Irwin et al., 2012)	Anger	1. My child felt mad. 2. My child was so angry s/he felt like yelling at somebody. 3. My child was so angry s/he felt like throwing something. 4. My child felt upset. 5. When my child got mad, s/he stayed mad.	Items 1, 4, 5 (tonic) Items 2, 3 (phasic)
State-Trait Anger Expression Inventory (STAXI-CA; del Barrio et al., 2004)	Anger	1. I am furious 2. I feel irritated 3. I feel angry 4. I feel like yelling at somebody 8. I feel like hitting someone 11. I feel annoyed 12. I feel like kicking someone 15. I want to smash something 20. I have a bad temper 21. I get angry very quickly 24. I fly off the handle 25. When I get mad, I say nasty things 30. I express my anger 37. I do things like slamming doors 40. I argue with others	Items 1, 2, 3, 11, 21 (tonic) Items 8, 12, 15, 20, 24, 25, 56 (phasic) <u>Not included:</u> Item 4 (redundant with PAS) Item 30 (nonspecific) Item 37 (redundant with IDA) Item 40 (nonspecific)

		56. I attack whatever makes me angry	
Difficulties in Emotion Regulation Scale—Parent Report (DERS-P; Bunford et al., 2018)	Emotion dysregulation	3. My child experiences his/her emotions as overwhelming and out of control 11. When my child is upset, he/she becomes angry with him/herself for feeling that way 14. When my child is upset, he/she becomes out of control 19. When my child is upset, he/she feels out of control 29. When my child is upset, he/she becomes irritated with him/herself for feeling that way 32. When my child is upset, he/she loses control over his/her behaviors 35. When my child is upset, it takes him/her a long time to feel better 36. When my child is upset, his/her emotions feel overwhelming	Items 3, 11, 19, 29, 35, 36 (tonic) Items 14, 32 (phasic)
Emotion Reactivity Scale (ERS; Nock et al., 2008)	Emotion reactivity	3. When I experience emotions, I feel them very strongly/intensely. 6. I experience emotions very strongly. 11. When I am angry/upset, it takes me much longer than most people to calm down. 12. I get angry at people very easily. 14. I am easily agitated. 15. My emotions go from neutral to extreme in an instant.	Items 3, 11, 12, 14, 16, 19 (tonic) Item 15 (phasic) <u>Not included:</u> Item 6 (redundant with Item 3)

		16. When something bad happens, my mood changes very quickly. People tell me I have a very short fuse. 19. My moods are very strong and powerful.	
Emotion Regulation Index for Children and Adolescents (ERICA; MacDermott et al., 2010)	Emotion regulation	6. When things don't go my way I get upset easily 8. I have angry outbursts 13. I get angry when adults tell me what I can and cannot do	Items 6, 13 (tonic) Item 8 (phasic)
Affective Reactivity Index (ARI; Stringaris et al., 2012)	Irritability (child)	1. Is easily annoyed by others. 2. Often loses his/her temper. 3. Stays angry for a long time. 4. Is angry most of the time. 5. Gets angry frequently. 6. Loses temper easily. 7. Overall irritability causes him/her problems.	Items 1, 2, 3, 4, 5, 6 (tonic) <u>Not included:</u> Item 7 (redundant with scale given assessment of global irritability)
Child Behavior Checklist (CBCL; Achenbach et al., 2001; Wiggins et al., 2014; Tseng et al., 2017)	Irritability (child)	86. stubborn, sullen or irritable 87. sudden changes in mood or feelings 95. temper tantrums or hot temper	Items 86, 87 (tonic) <u>Not included:</u> Item 95 (redundant with MAP-DB)
Irritability Inventory (II; Carlson et al., 2016)	Irritability (child)	2a. S/he rarely gets angry but when s/he does, the explosion is huge compared to the incident that provoked it.	Item 2b (tonic) Item 2a (phasic)

		2b. S/he is mostly reasonable but has days at a time where s/he is very touchy and gets very angry very easily. 2c. S/he has always been short-tempered and easily angered.	<u>Not included:</u> Item 2c (redundant with ERS, ARI)
Irritability, Depression, Anxiety Scale (IDA; Snaith et al., 1978)	Irritability (adult)	4. I lose my temper and shout or snap at others. 12. I feel I might lose control and hit or hurt someone. 16. People upset me so that I feel like slamming doors or banging about.	Items 4, 12, 16 (phasic)
Brief Irritability Test (BITe; Holtzman et al., 2015)	Irritability (adult)	1. I have been grumpy 2. I have been feeling like I might snap 3. Other people have been getting on my nerves 4. Things have been bothering me more than they normally do 5. I have been feeling irritable	Items 1, 2, 3, 4 (tonic) <u>Not included:</u> Item 5 (redundant with STAXI-CA, CBCL)
Irritability Questionnaire (IRQ; Craig et al., 2008)	Irritability (adult)	9. Lately I have felt frustrated. 11. When I am irritated, I need to vent my feelings immediately. 18. I've been feeling like a bomb, ready to explode.	Items 9, 11 (tonic) Item 18 (phasic)
Multidimensional Assessment of Preschool Disruptive Behavior—Temper	Temper loss	1. Have a temper tantrum, fall-out, or melt-down. 2. Stamp feet or hold breath during a temper tantrum, fall-out, or melt-down.	Items 1, 2, 3, 4, 5, 6 (phasic)

Loss (MAP-DB; Wakschlag et al., 2012, 2014)	3. Have a temper tantrum, fall-out, or melt-down that lasted more than 5 minutes. 4. Keep on having a temper tantrum, fall-out, or melt-down, even when you tried to help him/her calm down. 5. Break or destroy things during a temper tantrum, fall-out, or melt-down. 6. Have a temper tantrum, fall-out, or melt-down until exhausted. 7. Hit, bite, or kick during a temper tantrum, fall-out, or melt-down. 8. Lose temper or have a tantrum with you or other parent. 9. Lose temper or have a tantrum with other adults (e.g., teacher, babysitter, family member). 10. Lose temper or have a tantrum when frustrated, angry, or upset. 11. Lose temper or have a tantrum when tired, hungry, or sick. 12. Lose temper or have a tantrum to get something s/he wanted. 13. Lose temper or have a tantrum during daily routines, such as bedtime, mealtime, or getting dressed. 14. Lose temper or have a tantrum "out of the blue" or for no reason. 15. Become frustrated easily. 16. Yell angrily at someone. 17. Act irritable.	Item 20 (tonic) Items 7, 8, 9, 10, 11, 12, 13, 14, 21 (phasic) <u>Not included:</u> Item 15 (redundant with ERS, ERICA, IRQ) Item 16 (redundant with PAS) Item 17 (redundant with STAXI-CA, CBCL) Item 18 (redundant with DERS-P, ERS) Item 19 (redundant with ERS)

		18. Have difficulty calming down when angry. 19. Have a short fuse (become angry quickly). 20. Get extremely angry. 21. Have a hot or explosive temper. 22. Stay angry for a long time.	Item 22 (redundant with PAS, ERS)
Temperament in Middle Childhood Questionnaire (TMCQ; Simonds, 2006)	Temperament	53. Gets very angry when another child takes his/her toy away. 61. Gets angry when called in from play before s/he is ready to quit.	Items 53, 61 (tonic)
		87. Gets angry when s/he can't find something s/he is looking for. 94. Gets angry when s/he has trouble with a task. 146. Gets mad when provoked by other children.	Items 87, 94, 146 (tonic)

[a]Items contained in this column were identified as conceptually overlapping with tonic or phasic irritability. Items not contained in this column (i.e., additional items from full measure) were identified as not conceptually overlapping with tonic or phasic irritability.

[b]Items contained in this column were identified as nonredundant indicators for tonic or phasic irritability (labelled in parentheses). Items with redundant content, either within the same measure or compared to another measure, were identified as "not included" (with rationale given in parentheses).

Appendix B

Novel Irritability Instruments

Three novel instruments that assess youth irritability include two semi-structured interviews for toddlers/preschoolers (Early Childhood Irritability-Related Impairment Interview [E-CRI]; Wakschlag et al., 2020) and children/adolescents (Clinician Affective Reactivity Index [CL-ARI]; Haller et al., 2020) as well as one parent-report rating scale for youth with autism spectrum disorders (Aberrant Behavior Checklist Irritability Subscale [ABC-I]; Stoddard, Zik, Mazefsky, DeChant, & Gabriels, 2020). The E-CRI (Wakschlag et al., 2020) showed promising psychometric properties and found "mood" and "tantrum" factors that were moderately correlated, which lends further evidence to a tonic/phasic approach to measuring irritability. However, the young age ranges in their sample (12 to 18 months, 3 to 5 years) warrant caution when generalizing findings to school-age children. In addition, though the ABC-I (Stoddard et al., 2020) is an important contribution as a novel parent-report rating scale, the nature of the sample (youth with ASD) may not support its use for assessing irritability in typically-developing children. Finally, the CL-ARI (Haller et al., 2020) is a measure of clinician-rated irritability with published psychometric data for use with children and adolescents (ages 7 to 17), though it developed its three subscales (temper outbursts, irritable mood between outbursts, and impairment) without subjecting them to rigorous statistical analysis (e.g., did not conduct exploratory factor analysis). Each of the measures above demonstrate important strengths (e.g., conceptual overlap with tonic/phasic irritability) but have limitations (e.g., age of sample, population of interest, analytical framework) that underscore why Beauchaine

and Tackett (2020) contend that there remain no psychometrically validated measures of tonic and phasic irritability in youth.

Appendix C

Attention Check Questions

Variable	Item	Correct Response	Citation
attn1	Gender (parent)	Female (0)	-
attn2	Please answer “Yes.”	Yes (1)	Flessner et al., 2017
attn3	Please select the "Almost Never" response option.	Almost Never (2)	Parent et al., 2017
attn4	Are you still reading this?	Yes (1)	Flessner et al., 2017
attn5	Please drag the slider to 100.	100	-

Note. Attn1 and attn 2 were embedded in the parent demographic questionnaire; attn3 was embedded in the irritability rating scale; attn4 was embedded in the Disruptive Behavior Disorder Rating Scale; attn5 was embedded in the Impairment Rating Scale.

Appendix D

Preliminary Irritability Scale Item Pool – Tonic Items (*n* = 43)

Item	Measure
My child feels mad.	PAS[a]
My child feels upset.	PAS[a]
When my child gets mad, s/he stays mad.	PAS[a]
My child feels furious.	STAXI-CA[a]
My child feels irritated.	STAXI-CA[a]
My child feels annoyed.	STAXI-CA[a]
My child feels angry.	STAXI-CA[a]
My child gets angry very quickly.	STAXI-CA[a]
My child experiences his/her emotions as overwhelming and out of control.	DERS-P
When my child is [upset], s/he becomes angry with him/herself for feeling that way.	DERS-P
When my child is [upset], s/he feels out of control.	DERS-P
When my child is [upset], s/he becomes irritated with him/herself for feeling that way.	DERS-P
When my child is [upset], it takes him/her a long time to feel better.	DERS-P
When my child is [upset], his/her emotions feel overwhelming.	DERS-P
When my child experiences emotions, s/he feels them very strongly/intensely.	ERS
When my child is angry/upset, it takes him/her much longer than most people to calm down.	ERS
My child gets angry at people very easily.	ERS
My child is easily agitated.	ERS
When something bad happens, my child's mood changes very quickly. People tell me s/he has a very short fuse.	ERS
My child's moods are very strong and powerful.	ERS
When things don't go my child's way s/he gets upset easily	ERICA

Item	Measure
My child gets angry when adults tell him/her what s/he can and cannot do	ERICA
My child is easily annoyed by others.	ARI[a]
My child often loses his/her temper.	ARI[a]
My child stays angry for a long time.	ARI[a]
My child is angry most of the time.	ARI[a]
My child gets angry frequently.	ARI[a]
My child loses temper easily.	ARI[a]
My child is stubborn, sullen or irritable.	CBCL[a]
My child experiences sudden changes in mood or feelings.	CBCL[a]
My child is mostly reasonable but has days at a time where s/he is very touchy and gets very angry very easily.	II
My child has been grumpy.	BITe
My child has been feeling like s/he might snap.	BITe
Other people have been getting on my child's nerves.	BITe
Things have been bothering my child more than they normally do.	BITe
Lately my child has felt frustrated.	IRQ
When my child is irritated, s/he needs to vent his/her feelings immediately.	IRQ
My child gets extremely angry.	MAP-DB
My child gets very angry when another child takes his/her toy away.	TMCQ[a]
My child gets angry when called in from play before s/he is ready to quit.	TMCQ[a]
My child gets angry when s/he can't find something s/he is looking for.	TMCQ[a]
My child gets angry when s/he has trouble with a task.	TMCQ[a]
My child gets mad when provoked by other children.	TMCQ[a]

Note. All item wording modified to be consistent with parent-report language. PAS = PROMIS Pediatric Anger Scale; STAXI-CA = State-Trait Anger Expression Inventory; DERS-P = Difficulties in Emotion Regulation Scale–Parent Report; ERS = Emotion

Reactivity Scale; ERICA = Emotion Regulation Index for Children and Adolescents; ARI = Affective Reactivity Index; CBCL = Child Behavior Checklist; II = Irritability Inventory; BITe = Brief Irritability Test; IRQ = Irritability Questionnaire; MAP-DB = Multidimensional Assessment of Preschool Disruptive Behavior—Temper Loss; TMCQ = Temperament in Middle Childhood Questionnaire.

[a]Item was not included in the irritability scale because the authors indicated that replicating individual items without receiving written permission and/or without paying per use to administer the instrument would violate the copyright terms.

Appendix E

Preliminary Irritability Scale Item Pool – Phasic Items (*n* = 33)

Item	Measure
My child was so angry s/he felt like yelling at somebody.	PAS[a]
My child was so angry s/he felt like throwing something.	PAS[a]
My child feels like hitting someone.	STAXI-CA[a]
My child feels like kicking someone.	STAXI-CA[a]
My child wants to smash something.	STAXI-CA[a]
My child has a bad temper.	STAXI-CA[a]
My child flies off the handle.	STAXI-CA[a]
When my child gets mad, s/he says nasty things.	STAXI-CA[a]
My child attacks whatever makes him/her angry.	STAXI-CA[a]
When my child is [upset], s/he becomes out of control.	DERS-P
When my child is [upset], s/he loses control over his/her behaviors.	DERS-P
My child's emotions go from neutral to extreme in an instant.	ERS
My child has angry outbursts.	ERICA
S/he rarely gets angry but when s/he does, the explosion is huge compared to the incident that provoked it.	II
My child loses his/her temper and shouts or snaps at others.	IDA
My child feels like s/he might lose control and hit or hurt someone.	IDA
People upset my child so that s/he feels like slamming doors or banging about.	IDA
My child has been feeling like a bomb, ready to explode.	IRQ
My child has temper tantrums, fall-outs, or melt-downs.	MAP-DB
My child stamps their feet or holds their breath during a temper tantrum.	MAP-DB
My child has temper tantrums that lasted more than 5 minutes.	MAP-DB
My child keeps on having a temper tantrum, even when I tried to help him/her calm down.	MAP-DB
My child breaks or destroys things during a temper tantrum.	MAP-DB

Item	Measure
My child has a temper tantrum until s/he is exhausted.	MAP-DB
My child hits, bites, or kicks during a temper tantrum.	MAP-DB
My child loses temper or has a tantrum with me or other parent.	MAP-DB
My child loses temper or has a tantrum with other adults (e.g., teacher, babysitter, family member).	MAP-DB
My child loses temper or has a tantrum when frustrated, angry, or upset.	MAP-DB
My child loses temper or has a tantrum when tired, hungry, or sick.	MAP-DB
My child loses temper or has a tantrum to get something s/he wanted.	MAP-DB
My child loses temper or has a tantrum during daily routines, such as bedtime, mealtime, or getting dressed.	MAP-DB
My child loses temper or has a tantrum "out of the blue" or for no reason.	MAP-DB
My child has a hot or explosive temper.	MAP-DB

Note. All item wording modified to be consistent with parent-report language. PAS = PROMIS Pediatric Anger Scale; STAXI-CA = State-Trait Anger Expression Inventory; DERS-P = Difficulties in Emotion Regulation Scale–Parent Report; ERS = Emotion Reactivity Scale; ERICA = Emotion Regulation Index for Children and Adolescents; II = Irritability Inventory; IDA = Irritability, Depression, Anxiety Scale; IRQ = Irritability Questionnaire; MAP-DB = Multidimensional Assessment of Preschool Disruptive Behavior—Temper Loss.

[a]Item was not included in the irritability scale because the authors indicated that replicating individual items without receiving written permission and/or without paying per use to administer the instrument would violate the copyright terms.

Appendix F

Preliminary Irritability Rating Scale (N = 27)

Instructions: The following questions will ask you to compare the behavior of your child to others his/her age.

	Item	Domain
1.	When my child is upset, s/he becomes angry with him/herself for feeling that way.	Tonic
2.	When my child is upset, s/he becomes irritated with him/herself for feeling that way.	Tonic
3.	My child gets angry at people very easily.	Tonic*
4.	My child is easily agitated.	Tonic*
5.	When something bad happens, my child's mood changes very quickly. People tell me s/he has a very short fuse.	Tonic*
6.	When things don't go my child's way s/he gets upset easily.	Tonic*
7.	My child gets angry when adults tell him/her what s/he can and cannot do.	Tonic*
8.	My child has days at a time where s/he is touchy and gets angry easily.	Tonic*
9.	My child is grumpy.	Tonic*
10.	My child feels like s/he might snap.	Tonic
11.	Other people get on my child's nerves.	Tonic*
12.	Things bother my child more than they normally do.	Tonic*
13.	My child feels frustrated.	Tonic*

Item	Domain
14. When my child is irritated, s/he needs to vent his/her feelings immediately.	Tonic
15. My child gets extremely angry.	Tonic*
16. When my child is upset, s/he loses control over his/her behaviors.	Phasic*
17. My child has angry outbursts.	Phasic*
18. My child loses his/her temper and shouts or snaps at others.	Phasic*
19. My child slams doors or bangs about when people upset him/her.	Phasic*
20. My child feels like a bomb, ready to explode.	Phasic*
21. My child has temper tantrums or melt-downs.	Phasic*
22. My child loses their temper or has a tantrum with other adults (e.g., teacher, babysitter, family member).	Phasic*
23. My child loses their temper or has a tantrum when frustrated, angry, or upset.	Phasic*
24. My child loses their temper or has a tantrum when tired, hungry, or sick.	Phasic*
25. My child loses their temper or has a tantrum during daily routines, such as bedtime, mealtime, or getting dressed.	Phasic*
26. My child loses their temper or has a tantrum out of the blue or for no reason.	Phasic
27. My child has a hot or explosive temper.	Phasic*

Note. Parent participants completed all items as part of the full survey. Items were rated on a 5-point scale (1 = never, 2 = almost never, 3 = sometimes, 4 = often, 5 = almost always).

*Item included in final 22-item irritability scale.

Appendix G

Brief Screener

Instructions: Please answer the following questions about yourself and your child(ren).

1. What is your age?	a. under 18 years b. 18-25 years c. 26-35 years d. 36-45 years e. 46-55 years f. over 55 years
2. Please indicate how many children you have within the following age ranges:	
0 to 2 years old	a. none b. one (1) c. two (2) or more
3 to 5 years old	a. none b. one (1) c. two (2) or more
6 to 8 years old	a. none b. one (1) c. two (2) or more
10 to 12 years old	a. none b. one (1) c. two (2) or more
13 to 18 years old	a. none b. one (1) c. two (2) or more
3. Have you or your child(ren)'s teacher thought your child(ren) has/have behavior problems?	a. No b. Yes

Appendix H

Parent Demographic Questionnaire

Instructions: Please provide the following information about **yourself**.

MTurk ID number:	_____
Age (in years):	_____
Gender:	a. Female b. Male
Race: If Other, please specify:	a. African American/Black/African Origin b. Asian American/Asian Origin/Pacific Islander c. American Indian/Alaskan Native d. European Origin/White e. Latino-a/Hispanic f. Bi-racial/Multi-racial g. Other _____
Relationship to child: If Other, please specify:	a. Biological parent b. Stepparent c. Adoptive parent d. Foster parent e. Other _____
Marital status:	a. Single b. Married c. Divorced d. Separated e. Widowed
How many children do you have in your home (including the child on which you are reporting for this study)?	_____

What is the highest level of education you have completed?	a. Less than high school degree b. High school graduate or equivalent (GED) c. Some college (no degree) d. Associate degree e. Bachelor's degree f. Graduate degree (master's or doctoral)
Approximate household yearly income:	a. up to $10,000 b. $10,001-14,999 c. $15,000-24,999 d. $25,000-49,999 e. $50,000-74,999 f. $75,000-99,999 g. $100,000-149,999 h. $150,000-199,999 i. $200,000 or more

Instructions: Please provide the following information about **your child**.

Age (in years):	_____
Gender:	a. Female b. Male
Race: If Other, please specify:	a. African American/Black/African Origin b. Asian American/Asian Origin/Pacific Islander c. American Indian/Alaskan Native d. European Origin/White e. Latino-a/Hispanic f. Bi-racial/Multi-racial g. Other _____
Grade level in school:	a. Kindergarten b. 1st grade c. 2nd grade d. 3rd grade e. 4th grade f. 5th grade g. 6th grade

	h. 7th grade i. 8th grade j. 9th grade k. 10th grade l. 11th grade m. 12th grade
Has your child received a psychiatric and/or mental health diagnoses in the past?	a. no b. yes
If yes, please specify the name of the diagnosis(es):	_____

Appendix I

Impairment Rating Scale—Parent Version

Instructions: In the spaces below, please describe what you see as your child's primary problems in each area, both at home and at school, and describe the effects of your child's problems in that area. Then, mark an "X" on the lines at the points that describe how much the child's problems affect each area and *whether he or she needs treatment or special services for the problems*.

1. How your child's problems affect his or her relationship with playmates

No Problem Definitely does not need treatment or special services |_____| Extreme Problem Definitely needs treatment or special services

2. Regardless of whether your child is popular or unpopular with peers, does he or she have a special, close "best friend" that he or she has kept for more than a few months?

YES NO

3. How your child's problems affect his or her relationship with brothers or sisters

□ My child does not have siblings □ My child does not have regular contact with siblings

No Problem Definitely does not need treatment or special services |_____| Extreme Problem Definitely needs treatment or special services

4. How your child's problems affect his or her relationship with you (and a parenting partner if present)

No Problem Definitely does not need treatment or special services		Extreme Problem Definitely needs treatment or special services

5. How your child's problems affect his or her academic progress at school

No Problem Definitely does not need treatment or special services		Extreme Problem Definitely needs treatment or special services

6. How your child's problems affect his or her self-esteem

No Problem Definitely does not need treatment or special services		Extreme Problem Definitely needs treatment or special services

7. How your child's problems affect your family in general

No Problem Definitely does not need treatment or special services |_____| Extreme Problem Definitely needs treatment or special services

8. **Overall** severity of your child's problem in functioning and overall need for treatment.

No Problem Definitely does not need treatment or special services |_____| Extreme Problem Definitely needs treatment or special services

Appendix J

Parent DBD Rating Scale

Instructions: Check the column that best describes your child.

	Not at All	Just a Little	Pretty Much	Very Much	Don't Know
1. Often argues with adults					
2. Is often spiteful or vindictive					
3. Often blames others for his or her mistakes or misbehavior					
4. Often actively defies or refuses to comply with adults' requests or rules					
~~5. Is often angry and resentful~~					
~~6. Is often touchy or easily annoyed by others~~					
~~7. Often loses temper~~					
8. Often deliberately annoys people					

Note. Items 5, 6, and 7 were not be included in validation analyses due to their overlap with tonic/phasic irritability.

Appendix K

Means and Intercorrelations between Irritability Rating Scale Items – Cohort 1

	Irr1	Irr2	Irr3	Irr4	Irr5	Irr6	Irr7	Irr8	Irr9	Irr10	Irr11	Irr12	Irr13	Irr14	Irr15	Irr16	Irr17
Irr2	.91	--															
Irr3	.55	.53	--														
Irr4	.50	.53	.76	--													
Irr5	.52	.54	.76	.75	--												
Irr6	.42	.45	.66	.74	.76	--											
Irr7	.45	.40	.61	.58	.59	.62	--										
Irr8	.49	.51	.70	.75	.72	.66	.60	--									
Irr9	.40	.45	.58	.66	.61	.57	.47	.74	--								
Irr10	.52	.51	.66	.63	.65	.53	.59	.72	.63	--							
Irr11	.49	.53	.65	.66	.62	.56	.56	.64	.53	.69	--						
Irr12	.51	.54	.67	.70	.64	.65	.58	.73	.69	.70	.69	--					
Irr13	.47	.52	.60	.65	.54	.54	.54	.67	.67	.64	.65	.74	--				
Irr14	.44	.42	.47	.42	.44	.49	.47	.44	.33	.33	.37	.43	.49	--			
Irr15	.54	.49	.69	.67	.69	.64	.60	.69	.61	.71	.63	.68	.56	.49	--		
Irr16	.51	.49	.64	.66	.63	.62	.61	.64	.59	.64	.55	.66	.57	.49	.78	--	
Irr17	.54	.52	.70	.67	.69	.63	.58	.71	.62	.63	.65	.66	.62	.52	.84	.80	--
Irr18	.51	.54	.72	.67	.69	.61	.57	.67	.64	.62	.68	.66	.60	.49	.77	.77	.89
Irr19	.42	.42	.59	.53	.55	.45	.47	.57	.51	.51	.43	.47	.48	.47	.63	.67	.68
Irr20	.54	.51	.66	.63	.65	.55	.59	.68	.63	.79	.58	.64	.64	.40	.73	.73	.69
Irr21	.43	.43	.56	.69	.63	.66	.56	.64	.66	.53	.51	.62	.56	.41	.62	.73	.69
Irr22	.36	.35	.57	.59	.58	.54	.62	.60	.53	.58	.59	.60	.54	.42	.57	.68	.61
Irr23	.50	.50	.64	.67	.63	.67	.60	.67	.64	.60	.57	.69	.64	.56	.70	.78	.76
Irr24	.36	.38	.46	.56	.46	.49	.46	.59	.57	.46	.47	.55	.60	.47	.51	.55	.59
Irr25	.36	.37	.52	.59	.52	.57	.48	.63	.62	.59	.52	.65	.54	.38	.58	.64	.59
Irr26	.44	.43	.47	.53	.50	.44	.48	.63	.59	.70	.52	.60	.53	.32	.58	.59	.55
Irr27	.49	.47	.70	.68	.70	.61	.56	.66	.59	.71	.60	.66	.57	.43	.78	.75	.73

Item-scale	.62	.63	.80	.82	.80	.75	.70	.83	.75	.78	.74	.81	.75	.56	.83	.83	.85
M	2.66	2.68	2.76	3.02	2.88	3.25	2.94	2.80	2.73	2.16	2.74	2.62	3.05	3.16	2.38	2.59	2.54
SD	1.12	1.16	1.12	1.13	1.23	1.10	1.05	1.13	.94	1.06	.97	1.13	.91	1.04	1.08	1.15	1.05

	Irr18	Irr19	Irr20	Irr21	Irr22	Irr23	Irr24	Irr25	Irr26	Irr27
Irr19	.69	--								
Irr20	.70	.69	--							
Irr21	.64	.57	.64	--						
Irr22	.63	.54	.65	.69	--					
Irr23	.72	.63	.68	.80	.71	--				
Irr24	.55	.48	.55	.68	.58	.68	--			
Irr25	.57	.49	.64	.69	.66	.65	.68	--		
Irr26	.51	.50	.76	.62	.57	.61	.54	.73	--	
Irr27	.75	.64	.82	.65	.61	.70	.52	.66	.68	--
Item-scale	.84	.69	.83	.79	.74	.84	.67	.73	.71	.83
M	2.66	2.52	2.13	2.70	2.24	2.76	2.73	2.27	2.07	2.24
SD	1.04	1.12	1.04	1.00	.97	.98	1.08	1.05	.98	1.12

Note. $N = 200$. All correlations are significant at $p < .001$. Item-scale = corrected item-scale correlation; *M* = mean; *SD* = standard deviation.

Appendix L

Means and Intercorrelations between Irritability Rating Scale Items – Cohort 2

	Irr1	Irr2	Irr3	Irr4	Irr5	Irr6	Irr7	Irr8	Irr9	Irr10	Irr11	Irr12	Irr13	Irr14	Irr15	Irr16	Irr17
Irr2	.90	--															
Irr3	.54	.48	--														
Irr4	.54	.53	.77	--													
Irr5	.60	.54	.75	.70	--												
Irr6	.49	.43	.75	.66	.74	--											
Irr7	.56	.46	.61	.56	.64	.68	--										
Irr8	.56	.51	.60	.66	.65	.59	.61	--									
Irr9	.46	.42	.50	.62	.60	.51	.49	.64	--								
Irr10	.53	.48	.67	.67	.73	.67	.61	.64	.63	--							
Irr11	.59	.51	.57	.59	.62	.56	.57	.57	.57	.70	--						
Irr12	.55	.53	.60	.66	.62	.60	.58	.65	.67	.72	.69	--					
Irr13	.48	.46	.54	.57	.60	.59	.52	.60	.57	.61	.54	.67	--				
Irr14	.33	.36	.37	.40	.40	.44	.35	.33	.35	.36	.37	.44	.43	--			
Irr15	.56	.51	.73	.66	.74	.73	.62	.67	.58	.73	.61	.69	.60	.49	--		
Irr16	.50	.49	.61	.60	.69	.64	.53	.60	.51	.70	.47	.61	.57	.43	.74	--	
Irr17	.56	.49	.68	.68	.72	.70	.62	.63	.61	.72	.57	.69	.57	.47	.82	.77	--
Irr18	.56	.52	.69	.63	.70	.71	.60	.68	.60	.72	.62	.69	.62	.43	.83	.77	.81
Irr19	.47	.41	.63	.58	.61	.60	.52	.56	.49	.66	.53	.63	.58	.42	.68	.65	.71
Irr20	.58	.50	.65	.57	.75	.65	.53	.59	.60	.80	.63	.63	.57	.37	.77	.73	.73
Irr21	.51	.48	.61	.62	.66	.64	.57	.71	.61	.68	.52	.66	.55	.44	.77	.71	.75
Irr22	.57	.51	.63	.52	.64	.57	.60	.55	.45	.67	.50	.58	.51	.27	.69	.71	.67
Irr23	.49	.47	.63	.64	.69	.68	.56	.66	.60	.73	.52	.67	.64	.46	.75	.76	.78
Irr24	.39	.42	.50	.51	.58	.56	.42	.52	.42	.56	.47	.50	.51	.28	.53	.58	.57
Irr25	.47	.42	.54	.52	.60	.56	.49	.52	.46	.61	.49	.59	.54	.33	.63	.64	.63
Irr26	.59	.53	.56	.59	.67	.52	.54	.65	.57	.65	.58	.63	.53	.33	.70	.61	.67
Irr27	.61	.53	.68	.61	.77	.68	.64	.66	.58	.74	.57	.62	.60	.38	.82	.73	.79

Item-scale	.69	.63	.78	.77	.84	.78	.71	.77	.69	.84	.71	.79	.71	.49	.87	.80	.86
M	2.64	2.65	2.60	2.77	2.54	3.16	2.75	2.66	2.66	2.05	2.60	2.37	2.93	3.04	2.24	2.23	2.36
SD	1.12	1.10	1.02	1.03	1.21	1.10	1.11	1.13	.91	.98	1.05	1.05	.90	1.02	1.07	1.05	1.05

	Irr18	Irr19	Irr20	Irr21	Irr22	Irr23	Irr24	Irr25	Irr26	Irr27
Irr19	.72	--								
Irr20	.76	.66	--							
Irr21	.75	.66	.71	--						
Irr22	.70	.59	.73	.68	--					
Irr23	.78	.70	.70	.84	.70	--				
Irr24	.58	.51	.52	.61	.49	.70	--			
Irr25	.64	.63	.66	.67	.65	.67	.63	--		
Irr26	.67	.58	.71	.69	.65	.65	.52	.65	--	
Irr27	.78	.66	.84	.75	.72	.74	.51	.63	.74	--
Item-scale	.87	.76	.83	.83	.76	.85	.66	.73	.78	.86
M	2.42	2.48	1.93	2.35	1.95	2.47	2.48	2.02	1.84	2.03
SD	1.09	1.18	1.03	1.10	1.08	1.10	1.01	.99	.94	1.07

Note. $N = 197$. All correlations are significant at $p < .001$. Item-scale = corrected item-scale correlation; *M* = mean; *SD* = standard deviation.

Thesis and Dissertation Services