Assessment. Author manuscript; available in PMC 2015 February 05.

Published in final edited form as:

Assessment. 2015 February ; 22(1): 101–115. doi:10.1177/1073191114533814.

The Brief Irritability Test (BITe): A Measure of Irritability for Use Among Men and Women

Susan Holtzman¹, Brian P. O'Connor¹, Paula C. Barata², and Donna E. Stewart^{3,4}

¹University of British Columbia, Kelowna, British Columbia, Canada

²University of Guelph, Guelph, Ontario, Canada

³University Health Network, Toronto, Ontario, Canada

⁴University of Toronto, Toronto, Ontario, Canada

Abstract

Elevated levels of irritability have been reported across a range of psychiatric and medical conditions. However, research on the causes, consequences, and treatments of irritability has been hindered by limitations in existing measurement tools. This study aimed to develop a brief, reliable, and valid self-report measure of irritability that is suitable for use among both men and women and that displays minimal overlap with related constructs. First, 63 candidate items were generated, including items from two recent irritability scales. Second, 1,116 participants (877 university students and 229 chronic pain outpatients) completed a survey containing the irritability item pool and standardized measures of related constructs. Item response theory was used to develop a five-item scale (the Brief Irritability Test) with a strong internal structure. All five items displayed minimal conceptual overlap with related constructs (e.g., depression, anger), and test scores displayed negligible gender bias. The Brief Irritability Test shows promise in helping to advance the burgeoning field of irritability research.

Keywords

irritability; assessment; item response theory

Irritability is a mood characterized by a proneness to experience negative affective states, such as anger, annoyance, and frustration upon little provocation, and may be outwardly expressed in the form of aggressive behavior (Born & Steiner, 1999; Craig, Hietanen, Markova, & Berrios, 2008; Snaith & Taylor, 1985; Stringaris, 2011). Elevated levels of irritability have been reported across a wide array of medical and psychiatric conditions, ranging from chronic pain and nicotine withdrawal, to mood, anxiety, and neuropsychiatric disorders (Mangelli et al., 2006; Perlis et al., 2009; Sofaer & Walker, 1994; Youn et al.,

Corresponding Author: Susan Holtzman, Department of Psychology, University of British Columbia, ASC 283, 3333 University Way, Kelowna, British Columbia, V1V 1V7, Canada. susan.holtzman@ubc.ca.

Reprints and permissions: sagepub.com/journalsPermissions.nav

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

2011). Scientific interest in the construct of irritability has increased dramatically over the past several years, due in part to a number of recent findings in the literature. For example, at least 40% to 50% of depressed adults suffer from irritability and irritability has been linked with a greater lifetime persistence of depression, risk of suicide, and reduced quality of life (Fava et al., 2010; Perlis et al., 2009; Pickles et al., 2010). According to a recent 20-year prospective study, adolescent irritability can predict the development of mood and anxiety disorders during adulthood, as well as lower income and worse educational achievement (Stringaris, Cohen, Pine, & Leibenluft, 2009). High levels of irritability have also been shown to generate stressful interpersonal events (Sahl, Cohen, & Dasch, 2009), erode marital relationships (MacEwen & Barling, 1993), and are associated with greater cardiovascular reactivity to stress (Caprara et al., 1985). Taken together, these findings have triggered widespread calls for further research into the causes, consequences, and treatment of irritability (Born & Steiner, 1999; DiGiuseppe & Tafrate, 2007; Fava et al., 2010; Stringaris, 2011).

Despite the growing scientific interest in irritability, progress has been hindered by limitations in current measurement tools. In fact, the vast majority of irritability research has relied on single-item assessments (e.g., “how often have you felt irritable during the past week?”), which generally produce less reliable and valid results than multi-item scales (Bowling, 2005; Liu, 2003). Participants tend to give less consistent responses to single questions over time, and this may be due to changes in the interpretation of individual questions over time, as well as fluctuating situational factors. Indeed, our qualitative research has revealed that the ways in which the lay public defines and understands the term *irritability* varies not only between individuals but also within individuals over time (Barata, Holtzman, & Cunningham, 2012).

Several multi-item self-report scales have been designed specifically to assess irritability. The first published self-report measure of irritability was an 11-item subscale of the Buss–Durkee Hostility Inventory (Buss & Durkee, 1957). The Buss–Durkee Hostility Inventory conceptualizes irritability as a stable trait and uses a true–false response set. Twenty years later, Snaith and colleagues developed the Irritability, Depression, Anxiety Scale, which includes an eight-item irritability subscale assessing state irritability (Snaith, Constantopoulos, Jardine, & McGuffin, 1978). The lengthiest measure of irritability is the 30-item Caprara Irritability Scale (Caprara et al., 1985), which also conceptualizes irritability as a stable trait. The Minnesota Multiphasic Personality Inventory–2 (Butcher et al., 2001) and the Personality Assessment Inventory (Morey, 1991) each contain irritability subscales. However, the Personality Assessment Inventory subscale was designed to assess irritability within the context of mania or hypomania. As a testament to the growing scientific interest in irritability, four new measures have been published in the past 4 years. Craig et al. (2008) developed the Irritability Questionnaire (IRQ), a 21-item measure designed to reflect the multidimensional nature of irritability. The IRQ was developed on a small sample of patients with Alzheimer’s, Huntington’s, and affective disorders, and a control group. The 14-item Born–Steiner Irritability Scale (BSIS) was developed for use among women suffering from female-specific mood disorders (Born, Koren, Lin, & Steiner, 2008). The 18-item I-Epi was created specifically to measure irritability in epilepsy patients (Piazzini et al., 2011). Most recently, Stringaris et al. (2012) designed a seven-item parent

and self-report scale to assess childhood and adolescent irritability that focuses on the frequency and duration of angry feelings.

In general, multi-item irritability scales have demonstrated strong reliability and assess a range of thoughts, feelings, and behaviors associated with irritability. Despite these strengths, there are a number of reasons to be cautious in adopting these measures. A primary concern is that current measures tap into constructs other than irritability, such as *overt aggression* (e.g., “Sometimes I shout, hit, and kick and let off steam”), *hostility* (e.g., “I can’t help being a little rude to people I don’t like”), and *depression* (e.g., “I feel like harming myself”; Born et al., 2008; Caprara et al., 1985; Craig et al., 2008; Snaith et al., 1978). Irritability scales are most often confounded with anger items (e.g., “When I get angry, I use bad language and swear,” “I get angry frequently”). This is problematic since, unlike anger, irritability often lacks a direct cause, is less intense, less disturbed, longer in duration, and is associated with greater efforts to control its experience and expression (Barata et al., 2012; Beedie, Terry, & Lane, 2005; DiGiuseppe & Tafrate, 2010). Second, state measures of irritability are often confounded with items that reflect general dispositions or traits (e.g., “Arguments are a major cause of stress in my relationships,” “When I am right, I am right”; Caprara et al., 1985; Craig et al., 2008). Third, two-dimensional (inward- and outward-directed; Snaith et al., 1978) and multidimensional (behavioral, physiological, cognitive, affective; Craig et al., 2008) structures of irritability have been proposed; however, studies have failed to statistically test for underlying facets. One exception is a study by Caprara et al. (1985) who found one clear dominant factor of irritability using principal components analysis. Fourth, irritability scales may be unduly long, making them unappealing to researchers and clinicians who wish to integrate a scale into their protocols.

One final and significant concern is that item and test bias has not been explored in current irritability measures. Bias exists when persons from different groups have identical levels of a latent trait, but score differently on the same item or test. In measuring irritability, the issue of gender and gender bias is particularly relevant. Irritability has been described as a key feature of female-specific mood disturbances, such as premenstrual dysphoric disorder (Caplan, 2004), as well as the “male depressive syndrome” (Rutz, von Knorring, Pihlgren, Rihmer, & Walinder, 1995). Some studies have found more frequent and intense levels of irritability among women (Perlis et al., 2009; Piazzini et al., 2011), whereas others have failed to find significant gender differences (Fava et al., 2010; Verhoeven, Booij, Van der Wee, Penninx, & Van der Does, 2011; Marcus et al., 2008). Unfortunately, the extent to which gender differences are due to test bias or actual gender differences in irritability remains unclear (Smith & Reise, 1998).

Many of the aforementioned concerns regarding existing measures of irritability can be addressed using an item response theory (IRT) approach to scale development (Embretson & Reise, 2000; Streiner, 2010). Unlike classical test theory statistics, IRT allows researchers to select only the most informative test items, and keep scale length to a minimum. IRT methods also have the ability to investigate possible item and test bias based on gender, and to independently assess bias and group differences on the latent trait. Born et al. (2008) have suggested that men may experience and label irritability differently (e.g., grouchy, miserable, upset) than do women (e.g., impatient, intolerant, short). However, it remains

unclear whether scale items that use these supposed gender-specific descriptions display gender bias. Another advantage of IRT is that it can provide detailed information regarding the reliability and power of irritability items to discriminate between respondents across the full continuum of test scores. Lastly, it can be used to determine the usefulness of response scale options and whether total scale scores are accurate reflections of latent trait scores estimated by the IRT model.

In sum, there have been widespread calls for research into the causes, consequences, and treatment of irritability. However, measurement issues need to be addressed before research can move forward. The first objective of the current study was to develop a measure of irritability that is brief, reliable, and valid, and that displays minimal overlap with related constructs. Although most agree that irritability is conceptually distinct from anger, aggression, hostility, and depression, current measures do not reflect this distinction. According to the dispositional theory of moods (Siemer, 2009), irritability may temporarily predispose one to make angry appraisals or to feel angry, *but* it may also remain as a disposition (Lormand, 1985). In other words, irritability does not have to manifest as an angry appraisal or feeling to still be considered irritability. We hypothesized that irritability could be reliably assessed using fewer items than in past measures, and that items displaying the best psychometric properties would also be conceptually distinct from related constructs. The second objective was to explore the underlying dimensions of irritability. Although multiple facets have been proposed, this has been given little empirical attention. We tentatively hypothesized that irritability items would display a unidimensional structure (Caprara et al., 1985). Our third objective was to assess the suitability of our measure for use among men and women, as well as healthy and chronically ill samples. Specifically, we investigated the presence of item or test bias based on gender or test sample. We conceptualized irritability as a state given past research showing irritability to be a fluctuating symptom of various medical and psychiatric conditions (Stringaris, 2011; Youn et al., 2011) and that it can be temporarily induced by physical states (e.g., blood sugar imbalance; Warren, Deary, & Frier, 2003), and pharmacological interventions (e.g., interferon treatment; Russo et al., 2005).

Method

Initial Item Pool Generation

A total of 63 candidate items were generated for potential inclusion in our measure of irritability. Thirty-five items were obtained from two existing self-report measures of irritability: the 21-item IRQ (Craig et al., 2008) and the 14-item BSIS (Born et al., 2008). These were selected based on their demonstrated reliability in past research, and the quality and comprehensiveness of the scale items. By including these scales in their entirety, we were able to test whether our new scale represented a significant improvement over existing measures.

The remaining 28 candidate items were developed by our research team, based on a careful review of other irritability scales, a content analysis of published definitions of irritability, and qualitative interviews that we conducted with a diverse sample of 39 community-dwelling adults (59% female) who had experienced irritability during the past 2 weeks.

Almost half of our sample (44%) self-reported at least one Axis I psychiatric disorder (major depression [$n = 10$], bipolar disorder [$n = 4$], schizophrenia [$n = 2$], social anxiety disorder [$n = 2$], attention deficit hyperactivity disorder [$n = 2$], posttraumatic stress disorder [$n = 1$]). A detailed description of the methods and results is presented elsewhere (Barata et al., 2012). Briefly, participants were asked to describe their recent experiences with irritability (thoughts, feelings, behaviors, physical sensations), and, more generally, what irritability means to them. Transcripts were analyzed using a deductive thematic approach (Braun & Clarke, 2006). The first author and a graduate research assistant developed a preliminary pool of 48 items that reflected the main themes and subthemes from the thematic analysis, and were not already included in current measures of irritability. These 48 items were reviewed and revised by our research team. The strongest 28 items were selected based on consensus. Item selection was based on clarity and construct validity, and ensured that items were reflective of the multidimensional range of thoughts, feelings, behaviors, and sensations that have been used to describe and assess irritability in past research (e.g., Born & Steiner, 1999; Craig et al., 2008; Snaith & Taylor, 1985).

Scale Development

Participant Recruitment—An undergraduate student sample and an outpatient chronic pain sample were recruited. A community pain sample was chosen because of past studies showing elevated rates of psychiatric conditions (e.g., mood and anxiety disorders) and frequent negative mood in this population (Burns, Quartana, & Bruel, 2008; McWilliams, Cox, & Enns, 2003). Students, although generally in good physical health, may also be more likely to report problems with irritability (compared with those who are employed full-time, retired, or homemakers; Fava et al., 2010). Combining student and patient samples ensured heterogeneity with respect to levels of irritability (a heterogeneous sample that covers the full spectrum of the latent trait is important for IRT to properly estimate item parameters). This also helped to generate a sample that was diverse with respect to age, gender, socioeconomic status, and health status.

The student sample was recruited through the Psychology Research subject pools at two universities located in mid-sized Canadian cities. To be eligible, participants were required to be undergraduate students and fluent in English. Students received course credit for participating. Pain patients were recruited from two outpatient pain clinics (one public and one private) in a mid-sized Canadian city. Individuals were required to have been experiencing (nonmalignant) pain for at least 6 months, be at least 18 years old, and fluent in English. Recruitment strategies included advertisements in outpatient waiting rooms, approaching patients during clinic visits, and mailing recruitment letters to patients on the clinic waitlist. Patients were paid \$15 upon receipt of their completed questionnaire package.

Sample Characteristics—The final sample consisted of 1,116 participants (887 students and 229 patients with chronic pain). The student subsample was 64.3% female, 73.5% Caucasian (14.3% East or Southeast Asian, 6.9% South Asian, 8.7% other), and an average of 20 years of age ($SD = 3.5$). The chronic pain subsample was 61.6% female, 93.5% Caucasian, and an average of 57 years of age ($SD = 15$). The majority of pain patients was married (61.9%), had at least a high school education (81.1%), and was either on medical

leave/disability (40.3%) or retired (32.7%) at the time of study completion. More than half of patients (55.2%) reported an annual family income of less than \$40 000. Patients had been suffering from chronic pain for an average of 10.4 years ($SD = 10.4$, range = 0.4–74.5 years). Pain etiology varied widely and included: traumatic injury (27.4%), arthritis (16.1%), fibromyalgia (9.6%), degenerative disc disease (9.1%), and other known (e.g., complex regional pain syndrome, Crohn's disease) or unknown causes.

Survey Administration

Participants completed an online battery of questionnaires assessing their mood, personality, social relationships, well-being, and demographics. In anticipation that our pain sample would be older and have less comfort with, or access to, the Internet, they were given the option to complete a mail-in version of the survey (34% of patients took this option). Those who completed the written survey were more likely to be older ($p < .001$) and to report an annual income of less than \$40 000 ($p = .01$), but did not differ on gender or pain-related variables.

Measures

Irritability Item Pool—Participants were presented with the 63 candidate items and instructed: “Please indicate how often you have felt or behaved in the following ways, during the past two weeks, including today.” Each of the 63 items was rated on a 6-point Likert scale (1 = *never*, 2 = *rarely*, 3 = *sometimes*, 4 = *often*, 5 = *very often*, 6 = *always*). Twenty-eight of these items were developed for this study and the other 35 items were taken verbatim from the 21-item IRQ (Craig et al., 2008) and the 14-item BSIS (Born et al., 2008). We chose a 6-point scale (rather than a 4-point scale, used in the IRQ and BSIS) to obtain more fine-grained information, while keeping scale length to a minimum. The IRQ asks participants to rate the intensity *and* frequency of their irritability. However, since we expected these to be highly correlated, and again, since we were aiming for a short measure we chose to assess frequency only.

Beck Depression Inventory–II—The 21-item Beck Depression Inventory–II (Beck, Steer, & Brown, 1996) is a well-validated measure of depressive symptoms. Item responses range from 0 to 3. The Beck Depression Inventory–II has demonstrated good validity and internal consistency among college student samples (Steer & Clark, 1997) as well as among individuals with chronic pain (Harris & D'Leon, 2008).

NEO Five-Factor Inventory—The 12-item neuroticism subscale was used to assess the tendency to experience negative affect (e.g., “I often feel tense and jittery”). Items are rated on a 5-point scale (1 = *strongly disagree* to 5 = *strongly agree*). The NEO Five-Factor Inventory (Costa & McCrae, 1995) has demonstrated excellent reliability and validity in past research (Thalmayer, Saucier, & Eigenhuis, 2011).

State-Trait Anger Expression Inventory–2—The 10-item trait anger subscale of the State-Trait Anger Expression Inventory–2 (Spielberger, 1999) assessed the dispositional tendency to experience anger (e.g., “I am quick tempered”). This widely used measure asks participants to indicate how often they generally experience angry feelings using a 4-point

scale (1 = *almost never* to 4 = *almost always*). It has demonstrated strong psychometric properties in past research (Spielberger & Reheiser, 2009).

Aggression Questionnaire—The Aggression Questionnaire (Buss & Perry, 1992) is a 29-item scale of dispositional aggression and is comprised of four subscales: verbal aggression (e.g., “tell my friends openly when I disagree with them”), physical aggression (e.g., “If I have to resort to violence to protect my rights, I will”), trait anger (e.g., “I have trouble controlling my temper”), and trait hostility (e.g., “When people are especially nice, I wonder what they want”). Items are rated on a 5-point scale (1 = *extremely uncharacteristic of me* to 5 = *extremely characteristic of me*). The Aggression Questionnaire has moderate to high internal consistency and test–retest reliability, and has shown good convergent validity with other self-report scales of aggression (e.g., Harris, 1997).

Interpersonal Support Evaluation List–12—The Interpersonal Support Evaluation List–12 (Cohen & Hoberman, 1983) was used to assess perceived availability of support and has demonstrated good reliability and validity in past research. Participants indicate the amount and types of support they feel they have in their life (e.g., “There is someone I can turn to for advice about handling problems with my family”) on a 4-point Likert scale ranging from 1 (*definitely false*) to 4 (*definitely true*).

Satisfaction With Life Scale—Global life satisfaction was tested using five items (e.g., “In most ways my life is close to ideal”), rated on a 7-point scale (1 = *strongly disagree* to 7 = *strongly agree*). The Satisfaction With Life Scale (Diener, Emmons, Larsen, & Griffin, 1985) is the most widely used measure of life satisfaction and has strong psychometric properties (Pavot & Deiner, 1993).

Pain and Pain Interference—The pain subsample indicated their pain severity during the past 2 weeks using an 11-point numerical rating scale (0 = *no pain* to 10 = *pain as bad as you can imagine*; Jensen, Karoly, & Braver, 1986). Life interference due to pain was measured using the seven-item interference subscale of the Brief Pain Inventory (short form; Cleeland, 1991).

Demographic and Medical Information—All study participants responded to basic demographic questions. The pain sub-sample was asked more in-depth demographic questions, and questions about the location, duration, and cause of their chronic pain condition.

Statistical Analysis

Missing Data—The rate of missing data was <2%. Listwise deletion of cases with missing values was used rather than data imputation because our focus was on item-level responses.

Dimensionality—Item pools must be unidimensional for IRT analyses. Principal components analyses were therefore conducted on the matrices of polychoric correlations for the irritability item pools. The ratios of the first to second eigenvalues were then computed. A ratio greater than 3.0 indicates the existence of a dominant factor and provides sufficient evidence for unidimensionality for IRT analyses (Morizot, Ainsworth, & Reise,

2007). We also searched for evidence for possible additional dimensions, beyond the dominant factor, that might suggest a role for subscales of irritability. We conducted 1,000 random data set parallel analyses on the polychoric correlation matrices and we ran Velicer's minimum average partial (MAP) test, which focuses on the relative amounts of common and unique variance for the n -factor solutions (Zwick & Velicer, 1986). Goodness-of-fit index (GFI) coefficients for all possible n -factor solutions were also computed. GFI coefficients are useful in eliminating trivial factors that can sometimes just pass the minimum criteria in tests for numbers of factors. Finally, we conducted Schmid–Leiman bifactor analyses, which are ideally suited for revealing the existence and relative importance of group factors that account for residual variation and that are independent of the dominant dimension (Reise, Moore, & Haviland, 2010).

Item Response Theory Analyses

Graded Response Model—The implementation of Samejima's (1969) graded response model (GRM) in the *ltm* package in R (Rizopoulos, 2006) was used for the IRT analyses of the polytomous irritability items. The model provides latent trait estimates for respondents and two kinds of parameters for the individual items. The item discrimination parameter, a , reflects the magnitude of the association between an item and the latent trait. Higher discrimination values indicate that an item provides sharp discrimination between respondents whereas discrimination values near zero indicate that an item provides little information about the latent trait.

The threshold parameter, b , is an index of item difficulty that indicates where along the latent trait continuum a response to an item occurs. For polytomous data, the number of threshold parameters for each item is the number of response options minus one. The irritability items have a 6-point response scale, which results in the following five GRM response dichotomies: (1) Option 1 versus Options 2, 3, 4, and 5; (2) Options 1 and 2 versus Options 3, 4, and 5; (3) Options 1, 2, and 3 versus Options 4 and 5; (4) Options 1, 2, 3, and 4 versus Options 5 and 6; and (5) Options 1, 2, 3, 4, and 5 versus Option 6. The threshold parameter for each of these response dichotomies represents the location on the latent trait continuum where there is a 50% probability of endorsing the higher response option(s).

An information function can be obtained for each item, which reveals an item's ability to differentiate between respondents located at various points along the trait-level continuum. The information functions for the test items are then added to obtain the scale information function, which indicates how well the entire test functions in different latent trait ranges.

Differential Item and Test Functioning—The Differential Functioning of Items and Tests framework (DFIT; Flowers, Oshima, & Raju, 1999; Oshima & Morris, 2008) was used to examine possible measurement bias due to gender and sample type. The DFIT procedure involves estimating the IRT parameters separately for each group, computing linking constants to place the parameters in the same metric, and conducting preliminary DFIT analyses using the rescaled item parameters. Items with significant differential item functioning (DIF) were identified, item parameters for each group were reestimated using only those items without DIF, and revised linking constants were computed. All the original

item parameters for each group were then placed in the same metric using the revised linking constants, parameters were rescaled, and the final DFIT analyses were conducted. Statistical significance was assessed using the random data procedure recommended by Oshima, Raju, and Nanda (2006).

We report the noncompensatory DIF index, which quantifies DIF in a particular item under the assumption that all other items are DIF free. It is the average squared difference between the expected item endorsement probabilities. The DTF coefficient represents differential test functioning. The square root of the DTF coefficient (rDTF) represents the differences between the test characteristic curves expressed in the metric of the observed scores and is thus a revealing index of the effect size.

Results

We begin by presenting findings for four pools of items: (1) the 21 IRQ items (Craig et al., 2008), (2) the 14 BSIS items (Born et al., 2008), (3) the 28 items generated for this study, and (4) all three item pools combined. We then present findings for our new five-item measure.

Dimensionality Analyses

The results of the dimensionality tests for the four pools of irritability items are provided in Table 1. The findings were relatively consistent. The first eigenvalues in the matrices of polychoric correlations were large and much greater than the second eigenvalues (the first to second ratios ranged from 7 to 10), indicating a single dominant dimension in each item pool. The parallel analysis and MAP test results indicated more than one component for each item pool, but the GFI coefficients for the one-component models were all very high (.97 or greater), indicating that the additional factors identified by the parallel analyses and MAP test were trivial.

The results of the bifactor analyses for each of the item pools are too numerous to report, but there was a clear pattern in the findings that is easily summarized. The item loadings on the group factors were almost all lower, and usually distinctly lower, than the loadings the same items had on the general (dominant) factor. The loadings on the general factor in the bifactor analyses were also very similar to the loadings that the same items displayed in the one-component model. There was little attenuation. These findings were most pronounced in the all-items-combined pool. There were three items in the IRQ pool that had stronger group factor loadings than general factor loadings, but these items loaded on two different group factors. The evidence for additional dimensions or for possible subscales was thus weak.

Item Response Theory Analyses

Separate GRM analyses were conducted on the four pools of irritability items. For each item pool, perusal of the two-way and three-way margins revealed that very few of the many residuals were significant, indicating good fit of the models to the data. We also statistically removed (partialled out) the first, dominant dimension from the matrix of polychoric correlations for each item pool. Very few (10% or less) of the subsequent partial correlations

between the items in these matrices were above .20, indicating no serious violations of local independence.

The GRM threshold and discrimination parameters for all the item pools are too numerous to be reported. We therefore take a summary approach to presenting the findings from the IRQ and BSIS, and we provide more detailed results for our new measure.

The item and test information functions for the IRQ and BSIS are provided in Figures 1, 2, 3, and 4. The test information functions for the two measures were remarkably similar. Both measures provided high levels of discrimination between respondents from approximately $z = -1$ and higher on the latent trait continuum. Discrimination effectiveness tailed off below z values of -1 , only slightly for the IRQ but more noticeably for the BSIS.

There were also very similar patterns in the item information functions for the two measures. In both cases, just three or four items provided most of the psychometric information. The remaining items were minimally effective and/or redundant. The content of the most effective items is revealing. For the IRQ, the best items were (9) “I have felt frustrated,” (14) “I have felt bitter about things,” (16) “When I have looked back on how life has treated me, I’ve felt a bit disappointed and angry” and (18) “I have been feeling like a bomb, ready to explode.” The best four BSIS items were (1) “I have been feeling mad,” (2) “I have been feeling ready to explode,” (5) “I have been easily flying off the handle,” and (6) “It feels like there has been a cloud of anger over me.” Anger and readiness to explode were common themes in the best items.

GRM analyses were then conducted on the 28 items generated by our research team. The test information function for these items had the same pattern as the test information functions for the other two measures. The item information functions indicated that the following seven items provided most of the psychometric information: (2) “I have been grumpy,” (9) “I have been short with people,” (18) “I have been feeling like I might snap,” (22) “Other people have been getting on my nerves,” (25) “I have been getting upset more than usual,” (26) “Things have been bothering me more than they normally do,” and (28) “I have been feeling irritable.” Anger is much less prominent in the contents of these items than was the case for the best items from the other two irritability measures. Subsequent GRM analyses on these seven items revealed one item that was relatively weak and redundant with other items (“I have been short with people”). Another item, “I have been getting upset more than usual,” had an information function that was essentially identical to that for a semantically similar, but slightly stronger item, “Things have been bothering me more than they normally do.” These two items (feeling short, and getting upset) were therefore dropped, leaving us with the five best items from the 28-item pool.

The word “irritable” appears in one of our five best items (“I have been feeling irritable”) and in one item from the BSIS (“I have been irritable when someone touched me”). “Irritable” does not appear in any of the IRQ items. The Pearson correlation between scores on the IRQ and scores on the “I have been feeling irritable” item was .68, and the Pearson correlation between scores on the BSIS and scores on the “I have been feeling irritable” item was .73. The Pearson correlation between scores on the “I have been feeling irritable” item

and the total scores for the other four best items was .77. Therefore, the best items generated by our research team appear to tap into irritability just slightly better than do the items from the other measures.

A variety of further GRM and Pearson correlation analyses were then conducted on the three sets of best items described above, with the intention of developing a “super” measure of irritability. The results of these efforts are easily summarized. There were high levels of item information function overlap and redundancy between the four best items from the IRQ and the four best items from the BSIS. We were impressed with the face validity of the five best items generated from our qualitative interviews, and with the Pearson correlations between the “I have been feeling irritable” item and the other four items. We therefore used the five items as the basis for GRM analyses involving these items and the other eight best items, both individually and in various combinations. The results were consistent. Adding any of the eight best items from the other irritability measures did not improve the test information functions. Furthermore, the additional items were either highly redundant with, or weaker than, the five base items. There were also no meaningful differences in the Pearson correlations between various possible new irritability scales based on five, or more than five, of the best items and measures of depressive symptoms, neuroticism, trait anger and hostility, and physical and verbal aggression. In summary, we could find no reason to include in a measure of irritability any items other than the five best items that were generated from our qualitative interviews with our community sample.

Item and Test Characteristics of the Final Five-Item Scale

Results from the dimensionality analyses for these five items, hereby named the BITE, are in Table 1. There was one dominant dimension with a very large first to second eigenvalue ratio, 18.14. The parallel analyses and MAP test both indicated one factor. The intercorrelations between the five BITE items were .53 to .69. Cronbach’s alpha was .88.

GRM analyses were then conducted on the five items. The test and item information functions appear in Figures 1 and 4. The threshold and discrimination parameters appear in Table 2. The test information function for the BITE resembles the test information functions for the IRQ and BSIS, although it is slightly wavy because it is based on just five items. It would take an additional 10 or so items to smooth the function, but doing so would involve adding redundant and/or less powerful items from the existing pools. The time involved in completing a much longer test would not be worth the only mild improvements in test discrimination. The item with the best information function was “I have been feeling irritable.” As with the IRQ and BSIS, the BITE test information function tails off at the lower end of the latent trait continuum. However, the plot in Figure 1 suggests that the tailing-off is not serious.

The item response category characteristic curves, or response option curves for short, were examined for all five items of the BITE. Such curves can reveal response options that are redundant or not used often or properly by respondents. As an example, the response option curves for the “I have been feeling irritable” item appear in Figure 5. Each response option for this item was properly used. Option 6 (*always*) was used less frequently, but it should be retained for those few participants with very high levels of irritability. The curves for the

other BITE items sometimes varied slightly from those in Figure 5, but they indicated no serious problems.

Bivariate Analyses: Preliminary Tests of Convergent and Concurrent Validity

The Pearson correlations between scores on the BITE, IRQ, and BSIS ranged from .83 to .86 (Table 3). The correlations between the three measures of irritability and related constructs (e.g., anger, aggression) are also provided in Table 3. There were no meaningful differences between the three irritability measures in how they correlated with these other measures.

As a preliminary test of concurrent validity, Pearson correlations between the BITE and measures of life satisfaction, perceived social support, pain severity, and pain interference were calculated. All were significant at $p < .001$. The Pearson correlation between the latent trait and raw total BITE scores was .98.

Differential Item and Test Functioning

The noncompensatory DIF coefficients for the BITE are reported in Table 2. For gender, none of the five DIF coefficients were significant at the .05 level. Figure 6 displays the expected scores across the levels of the latent trait for the two items with the largest DIF. Levels of item bias were clearly small. At the test level, the DTF coefficient for gender was .0056 and significant, $p < .006$. However, the rDTF coefficient, .0746, which represents the differences between the test characteristic curves expressed in the metric of the observed scores, indicated very weak test bias due to gender. Males scored slightly, but consistently, higher than females (.0746 points higher on a 5- to 30-point scale) across the latent trait continuum. The plot of the expected test scores for males and females across levels of irritability displays the magnitude of test bias (Figure 6). Lines for males and females are virtually superimposed and indiscernible.

Similar findings emerged when DFIT analyses for the BITE were conducted for the pain versus student sample groups. Once again, none of the five DIF coefficients were significant at the .05 level (Table 2). The expected scores for the two items with the largest DIF are displayed in Figure 7. The levels of item bias were small. At the test level, the DTF coefficient for sample type, .0057 was significant, $p < .005$, but the rDTF value, .0754, was very small. The pain patients scored only .0754 points higher on a 5- to 30-point scale than did students who were at the same points on the irritability trait continuum. A plot of the expected test scores for the pain patient and student samples across levels of irritability also appears in Figure 7.

The same DFIT analyses examining gender bias were conducted for the IRQ and BSIS. Eleven of the 21 IRQ items displayed statistically significant gender bias. The corresponding rDTF value was 7.22 ($p < .001$), with females generally obtaining higher tests scores (7.22 points on a 21- to 126-point scale) than males who were at the same points of the latent trait continuum. None of the 14 BSIS items displayed statistically significant gender bias and the rDTF value for gender was 1.43 (1.43 points on a 14- to 84-point scale) and not significant.

Group Means on the BITe

The means on the BITe for the entire sample were 2.60 ($SD = 0.93$, range 1–5; item mean scoring) and 12.95 ($SD = 4.67$, range 6–30; item summation scoring). The item summation means for males and females were not significantly different (12.69 vs. 13.13), $t(1057) = 1.48$, $p = .14$, Cohen's $d = .10$. The rDTF value for gender reported above indicated that the observed difference between the means for males and females ($12.69 - 13.13 = -0.44$) is underestimated by .0746, and should thus be 0.51. This revised mean difference was also nonsignificant, $t(1057) = 1.73$, $p = .09$, and the effect size was small, Cohen's $d = .11$.

In contrast, means on the BITe for the student and pain patient samples were significantly different, $t(1097) = 5.73$, $p < .001$, Cohen's $d = .43$. The item summation means were 12.55 and 14.61, respectively. The rDTF value for sample type reported above indicates that the observed difference between the means for students and pain patients ($14.61 - 12.55 = 2.06$) is overestimated by .0754, and should thus be 1.98. This revised mean difference remained significant, $t(1097) = 5.52$, $p < .001$, Cohen's $d = .42$.

Discussion

In recent years, there has been a surge of interest among the psychiatric and medical communities regarding the specific causes and consequences of irritability. Yet research has been hindered by a heavy reliance on single-item measures to assess irritability, and limitations in existing multi-item scales. Using IRT methods, we developed the BITe, a five-item self-report measure of irritability that displays negligible bias with respect to gender. Despite its brevity, the BITe showed an excellent internal structure in this study. Cronbach's alpha was high (.88) and the scale demonstrated a strong ability to provide a meaningful rank ordering among respondents. All five BITe items are highly face valid and display minimal conceptual overlap with related constructs, including depression, anger, and hostility.

The extent to which the BITe offers advantages over existing scales is an important question. We directly compared our measure with two of the strongest and most recently published irritability scales. Compared with the 21-item IRQ (Craig et al., 2008) and the 14-item BSIS (Born et al., 2008), the five-item BITe offers the obvious advantage of being shorter. GRM analyses indicated that including any (or all) of the eight best items from the IRQ and BSIS did *not* result in any meaningful improvements in the test information function of the BITe. In fact, the additional items were either redundant or weaker than those already included in our scale, and several of those items tapped more strongly into anger, rather than irritability. All three scales demonstrated high internal consistency (Cronbach's alphas .88). Since the BITe contains face valid items that show minimal overlap with related constructs, it may allow researchers and clinicians to measure irritability with greater specificity. This greater specificity may be particularly important when working with patients who have prominent irritability in the context of comorbid mental or physical health problems (e.g., nicotine withdrawal, depression, traumatic brain injury), and may assist in better understanding the unique (and nonunique) causes and treatments of irritability.

This is also the first study to explore gender bias in existing measures of irritability and to purposefully develop a measure of irritability that is suitable for use among both men and women. Based on IRT analyses, there was no evidence of gender bias on any of the five BITE items and only negligible gender bias at the test score level. Men scored, on average, .0746 points higher on the 5- to 30-point scale than did women who were at the same points on the irritability trait continuum. Thus, scores on the BITE have a virtually identical meaning for men and women. This represents a major strength of our scale and stands in contrast to the IRQ, which displayed significant gender bias on 11 of the 21 items, and significant test level bias (women scored 7.22 points higher on the 21- to 126-point scale), resulting in an overestimation of irritability in females. The BSIS, which was designed specifically to assess irritability among women with female-specific mood disturbances, did *not* demonstrate any item or test bias with respect to gender. Since men and women with similar levels of irritability will likely respond to BSIS items in the same way, it may have broader applicability than was originally intended.

Several important findings also emerged regarding the nature of irritability itself. First, male and female participants had statistically equivalent levels of irritability. Thus, previous inconsistent findings regarding gender differences in irritability may be an artifact of test bias, but may also depend on the specific population of interest (Perlis et al., 2009; Piazzini et al., 2011).

Second, we found evidence for a single dominant factor of irritability in our sample of 1,116 participants using multiple statistical methods and testing across four item pools. The dominant factor evidence far surpassed the baseline required for IRT analyses. We then searched for the existence of smaller item pools that might suggest the importance of subscales, but the evidence was scattered and weak. Parallel analyses and the MAP test indicated more than one dimension in most of the item pools, but the GFI coefficients for one-dimension models were very high (greater than .97). The increments in GFI values for additional dimensions were negligible. The bifactor analyses revealed that item loadings on additional group factors were weaker than the loadings the same items had on the dominant factor. The items were thus more discriminating measures of the dominant factor than they were of their group factors. There were also no collections of items suggesting independent dimensions or subscales. The additional dimensions thus appeared to be either minor, nuisance factors that detract from the measurement of irritability, or “bloated specifics” caused by repetitive semantic similarities that exist within subsets of items. Our dimensionality findings contradict previous suggestions that irritability is a multidimensional construct (Craig et al., 2008).

Third, and not surprisingly, we found a high degree of correlation between measures of irritability, depressive symptoms, anger, hostility, and neuroticism (r s between .24 and .69). In our qualitative work on irritability, participants clearly distinguished irritability from other negative constructs, but often used those constructs to situate their own irritability (e.g., irritability leading to anger). This is consistent with work by Watson et al. (2007), who have pointed out that even the constructs of depression and anxiety cannot be “neatly separated” and that there is still much to be learned about the complex and overlapping structure of negative affective states. Last, our findings add to a growing body of literature regarding the

predictive validity of irritability measures. Scores on the BITE demonstrated significant correlations with perceived support, satisfaction with life, as well as pain severity and pain interference. Although we cannot infer direction of causality from our cross-sectional data, we suspect that these associations are bidirectional. Prospective studies with large samples are needed to determine whether the BITE can predict outcomes above and beyond measures of related constructs, such as the State-Trait Anger Expression Inventory for anger and Beck Depression Inventory for depression, and whether the BITE does a better job of predicting outcomes compared with other existing irritability scales. The incremental validity of the BITE will most likely be strongest when used in the context of patient populations known to suffer from prominent symptoms of irritability, but this remains to be explored.

Despite the strengths in our approach to developing the BITE, and in the psychometric properties of the measure itself, there are a number of limitations that deserve mentioning. We used all the items from two existing measures, plus 29 additional items of our own creation and did not find additional dimensions of irritability. However, it is possible that further research using homogeneous item composites might reveal stronger evidence for two or more dimensions. Another limitation is that our findings were based on cross-sectional self-report data. As a result, we were not able to calculate test-retest reliabilities for our scale, and we were not able to tease apart the temporal and causal relationships between irritability, related constructs, and well-being. The reliance on self-report data also means that we were not able to determine how well participants' scores on the BITE correlated with others' (e.g., significant others) reports. Since an individual may experience irritability without any outward signs or symptoms (Snaith & Taylor, 1985), we expect that other-rater scales of irritability would only be modestly related with self-report scales. Several other-rater scales have been published, but here again; many items reflect anger and aggression and not irritability (e.g., "Does he/she yell a lot," "He/she has been so enraged that he/she has hit someone"; Burns, Folstein, Brandt, & Folstein, 1990; Chatterjee, Anderson, Moskowitz, Hauser, & Marder, 2005; Craig et al., 2008). Another limitation of the BITE is that it does only a moderate job of discriminating between individuals at the lower end of the trait continuum (a limitation shared with other measures of psychological distress). Last, this scale may have limited utility in adult populations with high levels of irritability, but advanced cognitive impairment (e.g., later stage dementia and Huntington's disease).

Conducting a comprehensive evaluation of the reliability, validity, and utility of the BITE will be an unfolding process and the incremental validity of the BITE requires specific attention. Based on our previous work and others (Born & Steiner, 1999; Craig et al., 2008; Piazzini et al., 2011), we chose to conceptualize irritability as a state. However, the construct of irritability likely has both state and trait properties, similar to anger and anxiety (Spielberger, 1999). Future research is needed to examine the short- and long-term test-retest reliability of our scale, and to examine the extent to which levels of irritability remain stable over the lifespan. Scale instructions for the BITE can be easily modified to accommodate these types of questions (i.e., how irritable respondents "generally" feel vs. how irritable they have felt recently). The sample we used to develop the BITE was also used to evaluate its psychometric properties. Further research is needed to determine whether the strong psychometric properties displayed in our student and pain samples can be replicated in other populations (e.g., in those that are more ethnically diverse). Last, further research is

needed to determine the appropriateness of our scale for use among early and mid-adolescents, who are known to have elevated levels of irritability, but may not yet have the insight or capacity to self-report irritability. The strong psychometrics of our measure in our sample of predominantly 19- and 20-year-olds is promising in this regard.

In sum, the BITE is a brief self-report scale of irritability that demonstrated strong reliability and validity in this initial study. We found preliminary evidence that the BITE can be used successfully among both men and women, and in both healthy and clinical (pain) populations. Interestingly, based on a series of IRT analyses of more than 1,000 participants, we found that the construct of irritability can be effectively captured using only five items. This simple, easy-to-use tool may hold appeal to clinicians and researchers who are interested in measuring irritability, but who do not want to create further burden on patients and participants. Since irritability is not socially desirable, the inclusion of a brief, general measure of irritability within a broader clinical assessment may help encourage communication about irritability and pave the way for intervention. Clearly, there is much to be learned about the causes, consequences, and treatment of irritability. The BITE shows promise in helping to advance this burgeoning field of research.

Acknowledgments

The authors would also like to gratefully acknowledge Dr. Stephen Jefferys, Tom Beggs, Kyle Dyck, Raquel Graham, Lisa Landis, and Kara Turcotte for their assistance with participant recruitment and data collection.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a Catalyst Grant from the Canadian Institutes of Health Research–Institute of Gender and Health (FRN-103251).

References

- Barata, P., Holtzman, S., Cunningham, S. Distinguishing irritability from other negative emotional states. Poster presented at the Annual Meeting of the Canadian Psychological Association; Halifax, Nova Scotia, Canada. 2012 Jun. Retrieved from <http://www.cpa.ca/docs/File/convention/2012/abstracts/>
- Beck, AT., Steer, RA., Brown, GK. BDI-II manual. San Antonio, TX: Psychological Corporation; 1996.
- Beedie CJ, Terry PC, Lane AM. Distinctions between emotion and mood. *Cognition and Emotion*. 2005; 19:847–878. DOI: 10.1080/02699930541000057
- Born L, Koren G, Lin E, Steiner M. A new, female-specific irritability rating scale. *Journal of Psychiatry and Neuroscience*. 2008; 33:344–354. [PubMed: 18592028]
- Born L, Steiner MM. Irritability: The forgotten dimension of female-specific mood disorders. *Archives of Women's Mental Health*. 1999; 2:153–167. DOI: 10.1007/s007370050044\
- Bowling A. Just one question: If one question works, why ask several? *Journal of Epidemiology and Community Health*. 2005; 59:342–345. DOI: 10.1136/jech.2004.021204 [PubMed: 15831678]
- Braun V, Clarke V. Using thematic analysis in psychology. *Qualitative Research in Psychology*. 2006; 3:77–101. DOI: 10.1191/1478088706qp063oa
- Burns A, Folstein S, Brandt J, Folstein M. Clinical assessment of irritability, aggression, and apathy in Huntington and Alzheimer disease. *Journal of Nervous and Mental Disease*. 1990; 178:20–26. DOI: 10.1097/00005053-199001000-00004 [PubMed: 2136908]

- Burns JW, Quartana PJ, Bruehl S. Anger inhibition and pain: Conceptualizations, evidence and new directions. *Journal of Behavioral Medicine*. 2008; 31:259–279. DOI: 10.1007/s10865-008-9154-7 [PubMed: 18498056]
- Buss AH, Durkee A. An inventory for assessing different kinds of hostility. *Journal of Consulting Psychology*. 1957; 21:343–349. DOI: 10.1037/h0046900 [PubMed: 13463189]
- Buss AH, Perry M. The Aggression Questionnaire. *Journal of Personality and Social Psychology*. 1992; 63:452–459. DOI: 10.1037/0022-3514.63.3.452 [PubMed: 1403624]
- Butcher, JN., Graham, JR., Ben-Porath, YS., Tellegen, A., Dahlstrom, WG., Kaemmer, B. MMPI-2 (Minnesota Multiphasic Personality Inventory–2): Manual for administration, scoring, and interpretation. Minneapolis: University of Minnesota Press; 2001.
- Caplan PJ. The debate about PMDD and Sarafem: Suggestions for therapists. *Women & Therapy*. 2004; 27:55–67. DOI: 10.1300/J015v27n03_05
- Caprara G, Cinanni V, D'Imperio G, Passerini S, Renzi P, Travaglia G. Indicators of impulsive aggression: Present status of research on irritability and emotional susceptibility scales. *Personality and Individual Differences*. 1985; 6:665–674. DOI: 10.1016/0191-8869(85)90077-7
- Chatterjee A, Anderson KE, Moskowitz CB, Hauser WA, Marder KS. A comparison of self-report and caregiver assessment of depression, apathy, and irritability in Huntington's Disease. *Journal of Neuropsychiatry and Clinical Neurosciences*. 2005; 17:378–383. DOI: 10.1176/appi.neuropsych.17.3.378 [PubMed: 16179661]
- Cleeland, CS. Pain assessment in cancer. In: Osoba, D., editor. *Effect of cancer on quality of life*. Boca, Raton, FL: CRC Press; 1991. p. 293-305.
- Cohen S, Hoberman HM. Positive events and social supports as buffers of life change stress. *Journal of Applied Social Psychology*. 1983; 13:99–125. DOI: 10.1111/j.1559-1816.1983.tb02325.x
- Costa PT, McCrae RR. Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment*. 1995; 64:21–50. DOI: 10.1207/s15327752jpa6401_2 [PubMed: 16367732]
- Craig KJ, Hietanen H, Markova IS, Berrios GE. The Irritability Questionnaire: A new scale for the measurement of irritability. *Psychiatry Research*. 2008; 159:367–758. DOI: 10.1016/j.psychres.2007.03.002 [PubMed: 18374422]
- Diener E, Emmons RA, Larsen RJ, Griffin S. The Satisfaction With Life Scale. *Journal of Personality Assessment*. 1985; 49:71–75. DOI: 10.1207/s15327752jpa4901_13 [PubMed: 16367493]
- DiGiuseppe, R., Tafrate, RC. *Understanding anger disorders*. New York, NY: Oxford University Press; 2007.
- Embretson, SE., Reise, SP. *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum; 2000.
- Fava M, Hwang I, Rush AJ, Sampson N, Walters EE, Kessler RC. The importance of irritability as a symptom of major depressive disorder: Results from the National Comorbidity Survey Replication. *Molecular Psychiatry*. 2010; 14:1–12. DOI: 10.1038/mp.2009.20
- Flowers CP, Oshima TC, Raju NS. A description and demonstration of the polytomous-DFIT framework. *Applied Psychological Measurement*. 1999; 23:309–326. DOI: 10.1177/01466219922031437
- Harris CA, D'Leon JL. Psychometric properties of the Beck Depression Inventory-Second Edition (BDI-II) in individuals with chronic pain. *Pain*. 2008; 137:609–622. DOI: 10.1016/j.pain.2007.10.022 [PubMed: 18079063]
- Harris J. A further evaluation of the Aggression Questionnaire: Issues of validity and reliability. *Behavior Research and Therapy*. 1997; 35:1047–1053. DOI: 10.1016/S0005-7967(97)00064-8
- Jensen MP, Karoly P, Braver S. The measurement of clinical pain intensity: A comparison of six methods. *Pain*. 1986; 27:117–126. [PubMed: 3785962]
- Liu, C. Multilevel analysis. In: Lewis-Beck, M.Bryman, AE., Liao, TF., editors. *The SAGE encyclopedia of social science research methods*. Vol. 2. Thousand Oaks, CA: Sage; 2003. p. 673
- Lormand E. Toward a theory of moods. *Philosophical Studies*. 1985; 47:385–407.
- MacEwen K, Barling J. Type A behavior and marital satisfaction: Differential effects of achievement striving and impatience/irritability. *Journal of Marriage and the Family*. 1993; 55:1001–1010. DOI: 10.2307/352779

- Mangelli L, Fava GA, Grassi L, Ottolini F, Paolini S, Porcelli P, Sonino N. Irritable mood in Italian patients with medical disease. *Journal of Nervous and Mental Disease*. 2006; 194:226–228. DOI: 10.1097/01.nmd.0000202511.21925.a2 [PubMed: 16534442]
- Marcus SM, Kerber KB, Rush A, Wisniewski SR, Nierenberg A, Balasubramani GK, Trivedi MH. Sex differences in depression symptoms in treatment-seeking adults: Confirmatory analyses from the Sequenced Treatment Alternatives to Relieve Depression study. *Comprehensive Psychiatry*. 2008; 49:238–246. DOI: 10.1016/j.comp-psych.2007.06.012 [PubMed: 18396182]
- McWilliams LA, Cox BJ, Enns MW. Mood and anxiety disorders associated with chronic pain: An examination in a nationally representative sample. *Pain*. 2003; 106:127–133. DOI: 10.1016/S0304-3959(03)00301-4 [PubMed: 14581119]
- Morey, LC. *The Personality Assessment Inventory professional manual*. Odessa, FL: Psychological Assessment Resources; 1991.
- Morizot, J., Ainsworth, AT., Reise, SP. Toward modern psychometrics: Application of item response theory models in personality research. In: Robins, RW, Fraley, RC., Krueger, RF., editors. *Handbook of research methods in personality psychology*. New York, NY: Guilford Press; 2007. p. 407–423.
- Oshima TC, Morris SB. Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice*. 2008; 27:43–50. DOI: 10.1111/j.1745-3992.2008.00127.x
- Oshima TC, Raju NS, Nanda AO. A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement*. 2006; 43:1–17. DOI: 10.1111/j.1745-3984.2006.00001.x
- Pavot W, Diener E. Review of the Satisfaction With Life Scale. *Psychological Assessment*. 1993; 5:164–172. DOI: 10.1037/1040-3590.5.2.164
- Perlis RH, Fava M, Trivedi MH, Alpert J, Luther JF, Wisniewski SR, Rush AJ. Irritability is associated with anxiety and greater severity, but not bipolar spectrum features, in major depressive disorder. *Acta Psychiatrica Scandinavica*. 2009; 119:282–289. DOI: 10.1111/j.1600-0447.2008.01298.x [PubMed: 19207123]
- Piazzini A, Turner K, Edefonti V, Bravi F, Canevini M, Ferraroni M. A new Italian instrument for the assessment of irritability in patients with epilepsy. *Epilepsy & Behavior*. 2011; 21:275–281. DOI: 10.1016/j.yebeh.2011.04.051 [PubMed: 21624849]
- Pickles A, Aglan A, Collishaw S, Messer J, Rutter M, Maughan B. Predictors of suicidality across the life span: The Isle of Wight study. *Psychological Medicine*. 2010; 40:1453–1466. DOI: 10.1017/S0033291709991905 [PubMed: 19939326]
- Reise SP, Moore TM, Haviland MG. Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*. 2010; 92:544–559. DOI: 10.1080/00223891.2010.496477 [PubMed: 20954056]
- Rizopoulos D. ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*. 2006; 17:1–25.
- Russo S, Kema IP, Haagsma EB, Boon JC, Willemse PB, den Boer JA, Korf J. Irritability rather than depression during interferon treatment is linked to increased tryptophan catabolism. *Psychosomatic Medicine*. 2005; 67:773–777. DOI: 10.1097/01.psy.0000171193.28044.d8 [PubMed: 16204437]
- Rutz W, von Knorring L, Pihlgren H, Rihmer Z, Walinder J. Prevention of male suicides: Lessons from Gotland study. *Lancet*. 1995; 345:524. doi: 10.1016/S0140-6736(95)90622-3
- Sahl JC, Cohen LH, Dasch KB. Hostility, interpersonal competence, and daily dependent stress: A daily model of stress generation. *Cognitive Therapy and Research*. 2009; 33:199–210. DOI: 10.1007/s10608-007-9175-5
- Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*. 1969; 34:100–114.
- Siemer M. Mood experience: Implications of a dispositional theory of moods. *Emotion Review*. 2009; 1:256–263. DOI: 10.1177/1754073909103594
- Smith LL, Reise SP. Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction scale. *Journal of*

- Personality and Social Psychology. 1998; 75:1350–1362. DOI: 10.1037/0022-3514.75.5.135 [PubMed: 9866192]
- Snaith RP, Constantopoulos RP, Jardine MY, McGuffin P. A clinical scale for the self-assessment of irritability. *British Journal of Psychiatry*. 1978; 132:164–171. DOI: 10.1192/bjp.132.2.164 [PubMed: 623950]
- Snaith RP, Taylor CM. Irritability: Definition, assessment and associated factors. *British Journal of Psychiatry*. 1985; 147:127–136. DOI: 10.1192/bjp.147.2.127 [PubMed: 3840045]
- Sofaer BB, Walker J. Mood assessment in chronic pain patients. *Disability and Rehabilitation: An International, Multidisciplinary Journal*. 1994; 16:35–38. DOI: 10.3109/0-9638289409166434
- Spielberger, CD. *The State-Trait Anger Expression Inventory–2*. Lutz, FL: Psychological Assessment Resources; 1999.
- Spielberger CD, Reheiser EC. Assessment of emotions: Anxiety, anger, depression, and curiosity. *Applied Psychology: Health and Well-Being*. 2009; 1:271–302. DOI: 10.1111/j.1758-0854.2009.01017.x
- Steer R, Clark D. Psychometric characteristics of the Beck Depression Inventory-II with college students. *Measurement and Evaluation in Counseling and Development*. 1997; 30:128–136. DOI: 10.1037/t02025-000
- Streiner DL. Measure for measure: New developments in measurement and item response theory. *Canadian Journal of Psychiatry*. 2010; 55:180–186. [PubMed: 20370969]
- Stringaris A. Irritability in children and adolescents: A challenge for DSM-5. *European Child & Adolescent Psychiatry*. 2011; 20:61–66. DOI: 10.1007/s00787-010-0150-4 [PubMed: 21298306]
- Stringaris A, Cohen P, Pine DS, Leibenluft E. Adult outcomes of youth irritability: A 20-year prospective community-based study. *American Journal of Psychiatry*. 2009; 166:1048–1054. DOI: 10.1176/appi.ajp.2009.08121849 [PubMed: 19570932]
- Stringaris A, Goodman R, Ferdinando S, Razdan V, Muhrer E, Leibenluft E, Brotman MA. The Affective Reactivity Index: A concise irritability scale for clinical research settings. *Journal of Child Psychology and Psychiatry*. 2012; 53:1109–1117. DOI: 10.1111/j.1469-7610.2012.02561.x [PubMed: 22574736]
- Thalmayer A, Saucier G, Eigenhuis A. Comparative validity of Brief to Medium-Length Big Five and Big Six Personality Questionnaires. *Psychological Assessment*. 2011; 23:995–1009. DOI: 10.1037/a0024165 [PubMed: 21859221]
- Verhoeven FE, Booij L, Van der Wee NJ, Penninx BW, Van der Does AJ. Clinical and physiological correlates of irritability in depression: Results from the Netherlands study of depression and anxiety. *Depression Research and Treatment*. 2011; 2011:1–9. DOI: 10.1155/2011/126895
- Warren RE, Deary IJ, Frier BM. The symptoms of hyperglycaemia in people with insulin-treated diabetes: Classification using principal components analysis. *Diabetes/Metabolism Research and Reviews*. 2003; 19:408–414. DOI: 10.1002/dmrr.396 [PubMed: 12951649]
- Watson D, O'Hara MW, Simms LJ, Kotov R, Chmielewski M, McDade-Montez EA, Stuart S. Development and validation of the Inventory of Depression and Anxiety Symptoms (IDAS). *Psychological Assessment*. 2007; 19:253–268. DOI: 10.1037/1040-3590.19.3.253 [PubMed: 17845118]
- Youn J, Lee D, Jhoo J, Kim K, Choo I, Woo J. Prevalence of neuropsychiatric syndromes in Alzheimer's disease (AD). *Archives of Gerontology and Geriatrics*. 2011; 52:258–263. DOI: 10.1016/j.archger.2010.04.015 [PubMed: 20537736]
- Zwick WR, Velicer WF. Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*. 1986; 99:432–442. DOI: 10.1037/0033-2909.99.3.432

Figure 1. Test information functions for the (A) Irritability Questionnaire (Craig, Hietanen, Markova, & Berrios, 2008), (B) Born-Steiner Irritability Scale (Born, Koren, Lin, & Steiner, 2008), and (C) Brief Irritability Test.

Figure 2. Irritability Questionnaire (Craig, Hietanen, Markova, & Berrios, 2008) item information functions.

Note. The best four items were as follows: “I have felt frustrated” (9), “I have felt bitter about things” (14), “When I have looked back on how life has treated me, I’ve felt a bit disappointed and angry” (16), and “I have been feeling like a bomb, ready to explode” (18).

Figure 3.

Born–Steiner Irritability Scale (Born, Koren, Lin, & Steiner, 2008) item information functions.

Note. The best four items were as follows: “I have been feeling mad” (1), “I have been feeling ready to explode” (2), “I have been easily flying off the handle” (5), and “It feels like there has been a cloud of anger over me” (6).

Figure 4.

Brief Irritability Test item information functions.

Note. The five items are as follows: “I have been grumpy” (1), “I have been feeling like I might snap” (2), “Other people have been getting on my nerves” (3), “Things have been bothering me more than they normally do” (4), and “I have been feeling irritable” (5).

Figure 5.
Response option curves for the Brief Irritability Test item “I have been feeling irritable.”

Figure 6.

Expected Brief Irritability Test (BITe) item and test scores for males and females.

Note. Graphs for the two BITe items with the largest differential item functioning are presented here (solid line = females; dashed line = males).

Figure 7.

Expected Brief Irritability Test (BITE) item and test scores for the pain and student samples.

Note. Graphs for the two BITE items with the largest differential item functioning are presented here (solid line = students; dashed line = pain patients).

Irritability Item Pools: Eigenvalues, Numbers of Components, Goodness of Fit, H Scalability Coefficients, and Cronbach's Alphas.

Table 1

Measure	First eigen value	Second eigen value	Ratio	No. of factors			GFI	α
				Parallel analysis	MAP test			
IRQ (21 items)	7.49	1.07	7.00	4	2		.97	.90
BSIS (14 items)	7.45	0.99	7.46	2	2		.98	.91
Additional items generated for this study (28 items)	13.38	1.27	10.53	2	3		.99	.95
All irritability items combined	26.45	2.53	10.47	5	9		.98	.97
BITe (5 items)	3.36	0.19	18.14	1	1		.99	.88

Note. MAP = minimum average partial; GFI = goodness-of-fit index; IRQ = Irritability Questionnaire (Craig, Hietanen, Markova, & Berrios, 2008); BSIS = Born-Steiner Irritability Scale (Born, Koren, Lin, & Steiner, 2008); BITe = Brief Irritability Test.

Item Response Theory Results: Response Option Thresholds (*b*), Discrimination Parameters (*a*), and Differential Item Functioning (DIF) Coefficients for Gender and Sample Type for the Brief Irritability Test (BITe) Items.

Table 2

	<i>b</i> 1	<i>b</i> 2	<i>b</i> 3	<i>b</i> 4	<i>b</i> 5	<i>a</i>	DIF-Gender	DIF-Sample type
1. I have been grumpy	−1.80	−0.40	1.03	1.81	2.93	2.23	.009	.002
2. I have been feeling like I might snap	−0.41	0.55	1.52	2.20	3.01	2.25	.008	.021
3. Other people have been getting on my nerves	−1.36	−0.03	1.27	2.11	3.18	2.15	.006	.056
4. Things have been bothering me more than they normally do	−1.00	0.07	0.89	1.70	2.58	2.74	.001	.008
5. I have been feeling irritable	−1.41	−0.22	0.82	1.56	2.36	3.31	.015	.058

* *p* < .05.

Table 3

Pearson Correlations Among Three Measures of Irritability and Related Constructs.

	BITe	IRQ	BSIS
IRQ	.802	1.000	.856
BSIS	.864	.856	1.000
STAXI-2 (Trait Anger)	.509	.560	.572
AQ (Anger)	.559	.591	.607
AQ (Verbal Aggression)	.245	.327	.301
AQ (Physical Aggression)	.254	.243	.251
AQ (Hostility)	.522	.582	.518
BDI-II (Depressive Symptoms)	.668	.690	.674
NEO-FFI (Neuroticism)	.577	.658	.586
BPI–Pain Intensity	.336	.325	.291
BPI–Pain Interference	.478	.556	.493
ISEL	–.321	–.347	–.345
SWLS	–.455	–.553	–.470

Note. BITe = Brief Irritability Test; IRQ = Irritability Questionnaire; BSIS = Born–Steiner Irritability Scale; STAXI-2 = State Trait Anger Expression Inventory–2; AQ = Aggression Questionnaire; BDI-II = Beck Depression Inventory–II; NEO-FFI = NEO Five-Factor Inventory; BPI = Brief Pain Inventory; ISEL = Interpersonal Support Evaluation List; SWLS = Satisfaction With Life Scale. All correlations $p < .001$.