Journal of Affective Disorders 212 (2017) 38–45

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

# Test-retest reliability and validity of a frustration paradigm and irritability measures

Wan-Ling Tseng[a,*], Elizabeth Moroney[a], Laura Machlin[a], Roxann Roberson-Nay[b], John M. Hettema[b], Dever Carney[b], Joel Stoddard[a], Kenneth A. Towbin[a], Daniel S. Pine[a], Ellen Leibenluft[a], Melissa A. Brotman[a]

[a] Emotion and Development Branch, National Institute of Mental Health, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, USA
[b] Department of Psychiatry and the Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA

ARTICLE INFO

*Keywords:*
Test-retest reliability
Validity
Irritability
Frustration
Parent-child agreement

ABSTRACT

*Background:* Data on the reliability and validity of assessments for irritability, particularly behavioral paradigms, are limited. This study examined the test-retest reliability and validity of a frustration paradigm (the Affective Posner 2 task) and two irritability measures [the Affective Reactivity Index (ARI) and Child Behavior Checklist (CBCL) irritability].

*Methods:* Participants were 109 youth from a general population sample of twins (aged 9–14 years). Participants completed two visits that were 2–4 weeks apart. At both visits, participants completed the Affective Posner 2 task and self-reported their irritability using the ARI. Parents reported their child's irritability using the ARI and completed the CBCL.

*Results:* The Affective Posner 2 task demonstrated good test-retest reliability, with intraclass correlations (ICCs) ranging from .44 to .78. The task effectively evoked negative affect (frustration and unhappiness) at both test and retest, demonstrating its construct validity. Moreover, self-rated frustration and unhappiness during the frustration components of the task correlated positively with self-reported but not parent-reported irritability, providing modest support for convergent validity. Parent- and child-reports of the ARI and parent-reports of the CBCL irritability measure showed excellent test-retest reliability, with ICCs ranging from .88 to .90.

*Limitations:* The sample consists of mostly twins aged 9–14 years from the communities. Thus, results may not generalize to non-twin samples or clinical samples outside of this age range.

*Conclusions:* The Affective Posner 2 paradigm and the ARI and CBCL irritability scales may be useful tools for longitudinal or treatment research on irritability.

## 1. Introduction

Recently, childhood irritability has received increased scientific attention (Leibenluft, 2011; Stringaris, 2011), due in part to the addition of disruptive mood dysregulation disorder (DMDD) to the Diagnostic and Statistical Manual of Mental Disorders, fifth edition (DSM-5; American Psychiatric Association, 2013). Children with irritability often have a low threshold for frustration, defined as the emotional and behavioral responses to blocked goal attainment (Leibenluft, 2011). Paradigms that evoke frustration thus provide an effective means for studying irritability in a laboratory setting (Leibenluft and Stoddard, 2013). Indeed, irritability has been linked with frustration that arises when a goal or reward is blocked (Berkowitz, 1989), and it aligns with the frustrative nonreward construct in the National Institute of Mental Health Research Domain Criteria (RDoC; Insel et al., 2010) matrix. However, data on the reliability and validity of behavioral paradigms that probe frustration are limited. This study is the first to examine test-retest reliability and validity of a behavioral paradigm used to elicit frustration (the Affective Posner 2 task). This study also reports test-retest reliability and validity of two commonly used clinical rating scales for childhood irritability [the Affective Reactivity Index (ARI) and the Child Behavior Checklist (CBCL) irritability].

In the past few years, studies on irritability have employed

* Correspondence to: Section on Bipolar Spectrum Disorders, Emotion and Development Branch, National Institute of Mental Health, National Institutes of Health, Building 15, Room 208, 15K North Drive, Bethesda, MD 20892, USA.
*E-mail address:* wan-ling.tseng@nih.gov (W.-L. Tseng).

http://dx.doi.org/10.1016/j.jad.2017.01.024
Received 17 October 2016; Accepted 22 January 2017
Available online 23 January 2017
0165-0327/ 

behavioral paradigms that induce frustration by withholding an expected reward through either increased task difficulty or rigged information (Deveney et al., 2013; Perlman et al., 2015). These studies demonstrate that frustration can be elicited in the laboratory. For example, in Deveney et al.'s study (2013), a previous version of the frustration paradigm described in this study was used. Deveney et al. (2013) found that children with severe irritability reported more frustration than healthy children during the frustration condition, suggesting discriminant validity of such paradigm. However, the test-retest reliability of subjects' behavior on frustration paradigms is unknown. Establishing the longitudinal reliability (e.g., test-retest reliability) of an assessment technique is important for psychological and psychiatric research. In particular, good test-retest reliability is essential for longitudinal studies or treatment trials, i.e., studies in which the research focus is changes that occur with development or in response to treatment.

In addition to behavioral paradigms, studies are increasingly using scales designed to assess childhood irritability (e.g., Roberson-Nay et al., 2015; Savage et al., 2015; Stoddard et al., 2014; Stringaris et al., 2012a, 2012b; Wiggins et al., 2014). Two commonly used scales are the ARI (Stringaris et al., 2012a) and three items extracted from the CBCL (see the Methods section; Aebi et al., 2013; Roberson-Nay et al., 2015; Savage et al., 2015; Wiggins et al., 2014). Studies report good internal consistency, construct validity, and discriminant validity of the ARI (DeSousa et al., 2013; Mulraney et al., 2014; Stringaris et al., 2012a), as well as good internal consistency of the CBCL irritability score (Aebi et al., 2013; Stringaris et al., 2012b; Wiggins et al., 2014). Although some test-retest reliability data are available for the ARI, they are based on small samples ($N$s≤40). For example, Mulraney et al. (2014) reported good test-retest reliability of self-reported ARI over a 1-week period in an adult sample [intraclass correlation (ICC)=.80] (Mulraney et al., 2014). Stringaris et al. (2012a) found good stability of parent-ARI ($r$=.88), but poor stability of child-ARI ($r$=.29), over a 1-year interval in a youth sample. No studies have examined the test-retest reliability of the CBCL irritability score.

Therefore, to address these limitations in the literature, we evaluated the test-retest reliability and validity of a frustration paradigm (i.e., the Affective Posner 2 task); that is, whether the task elicits frustration in youth reliably (i.e., twice over a 2–4 week period). We also examined the test-retest reliability and validity of two irritability measures (i.e., the ARI and the CBCL irritability scale) over the same period.

## 2. Methods

### 2.1. Participants

The participants were part of a general population sample of twins aged 9–14 years recruited through the Mid-Atlantic Twin Registry (Lilley and Silberg, 2013) for the Virginia Commonwealth University (VCU) Juvenile Anxiety Study (Carney et al., 2016). A total of 109 participants who completed the Affective Posner 2 task at both visits were included. The two visits were two to four weeks apart ($M$=23.15 days, $SD$=7.11 days). The mean age of the sample was 11.26 years ($SD$=1.33); the mean IQ, measured by the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999) two-subtest forms, was 110.72 ($SD$=13.87); 55% of the sample were female. A correlation table with these demographic variables and all the study variables at both visits is presented in Table 1. The sample included 48 complete twin pairs (28 monozygotic twins and 20 dizygotic twins) and 13 singletons. Participants were excluded from the study if they had an IQ less than 70, serious physical or neurological symptoms, thoughts of suicide or homicide, substance abuse, a diagnosis of autism or a severe pervasive developmental disorder, psychotropic medication use other than stimulants for attention-deficit/hyperactivity disorder (ADHD), or current or past episodes of psychosis. The study was approved by the VCU Institutional Review Board. Participants provided informed assent, and parents provided consent prior to their participation in the study. At both visits, participants completed the Affective Posner 2 task and self-reported their irritability using children's self-report of the ARI (Stringaris et al., 2012a); parents reported their child's irritability using parent-reports of the ARI and completed the CBCL (Achenbach, 1991; Achenbach and Rescorla, 2001). Participants received $25 for participation plus task winnings during the non-frustration part of the task (up to $50).

### 2.2. Affective Posner 2 task

This task was adapted from the Affective Posner task used in previous studies (Deveney et al., 2013; Rich et al., 2005, 2007, 2010). Trials consisted of (1) a fixation cross, (2) two boxes, (3) a blue cue appearing in one of the boxes, (4) a white target appearing in one of the boxes, (5) a jittered inter-stimulus interval (a blank screen), and (6) feedback (Fig. 1a). Participants were asked to identify the target's location (left or right) by button press. For 75% of trials, the white target appeared in the same box as the blue cue (valid trials); for 25% of trials, the white target appeared in the opposite location (invalid trials).

Prior to completing the task, participants underwent a 50-trial training during which they received accurate feedback on their performance but did not win or lose money. The task itself consisted of two *non-frustration* runs, a *pilot frustration* run, and two *frustration* runs (Fig. 1b). During the two non-frustration runs (Fig. 1b), participants received accurate feedback on their performance, earning $.50 for every correct response and losing $.50 for every incorrect response. During the pilot frustration run and the two frustration runs, participants were instructed that they needed to respond both correctly and quickly in order to win money. The pilot frustration run of the task (Fig. 1b) consisted of 32 trials and was used to acclimate participants to the "real" frustration runs; participants received rigged feedback on 10% of correct trials. Specifically, on trials with rigged feedback, participants were informed that they were "too slow," and lost money regardless of their actual reaction time. In the two "real" frustration runs (Fig. 1b), participants received rigged feedback on 60% of correct trials. After each run of the task, participants rated their feelings of unhappiness and frustration using 9-point Likert scales (i.e., 1="*happy*" or "*not at all frustrated*"; 9="*sad*" or "*extremely frustrated*"). Task variables [i.e., accuracy, reaction time (RT) on correct trials] and valence ratings (unhappiness, frustration) from the two non-frustration runs and two frustration runs were used in the reliability and validity analyses.

It is important to examine attention shifting in the context of frustration because the ability to orient attention away from frustrating stimuli is a critical component of emotion regulation. When frustrated and confronted with a negative event, irritable children may have difficulty disengaging from the blocked goal and the associated negative affect and shifting their attention to helpful distractors or emotion regulation strategies (Deveney et al., 2013). Indeed, previous research has shown that children with severe irritability responded more slowly to invalid trials vs. valid trials, suggesting difficulties in attention shifting (Deveney et al., 2013). Given this, we also examined response cost (i.e., validity effect) on accuracy and RT, as a measure of attention orienting/shifting. Specifically, we tested the test-retest reliability and validity of the accuracy and RT differences between invalid and valid trials (invalid – valid). Data from the pilot frustration runs were not included in the analyses.

After completing the task for the second time, participants filled out a self-report questionnaire to assess whether they were deceived. Participants were then debriefed about the use of deception in the task (i.e., they were told that the "too slow" feedback was not based on their reaction time). No participant reported marked distress due to frustration or deception.