Research paper

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

Wan-Ling Tsenga,⁎ Elizabeth Moroneya, Laura Machlina, Roxann Roberson-Nayb, John M. Hettemab, Dever Carneyb, Joel Stoddardc, Kenneth A. Towbina, Daniel S. Pinea, Ellen Leibenlufta, Melissa A. Brotmana

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...
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; Mulroney 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<40). For example, Mulroney 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] (Mulroney 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=7.71 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.
Table 1
Correlations between study variables.

<table>
<thead>
<tr>
<th>Visit 1</th>
<th>NF_ACC</th>
<th>NF_RT</th>
<th>NF_happy</th>
<th>NF_frust</th>
<th>F_ACC</th>
<th>F_RT</th>
<th>F_happy</th>
<th>F_frust</th>
<th>NF_ACC.RC</th>
<th>NF_RT.RC</th>
<th>F_ACC.RC</th>
<th>F_RT.RC</th>
<th>ARI_C</th>
<th>ARI_P</th>
<th>CBCL</th>
<th>Age</th>
<th>Gender</th>
<th>IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visit 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NF_ACC</td>
<td>0.14</td>
<td>0.00</td>
<td>-0.07</td>
<td>-0.30</td>
<td>-0.14</td>
<td>-0.1</td>
<td>0.1</td>
<td>0.00</td>
<td>-0.16</td>
<td>0.07</td>
<td>-0.25</td>
<td>-0.13</td>
<td>-0.01</td>
<td>0.07</td>
<td>0.15</td>
<td>0.08</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>NF_RT</td>
<td>0.13</td>
<td>0.05</td>
<td>-0.01</td>
<td>-0.34</td>
<td>-0.26</td>
<td>-0.23</td>
<td>-0.09</td>
<td>0.26</td>
<td>-0.19</td>
<td>0.35</td>
<td>-0.19</td>
<td>-0.10</td>
<td>-0.03</td>
<td>0.20</td>
<td>-0.21</td>
<td>0.07</td>
<td>-0.13</td>
<td></td>
</tr>
<tr>
<td>NF_frust</td>
<td>-0.18</td>
<td>-0.07</td>
<td>-0.09</td>
<td>-0.00</td>
<td>-0.04</td>
<td>-0.22</td>
<td>-0.12</td>
<td>-0.07</td>
<td>-0.12</td>
<td>0.04</td>
<td>-0.14</td>
<td>-0.12</td>
<td>-0.01</td>
<td>0.12</td>
<td>-0.03</td>
<td>0.23</td>
<td>-0.07</td>
<td></td>
</tr>
<tr>
<td>F_ACC</td>
<td>0.27</td>
<td>0.24</td>
<td>-0.03</td>
<td>-0.13</td>
<td>-0.42</td>
<td>-0.20</td>
<td>-0.11</td>
<td>0.05</td>
<td>0.63</td>
<td>-0.09</td>
<td>-0.09</td>
<td>-0.09</td>
<td>-0.10</td>
<td>-0.15</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F_RT</td>
<td>-0.18</td>
<td>0.20</td>
<td>-0.08</td>
<td>-0.05</td>
<td>0.15</td>
<td>-0.32</td>
<td>-0.28</td>
<td>0.03</td>
<td>0.24</td>
<td>0.34</td>
<td>-0.12</td>
<td>-0.15</td>
<td>0.19</td>
<td>0.33</td>
<td>-0.18</td>
<td>0.18</td>
<td>-0.18</td>
<td></td>
</tr>
<tr>
<td>F_happy</td>
<td>0.16</td>
<td>-0.04</td>
<td>-0.46</td>
<td>-0.05</td>
<td>-0.20</td>
<td>-0.83</td>
<td>-0.13</td>
<td>-0.13</td>
<td>-0.13</td>
<td>0.32</td>
<td>-0.06</td>
<td>0.08</td>
<td>0.22</td>
<td>-0.09</td>
<td>0.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F_frust</td>
<td>0.18</td>
<td>0.09</td>
<td>-0.39</td>
<td>-0.08</td>
<td>-0.25</td>
<td>-0.88</td>
<td>-0.01</td>
<td>-0.14</td>
<td>-0.23</td>
<td>-0.16</td>
<td>-0.02</td>
<td>0.07</td>
<td>0.04</td>
<td>-0.08</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NF_ACC.RC</td>
<td>0.83**</td>
<td>0.35**</td>
<td>-0.14</td>
<td>-0.07</td>
<td>0.25</td>
<td>0.06</td>
<td>0.14</td>
<td>0.15</td>
<td>-0.14</td>
<td>0.22</td>
<td>-0.23</td>
<td>-0.16</td>
<td>-0.02</td>
<td>0.07</td>
<td>0.04</td>
<td>-0.08</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>NF_RT.RC</td>
<td>-0.22</td>
<td>-0.07</td>
<td>0.04</td>
<td>0.14</td>
<td>0.24</td>
<td>-0.06</td>
<td>-0.13</td>
<td>-0.18</td>
<td>-0.05</td>
<td>0.27</td>
<td>-0.06</td>
<td>0.05</td>
<td>0.07</td>
<td>-0.28</td>
<td>0.09</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F_ACC.RC</td>
<td>0.13</td>
<td>0.36</td>
<td>-0.01</td>
<td>0.13</td>
<td>0.57</td>
<td>-0.65</td>
<td>-0.04</td>
<td>0.34</td>
<td>0.02</td>
<td>-0.06</td>
<td>-0.09</td>
<td>0.15</td>
<td>-0.23</td>
<td>0.09</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F_RT.RC</td>
<td>-0.09</td>
<td>0.20</td>
<td>0.13</td>
<td>0.09</td>
<td>0.24</td>
<td>0.15</td>
<td>-0.08</td>
<td>-0.15</td>
<td>0.17</td>
<td>-0.13</td>
<td>-0.01</td>
<td>-0.08</td>
<td>-0.30</td>
<td>0.23</td>
<td>-0.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARI_C</td>
<td>-0.08</td>
<td>-0.08</td>
<td>0.17</td>
<td>0.13</td>
<td>-0.03</td>
<td>-0.11</td>
<td>-0.33</td>
<td>-0.26</td>
<td>-0.05</td>
<td>0.19</td>
<td>-0.17</td>
<td>0.05</td>
<td>0.37</td>
<td>0.28</td>
<td>0.00</td>
<td>0.02</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>ARI_P</td>
<td>-0.05</td>
<td>-0.05</td>
<td>0.04</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.15</td>
<td>-0.08</td>
<td>-0.06</td>
<td>0.08</td>
<td>0.07</td>
<td>0.14</td>
<td>0.22</td>
<td>0.64</td>
<td>-0.21</td>
<td>-0.03</td>
<td>0.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBCL</td>
<td>0.01</td>
<td>0.03</td>
<td>0.10</td>
<td>0.17</td>
<td>-0.07</td>
<td>0.09</td>
<td>0.02</td>
<td>0.03</td>
<td>0.04</td>
<td>-0.08</td>
<td>0.17</td>
<td>0.29</td>
<td>0.54</td>
<td>0.02</td>
<td>-0.12</td>
<td>-0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.17</td>
<td>0.18</td>
<td>0.01</td>
<td>0.01</td>
<td>-0.08</td>
<td>-0.14</td>
<td>0.15</td>
<td>0.01</td>
<td>-0.20</td>
<td>-0.28</td>
<td>-0.09</td>
<td>0.01</td>
<td>-0.21</td>
<td>-0.08</td>
<td>-0.16</td>
<td>0.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>0.05</td>
<td>0.11</td>
<td>0.00</td>
<td>0.02</td>
<td>0.41**</td>
<td>-0.10</td>
<td>-0.12</td>
<td>-0.13</td>
<td>0.03</td>
<td>0.17</td>
<td>0.25</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>-0.15</td>
<td>-0.15</td>
<td>-0.06</td>
<td></td>
</tr>
<tr>
<td>IQ</td>
<td>0.14</td>
<td>0.10</td>
<td>0.12</td>
<td>0.10</td>
<td>-0.01</td>
<td>0.06</td>
<td>-0.16</td>
<td>0.18</td>
<td>0.04</td>
<td>-0.03</td>
<td>-0.10</td>
<td>0.01</td>
<td>-0.07</td>
<td>0.03</td>
<td>0.08</td>
<td>-0.06</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

Note. The within-time correlations between Visit 1 variables are above the diagonal; the within-time correlations between Visit 2 variables are below the diagonal. NF_ACC=Non-Frustration Runs Accuracy; NF_RT=Non-Frustration Runs Reaction Time; NF_frust=Non-Frustration Runs Frustration Ratings; F_ACC=Non-Frustration Runs Happy/Unhappy Ratings; F_RT=Non-Frustration Runs Reaction Time; F_happy=Non-Frustration Runs Happy/Unhappy Ratings; F_frust=Non-Frustration Runs Frustration Ratings; NF_ACC.RC=Non-Frustration Runs Accuracy Response Cost; NF_RT.RC=Non-Frustration Runs Reaction Time Response Cost; F_ACC.RC=Non-Frustration Runs Happy/Unhappy Ratings Response Cost; F_RT.RC=Non-Frustration Runs Frustration Ratings Response Cost; ARI_C=Affective Reactivity Index Child Report; ARI_P=Affective Reactivity Index Parent Report; CBCL=Child Behavior Checklist. Gender was coded as 0 for male and 1 for female.  
* p < .05.  
** p < .01.  
*** p < .001.
temper easily.

Past research has demonstrated good internal consistency (Aebi et al., 2012a). The child-reported and parent-reported ARI in this sample (DeSousa et al., 2013; Mulraney et al., 2014; Stringaris et al., 2013) included .86 for child ARI; .89 and .84 for parent ARI). The sum scores of irritability (Aebi et al., 2013; Stringaris et al., 2012b; Wiggins et al., 2014) have been demonstrated in past research. For the current sample, the internal consistency of this measure was good at both visits (α's=.84 at Visit 1 and .78 at Visit 2). The sum of the three items was used in the analyses (a possible score of 0–12).

2.3. Irritability measures

2.3.1. ARI

Children’s irritability was assessed using the parent- and child-reports of the ARI (Stringaris et al., 2012a). This is a six-item scale assessing the frequency, duration, and threshold of irritability in the past 6 months (Stringaris et al., 2012a). Each item was rated on a 3-point Likert scale from 0 (not true) to 2 (certainly true). Sample items include “get angry frequently,” “stay angry for a long time,” and “lose temper easily.” Past research has demonstrated good internal consistency (α’s=.80), good validity (i.e., construct validity and discriminant validity), and a single-factor structure in clinical and community samples (DeSousa et al., 2013; Mulraney et al., 2014; Stringaris et al., 2012a). The child-reported and parent-reported ARI in this sample also showed good internal consistency at both visits (α’s=.86 and .89 for child ARI; .89 and .84 for parent ARI). The sum scores of the six items on the child- and parent-report were used in the analyses (a possible score of 0–12).

2.3.2. CBCL

Parents completed the CBCL (Achenbach, 1991). Three items extracted from the parent-reported CBCL (“temper tantrums or hot temper,” “stubborn, sullen or irritable,” and “sudden changes in mood or feelings”) were also used to assess childhood irritability in the past 6 months. Past studies have also used scores derived from these items to assess childhood irritability (Aebi et al., 2013; Roberson-Nay et al., 2015; Stringaris et al., 2012b; Wiggins et al., 2014). Each item was rated on a 3-point Likert scale from 0 (not true) to 2 (very true or often true). Good internal consistency (Aebi et al., 2013; Stringaris et al., 2012b; Wiggins et al., 2014) and a single-factor structure (Roberson-Nay et al., 2015; Stringaris et al., 2012b; Wiggins et al., 2014) have been demonstrated in past research. For the current sample, the

3. Data analysis

All analyses were performed using SPSS. There were no missing data in the task variables at both visits; irritability measures were missing in some participants (n=4–9 at Visit 1; n=10–16 at Visit 2). Missing data were handled with pairwise deletion in the analyses.

3.1. Test-retest reliability

We computed Pearson’s correlation (r) and ICC (two-way mixed effect model, ICC (C, k); McGraw and Wong (1996)) to evaluate the test-retest reliability of the Affective Posner 2 task variables (accuracy, RT) and valence ratings (happy/unhappy, frustration). We also computed Pearson’s correlation and ICC to determine the test-retest reliability of response cost (i.e., invalid vs. valid trials) on accuracy and RT. All of these analyses were conducted separately for non-frustration and frustration runs. To determine the test-retest reliability of the irritability measures (i.e., child ARI, parent ARI, and the CBCL irritability), Pearson’s r and ICC were also calculated. We evaluated reliability using Cicchetti and Sparrow’s (1981) definition for judging the clinical significance of ICC values: <.40= poor; .40–.59= fair; .60–.74= good; >.74= excellent (Cicchetti and Sparrow, 1981).

3.2. Construct validity and convergent validity

To test the construct validity of the Affective Posner 2 task, i.e., whether the task elicits negative affect effectively at both visits, we examined changes in valence ratings over the course of the task and across visits. To this end, we conducted two separate two-way Runs (5)×Visit (2) repeated-measures analyses of variance (ANOVA), one for happy/unhappy ratings and the other for frustration ratings. We also evaluated convergent validity of the task by examining correlations between task variables (i.e., accuracy, RT, response cost, ratings) and child- and parent-reported irritability. To determine the convergent validity of the irritability measures (i.e., child ARI, parent ARI, and the CBCL irritability) and the agreement between these measures, Pearson’s r and ICC were also calculated.

4. Results

4.1. Test-retest reliability of task ratings and behavior

Accuracy, RT, happy/unhappy ratings, and frustration ratings during both non-frustration runs and frustration runs were reliable and stable over 2–4 weeks (r’s=.28 – .64, p’s < .01; ICCs =.44 to .78, p’s < .001; Table 2). Most measures had good to excellent test-retest reliability (ICC >.60); during non-frustration runs, the happy/unhappy ratings and frustration ratings had fair test-retest reliability (ICCs =.44 and .56). Test-retest reliability for happy/unhappy ratings during frustration runs was better than those during non-frustration runs (r’s=.63 vs. r=.28, p=.001, using Fisher’s r-to-z transformation); similar finding was evident for frustration ratings but at a trend level (r =.58 vs. r =.39, p=.07). Of note, the lower reliability of ratings during non-frustration runs compared to that during frustration runs may be due to a restricted range of variances in the ratings.

Response cost in accuracy and RT during both non-frustration runs and frustration runs were reliable and stable over 2–4 weeks (r’s=.28 – .55, p’s < .01; ICCs =.44 to .70, p’s < .001; Table 3). Specifically, response cost in accuracy had good test-retest reliability (ICCs =.70 and .64), while response cost in RT had fair test-retest reliability (ICCs =.58 and .44).
Test-retest reliability of task variables.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Visit 1 Mean (SD)</th>
<th>Visit 2 Mean (SD)</th>
<th>Pearson’s r</th>
<th>Intraclass Correlations (ICC)</th>
<th>Difference Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Frustration Runs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>97.00 (3.08)</td>
<td>96.15 (3.75)</td>
<td>.54***</td>
<td>.69***</td>
<td>−.85</td>
</tr>
<tr>
<td>RT (ms)</td>
<td>475.04 (108.19)</td>
<td>456.50 (120.91)</td>
<td>.57***</td>
<td>.72***</td>
<td>−18.54</td>
</tr>
<tr>
<td>Happy/unhappy rating</td>
<td>2.61 (1.64)</td>
<td>2.14 (1.45)</td>
<td>.28**</td>
<td>.44***</td>
<td>−.47</td>
</tr>
<tr>
<td>Frustration rating</td>
<td>1.96 (1.44)</td>
<td>1.95 (1.42)</td>
<td>.39**</td>
<td>.56**</td>
<td>−.01</td>
</tr>
<tr>
<td>Frustration Runs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>82.34 (11.97)</td>
<td>83.48 (11.77)</td>
<td>.64***</td>
<td>.78***</td>
<td>1.14</td>
</tr>
<tr>
<td>RT (ms)</td>
<td>334.24 (70.09)</td>
<td>351.96 (89.92)</td>
<td>.64***</td>
<td>.76***</td>
<td>17.72</td>
</tr>
<tr>
<td>Happy/unhappy rating</td>
<td>4.74 (2.53)</td>
<td>4.00 (2.47)</td>
<td>.63***</td>
<td>.77***</td>
<td>−.74</td>
</tr>
<tr>
<td>Frustration rating</td>
<td>5.49 (2.66)</td>
<td>4.38 (2.74)</td>
<td>.58**</td>
<td>.74**</td>
<td>−1.11</td>
</tr>
</tbody>
</table>

Note. RT = reaction time; ms = milliseconds.
***p < .001.
**p < .01.
*p < .05.

4.2. Test-retest reliability of, and agreement between, child and parent irritability measures

Child-reported ARI, parent-reported ARI, and CBCL irritability scores were all reliable and stable over 2–4 weeks (r’s=.79 – .82, p’s < .001; Table 4); ICCs=.88, .90, and .90, respectively (p’s<.001), suggesting excellent test-retest reliability for these measures. At both visits, the agreement between child-reported and parent-reported ARI was only low to moderate (r’s=.37 and .22, p’s<.05), as was the agreement between child-reported ARI and parent-reported CBCL irritability (r’s=.28 and .29, p’s<.01; Table 4). At both visits, parent-reported ARI was highly correlated with parent-reported CBCL irritability (r’s=.64 and .54, p’s<.001; Table 4), providing support for the convergent validity of these measures.

4.3. Construct validity and convergent validity of the task

As shown in Fig. 2a and b, participants felt less happy and more frustrated in frustration runs than in non-frustration runs, demonstrating the construct validity of the task, i.e., the ability of the task to elicit negative affect. This was true for both Visit 1 and Visit 2. Happy/unhappy ratings and frustration ratings during frustration runs dropped from Visit 1 to Visit 2; however, at Visit 2, the ratings during frustration runs were still higher than those during non-frustration runs.

At both visits, accuracy, RT, and response cost did not correlate with child- or parent-reported irritability (r’s=.17 to .19, p’s>.07), with the exception of a positive, yet small, correlation between parent-reported CBCL irritability and RT in non-frustration runs at Visit 1 (r=.20, p<.05). The only consistently significant findings were between child-reported ARI and self-reported ratings during task (unhappiness and frustration) in frustration runs (r’s=.32 and .33, p’s<.001, at Visit 1 and 2, between child ARI and unhappiness; r’s=.23 and .26, p’s<.05, at Visit 1 and 2, between child ARI and frustration).

4.4. Additional analyses

Because this was predominately a twin sample, we conducted analyses to account for non-independence in our data due to including twins from the same family. Specifically, we conducted mixed models with family unit as the random effect to evaluate reliability and validity of task and irritability measures. We also adjusted for gender, days between visits, and IQ in the mixed models, and calculated partial correlations after covarying gender, days between visits, and IQ. All the correlations remained similar and significant, suggesting that reliability and validity were not inflated due to nested family effect and was not influenced by the gender of the child, days between visits, or IQ.

5. Discussion

The present study examined test-retest reliability and validity of a frustration paradigm, i.e., the Affective Posner 2 task, and two measures of irritability, i.e., the ARI and the CBCL irritability scale. We found that the accuracy, RT, response cost (validity effect), happy/unhappy ratings, and frustration ratings of the Affective Posner 2 task demonstrated good test-retest reliability over 2–4 weeks. Child-reported ARI, parent-reported ARI, and CBCL irritability scores all showed excellent test-retest reliability over 2–4 weeks. In addition, our frustration paradigm demonstrated validity, i.e., it effectively elicited negative affect in youth. Specifically, participants felt less happy and more frustrated after frustration, compared to non-frustration, runs. This is true even when participants performed the task for the second time after 2–4 weeks. Moreover, self-rated unhappiness and frustration during frustration runs correlated positively with self-reported but not...
parent-reported irritability, providing only modest support for the convergent validity of the task with other irritability measures. Overall, these findings suggest that the Affective Posner 2 task and the irritability measures are reliable and potentially valid and may be useful for future studies on childhood irritability and frustration.

It should be noted that the ICCs and Pearson’s correlations for the variables in the non-frustration runs were lower and less reliable than those in the frustration runs. This is expected because the variances of the variables in the non-frustration runs were low. Compared to the frustration runs, the non-frustration runs were relatively easy; participants’ mean accuracy was over 96% during the non-frustration runs for both visits, and most participants were happy and not frustrated because they were winning money. As a result, the variables in the non-frustration runs were less reliable due to the restricted range of variance.

Table 4
Reliability of and Correlations between Irritability Measures.

<table>
<thead>
<tr>
<th>Visit</th>
<th>Child ARI (n=104)</th>
<th>Parent ARI (n=105)</th>
<th>CBCL Irritability (n=100)</th>
<th>Visit</th>
<th>Child ARI (n=99)</th>
<th>Parent ARI (n=98)</th>
<th>CBCL Irritability (n=98)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visit 1</td>
<td>Child ARI</td>
<td>—</td>
<td>.37***</td>
<td>.28**</td>
<td>.79***</td>
<td>.26</td>
<td>.26</td>
</tr>
<tr>
<td>Parent ARI</td>
<td>—</td>
<td>—</td>
<td>.64***</td>
<td>.29**</td>
<td>.82***</td>
<td>.56</td>
<td>.56</td>
</tr>
<tr>
<td>CBCL Irritability</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.23</td>
<td>.55</td>
<td>.82</td>
<td></td>
</tr>
<tr>
<td>Visit 2</td>
<td>Child ARI</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.22</td>
<td>.29</td>
<td></td>
</tr>
<tr>
<td>Parent ARI</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBCL Irritability</td>
<td>Mean (SD)</td>
<td>3.43 (3.00)</td>
<td>1.55 (2.49)</td>
<td>.88 (1.44)</td>
<td>2.94 (3.13)</td>
<td>1.49 (2.15)</td>
<td>.74 (1.26)</td>
</tr>
</tbody>
</table>

Note. ARI=Affective Reactivity Index; CBCL=Child Behavior Checklist.

Although overall our frustration paradigm demonstrated fair to good test-retest reliability, it is less stable compared to parent- or child-report of irritability. This may be because the two types of assessments (i.e., clinical rating scale vs. task paradigm) elicit information about two different aspects of irritability, e.g., trait vs. state. Irritability as a trait has been found to be relatively stable (Caprara et al., 2007; Leibenluft and Stoddard, 2013). It is possible that our frustration paradigm may provide information about irritability that is more state-like (e.g., aberrant responses to blocked goals such as anger and frustration). In that case, one would expect its measures to be more context-dependent and less reliable than parent- and child-reports, which also assess trait-like irritability, and thus may be more context-independent and hence more reliable.

Our reliability findings have important implications for treatment and longitudinal research. Test-retest reliability sets an upper limit for longitudinal stability because stability is reduced by both test-retest unreliability and true score changes (McCrae et al., 2011). As such, measures with poor test-retest reliability have limited ability to detect reliable and true scores changes, i.e., changes that reflect meaningful developmental alterations, rather than variation due to unreliability. Therefore, a reliable measure is fundamental to detect meaningful changes associated with development or intervention (Leon, 2008). Similarly, test-retest reliability is important for behavioral genetic research because poor test-retest reliability may attenuate heritability estimates (McCrae et al., 2011). Of note, the fact that participants were still frustrated when they completed the task for the second time after 2–4 weeks indicated that the Affective Posner 2 may be an effective probe to use in clinical trials (i.e., pre- and post-treatment) in which irritability and aberrant responses to frustration are a target of treatment.

At both visits, the correlations between irritability measures was high within informants (i.e., parent ARI and parent CBCL irritability), but low to moderate between informants (i.e., child ARI and parent ARI or child ARI and parent CBCL irritability). This is consistent with previous research on youth’s behavioral and emotional symptoms, which typically finds high agreement within the same informant or similar informants (e.g., pairs of parents) and low agreement between informants (e.g., subjects and parents; see Achenbach et al., 1987 for a meta-analysis; De Los Reyes and Kazdin, 2005; Youngstrom et al., 2000). The low agreement between parent- and child-reports is worth noting because it suggests that parents and children each contribute variance not accounted for by the other; thus, children’s self-report cannot substitute for reports by the parents, and vice versa. At both visits, the correlation between parent ARI and parent CBCL irritability was high (r=.54 and .64), suggesting convergent validity of the two
measures. However, there are still non-shared variances between these two measures, suggesting that the ARI and CBCL irritability may each provide unique information about irritability that is not measured by the other. Given the importance of a multi-method approach to minimize measurement error from a single source (Cicchetti and Cohen, 1995), future research may benefit from including both measures to assess irritability.

Of note, at both visits, child-reported ARI positively correlated with frustration and unhappy ratings during frustration runs. However, child-reported ARI was not related to other task variables (i.e., accuracy, RT, response cost), and parent-reported irritability was not related to either performance or ratings during the task. The null associations between parent-reported irritability and the task variables may be attributed to the non-clinical nature of the sample, i.e., low levels of parent-reported irritability in this community sample. Such poor agreement between parent-reported irritability and the task variables is consistent with the low cross-informant agreement in the child development and psychopathology literature mentioned above (Achenbach et al., 1987; De Los Reyes and Kazdin, 2005; Youngstrom et al., 2000). The correlations between child ARI and task ratings, although significant, were moderate and may be due to shared method variance, i.e., both were reported by the same informant (the child). Although modest associations between questionnaire measures and behavioral tasks are to be expected, this finding suggests that the two types of assessments may be assessing different aspects of irritability (i.e., trait vs. state as described earlier) or even different constructs. Thus, our findings only provide modest support for the convergent validity of the task with other irritability measures and call for validations from future research.

The present findings should be evaluated in light of the following limitations. First, the sample consists of mostly twins, although we accounted for this effect in the analyses. Also, the age range of the sample is narrow (i.e., age 9–14). Thus, results may not generalize to non-twin samples or youth outside of this age range. Second, test-retest reliability may be decreased in clinical populations due to disease progression or symptom fluctuation (Duﬀ, 2012). Future research with clinical samples is needed to assess whether the current task and measures have acceptable reliability in youth with disorders. Third, children who were on medications for ADHD were included in the study. Medications such as stimulants may have affected performance (e.g., RT) or ratings during the task. Moreover, it is unclear whether parents and children provided irritability ratings based on behaviors on or off medications. Finally, the Aﬀective Posner 2 paradigm does not allow us to parse out different aspects of reward processes (e.g., reward anticipation and outcomes). This paradigm may be measuring frustration in a speciﬁc context (i.e., receipt of nonreward). It is unclear whether frustration arose from loss of expected reward (frustrative nonreward), or anticipation of nonreward, or both.

In conclusion, the Aﬀective Posner 2 task demonstrated good test-retest reliability and modest validity, although it successfully evoked frustration in youth at both test and retest. The parent- and child-reports of the ARI and the parent-report of the CBCL irritability scale also showed good test-retest reliability. These reliability data provide an important basis for future studies to further investigate the Aﬀective Posner 2’s validity, as validity cannot be achieved without adequate reliability. Together, our study suggests that the Aﬀective Posner 2 paradigm, as well as the ARI and CBCL irritability scales, may be useful tools for research on irritability.

Contributors

Tseng, W.L. prepared the data, conducted the analyses, interpreted the results, and drafted the manuscript. Moroney, E. and Machlin, L. prepared the data and assisted in data analyses and drafting of the manuscript. Roberson-Nay, R. and Hettema, J.M. contributed to the conceptualization and design of the study, data interpretation, and provided critical feedback on the manuscript. Carney, D. assisted in data collection and data preparation, and provided critical feedback on the manuscript. Stoddard, J. and Towbin, K.A. assisted in data interpretation and provided critical feedback on the manuscript. Pine, D.S., Leibenluft, E., and Brotman, M.A. contributed to the conceptualization and design of the study, data interpretation, and provided critical feedback on the manuscript. All authors approved the manuscript prior to its submission.

Financial disclosures

The authors disclose no conﬂicts of interest related to this work.

Acknowledgements

This research was in part supported by the Intramural Research Program of the National Institute of Mental Health (NIMH), National Institutes of Health (NIH). This study was also supported by R01MH098055 (JMH) and R01MH101815 (RRN).

References

McCrae, R.R., Kurz, J.E., Yamagata, S., Terracciano, A., 2011. Internal consistency,


